Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2022 SCPE. Volume 23, Issues 4, pp. 193–210, DOI 10.12694/scpe.v23i4.2019

AN EFFICIENT NOVEL APPROACH WITH MULTI CLASS LABEL CLASSIFICATION THROUGH MACHINE LEARNING MODELS FOR PANCREATIC CANCER

P. SANTOSH REDDY *AND M. CHANDRA SEKHAR †

Abstract. Pancreatic cancer is right now the fourth largest cause of cancer-related deaths. Early diagnosis is one good solution for pancreatic cancer patients and reduces the mortality rate. Accurate and earlier diagnosis of the pancreatic tumor is a demanding task due to several factors such as delayed diagnosis and absence of early warning symptoms. The conventional distributed machine learning techniques such as SVM and logistic regression were not efficient to minimize the error rate and improve the classification of pancreatic cancer with higher accuracy. Therefore, a novel technique called Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System (DHEGQDRLCS) is developed in this paper. First, the number of data samples is collected from the repository dataset. This repository contains all the necessary files for the identification of prognostic biomarkers for pancreatic cancer. After the data collection, the separation of training and testing samples is performed for the accurate classification of pancreatic cancer samples. Then the training samples are considered and applied to Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System. The proposed hybrid classifier system uses the Kernel Quadratic Discriminant Function to analyze the training samples. After that, the Elitism gradient gene optimization is applied for classifying the samples into multiple classes such as non-cancerous pancreas, benign hepatobiliary disease i.e., pancreatic cancer, and Pancreatic ductal adenocarcinoma. Then the Reinforced Learning technique is applied to minimize the loss function based on target classification results and predicted classification results. Finally, the hybridized approach improves pancreatic cancer diagnosing accuracy. Experimental evaluation is carried out with pancreatic cancer dataset with Hadoop distributed system and different quantitative metrics such as Accuracy, balanced accuracy, F1-score, precision, recall, specificity, $TN, TP, FN, FP, ROC_{AUC}, PRC_{AUC}$, and PRC_{APS}. The performance analysis results indicate that the DHEGQDRLCS provides better diagnosing accuracy when compared to existing methods.

Key words: Pancreatic cancer diagnosis, Kernel Quadratic Discriminant analysis, Elitism gradient gene optimization, Reinforced Learning technique, Hadoop distributed system.

AMS subject classifications. 68T05

1. Introduction.

PDAC (pancreatic ductal adenocarcinoma) is still an incurable cancer that kills a lot of people. Despite breakthroughs in the identification and treatment of other gastrointestinal cancers, such as colorectal and gastric cancers, PDAC mortality rates only slightly exceed the number of newly diagnosed cases, with 5-year survival rates as low as 3-15 percent 3,4. These poor results are due to late and incurable stage diagnosis, as well as significant chemo-resistance in tumours[7][8]. Most treatment approaches are rendered useless due to the latter. The importance of early diagnosis is well recognised, and various healthcare groups have advocated for a move toward early detection. A lot of findings indicate the advantages of early PDAC detection.

Patients diagnosed with stage I illness have a higher chance of surviving than those diagnosed with later stages. Similarly, when compared to PDAC [2] detected when individuals have symptoms, incidentally diagnosed PDAC is related with longer median survival. Patients diagnosed at an operable stage have a significantly higher chance of survival than those detected at an inoperable stage. Unfortunately, up to 85% of instances are not surgically resectable when they are discovered, and in the United Kingdom, more than half of cases are detected after a non-specific illness course that results in an emergency hospital admission 10-13.

Biomarkers may play an important role in the early detection of PDAC by enriching for persons in high-risk groups who are more likely to develop cancer, allowing doctors to prioritize individuals for screening. Despite

*Department of Computer Science and Engineering, Presidency University, Bengaluru, India (santosh.reddyp@presidencyuniversity.in)

†Department of Computer Science and Engineering, B. N. M. Institute of Technology, Bengaluru, India (mchandrasekhar@presidencyuniversity.in)

thousands of articles, no one candidate biomarker for the early identification of PDAC has been translated into clinical application. Indeed, developing biomarkers for this disease has distinct hurdles. Due to the low prevalence of PDAC, obtaining the sufficient number of samples for biomarker development is difficult, necessitating extensive national and international collaborations. Pancreatic tumors are extremely diverse, both within and between people.

As a result, single biomarkers are unlikely to be sensitive enough to detect PDAC, and compre-hensive panels of biomarkers will be necessary. PDAC biomarker research requires an understanding of and accounting for any confounding factors68. The majority of important studies now include illness controls, such as chronic pancreatitis. 68 Furthermore, people are becoming more aware that obstruc-tive jaundice can contribute to false-positive biomarker results69-71. The fact that a large percentage of PDAC patients have diabetes mellitus (DM), 50 is not adequately accounted for in biomarker re-search, and developing biomarkers risk being associated with DM rather than PDAC.

Finally, a key limitation in early detection studies for PDAC has been the lack of specialized pre-diagnostic groups. When weighing the cost vs. value of a biomarker-assisted screening programme, take into account the costs of both the first screening and the subsequent tests required to confirm the diag-nosis. Since both genuine positive and false positive tests necessitate additional investigation, high spec-ificity biomarkers are required.

Pancreatic cancer is the fourth most prevalent cause of cancer death and the second most common cause of death from neoplasms that disrupt digestion. Regular pancreas segmentation, on the other hand, remains a point of contention for the following reasons: 1) CT scans with low soft tissue contrast. 2) Significant anatomical differences. In terms of size and placement in the abdominal cavity of patients, the pancreas is very unpredictable anatomically [6][9]. The pancreas is a flexible tissue that yields. As a result, the shape and appearance of the pancreas fluctuate significantly between individu-als.

1.1. Contributions of the paper. In order to solve the existing issues, a novel DHEGQDRLCS is introduced. The contribution of the proposed DHEGQDRLCS is listed below:

- To improve the pancreatic cancer diagnosis accuracy, a novel DHEGQDRLCS technique is introduced as a reliable diagnostic tool to improve the clinical practicality for diagnosing pancreatic cancer early based on hybridization of Kernel Quadratic Discriminant Function, Elitism gradient gene optimization, and Reinforced Learning technique.
- A Kernel Quadratic Discriminant Function is applied to the DHEGQDRLCS technique for analyzing the correlation between the data samples and the class means value by using the kernel function. Then the Elitism gradient gene optimization technique finds the maximum correlated results for classifying the samples into a particular class. After that, reinforcement learning is applied to find the minimum loss function. This process increases the accuracy and minimizes incorrect pancreatic classification.
- Extensive experimentation is conducted with Hadoop distributed system to measure the performance of the DHEGQDRLCS technique and other related works. The obtained result shows that our proposed, DHEGQDRLCS technique provides better performance than the existing Distributed eSVM (DeSVM) and Distributed eLR (DeLR) Models.

2. Literature Survey. In a recent analysis of 3.9 million cancer patients from seven countries (Australia, Canada, Denmark, Ireland, New Zealand, Norway, and the United Kingdom) assessing seven cancer sites (esophagus, stomach, colon, rectal, pancreas, lung, and ovary), PDAC was found to have the lowest 5-year survival rates (ranging from 7.9 percent in the UK to 14.6 percent in Australia). 4 Early discovery will undoubtedly aid in the improvement of these statistics, and as our review demonstrates, progress has already been made. The development and validation of biological and epidemiological markers will benefit from the construction of new customized cohorts (of people with NOD or symptoms).

Careful and ethical use of existing data, whether via social media or electronic health records, can help prediction models, while AI applied to imaging can help discover lesions early. Identifying the small number of people with MCLs who are at the highest risk of developing PDAC is still a major knowledge gap in the field of mucinous cysts. Much work, including the ongoing trials discussed here, has to be done to enhance the early detection of PDAC. The ongoing cohort studies are critical in many ways, not least because they help to raise awareness of PDAC symptoms and their relation to NOD among healthcare practitioners and patients.

Early detection advances will be combined with therapy advances to help people with PDAC live longer. According to the concept of regular automation algorithms, Support Vector Machine is an urgent classification algorithm for categorising data connected to the calculation of Wisconsin Breast Cancer data in a short amount of time [10]. Proportion of relative outcomes for four alternative algo-rithms for data retrieval and automatic automation in terms of efficacy and effectiveness. Initiates a new function for the benefit of the medical health system, with the purpose of predicting an average patient's outcome in the analysis of electronic medical processes and the recognized parameters of parameters developed for optimal operation [11].

The application coordination with the estimation of data for variable effects, categories of ef-fects, and the threshold parameter to identify the disease diagnosis generally provides efficient prognos-tic data. They use and cover medical proceedings for blood cancer, heart failure, and diabetes [12]. Using data from combination radiation (EBRT) and brachytherapy (BT), construct a function based on the red convolution neural representation to analyse rectal prescribed amount sharing and forecast rectal toxicity in patients with uterine cancer [13]. To have an impact on patient data, they established a place to live and transfer approach.

To increase the dates for video data losses and loss factors, the adaptive synthetic model technique is applied. To construct RSDM discriminate regions with the calculation model, create Gradi-ent Activation Weight Map (Grad-CAM) classes[19][20]. A combination of experimental results is used to examine the CNN-based representation for predicting rectal dose by means of transfer therapy for uterine cancer radiation.

A novel machine learning technique with the twin support vector machine (TWSVM) was developed in [21] (i.e. distributed extended SVM) for identifying pancreatic cancer early. However, the higher pancreatic cancer detection accuracy was not achieved. Unconditional and conditional logistic regression models were developed in [22] (i.e. Distributed extended LR) for pancreatic cancer diagnosis by measuring the relationship between fasting glucose levels and pancreatic cancer risk. But, the higher precision and recall analysis was not achieved.

A Fully Automatic Deep Learning Framework was developed in [23] for Pancreatic Ductal Adenocarcinoma Detection. But the Framework was not efficient to perform the multiclass classification. Multi-Omics DEep Learning for Prognosis-correlated subtyping (MODEL-P) was developed in [24] to identify Pancreatic Ductal Adenocarcinoma Pancreatic But it failed to improve the robustness of the deep learning model while handling more samples for pancreatic cancer detection. Three classification algorithms namely linear discriminant, analysis (LDA), support vector machine (SVM), and k-nearest neighbor (KNN) were developed in [25] for the classification of benign and malignant pancreatic tumors. However, the designed algorithms failed to establish a more reliable classification model for the accurate detection of pancreatic cancer.

Different Blood biomarkers classes were analyzed in [26] for differential diagnosis and early detection of pancreatic cancer. But the analysis was not considered with a large number of samples. A grouped" neural network (GrpNN) architecture was designed in [27] to generate a dimensionally reduced vector for early detection of pancreatic cancer. But it failed to apply to multi-modal clinical data sets. A combinatorial approach consisting of Particle Swarm Optimization (PSO), Artificial Neural Network (ANN), and Neighborhood Component Analysis (NCA) iterations was developed in [28] for pancreatic cancer diagnoses. But, the performance of pancreatic cancer detection was not improved.

3. System architecture. The system architecture is the conceptual model that describes the structural views. A system architecture also consists of the number of processes involved and works together to implement the overall system.

Figure 3.1 illustrates the system design architecture. First, the input is collected from the corresponding dataset. Then the input data gets preprocessed. Pre-processing is an essential step in the data mining process for transforming the raw data into a structured format that helps to improve accuracy and minimize time complexity. Next, data separation is performed to split the preprocessed dataset into the testing and training data. In machine learning, data separation is used to avoid overfitting. It is a modeling error that occurs when a function is too strongly fit to a limited set of data points. Then the testing and training samples are given to the machine learning techniques such as Distributed Extended Support Vector Machines (i.e.: Distributed eSVM), Distributed Extend Logistic Regression Model, and Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System. These techniques train a model on known input data and provide future outputs. The processes of three machine learning techniques are described as given below.

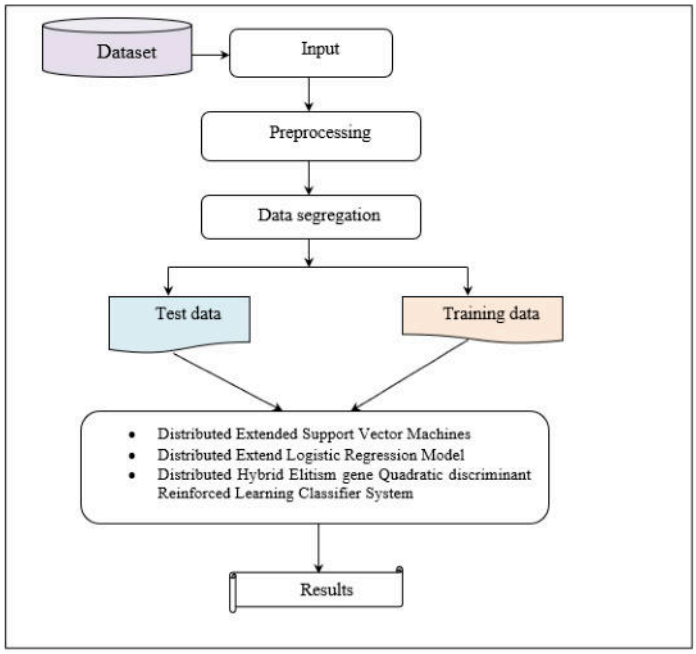


Fig. 3.1: System Design

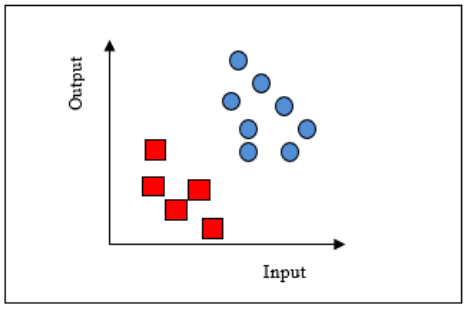


Fig. 3.2: Distributed Extended Support Vector Machines

3.1. Distributed eSVM. Distributed Extended Support Vector Machines (DeSVM) [21] is a data classification approach that uses hyperplanes to categorize the data into different classes. In DeSVM, Distributions are considered any population of data that has scattered in the two-dimensional space. The DeSVM technique is often useful for data with non-regularity, or data with an unknown distribution. Let us consider the DeSVM has two types of values, which are shown in figure 3.2.

Figure 3.2 illustrates the DeSVM principle used to build numerous separating hyperplanes from labeled data, dividing the data space into different classes where only one type of data is distributed in each segment. The DeSVM technique is often useful for data with non-regularity, or data with an unknown distribution [18]. To solve the classifier, then draw the straight line $y = ax + b$ that separates the red and blue items to classify the data above and below.

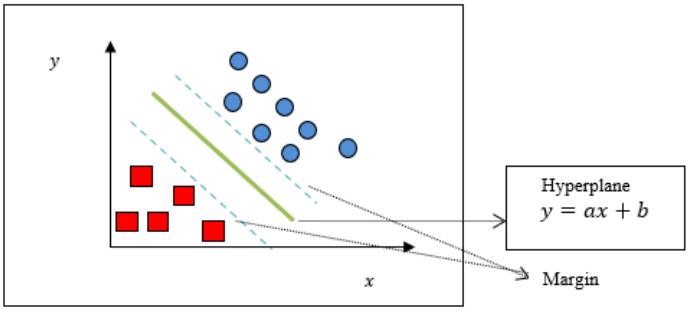


Fig. 3.3: Classification results

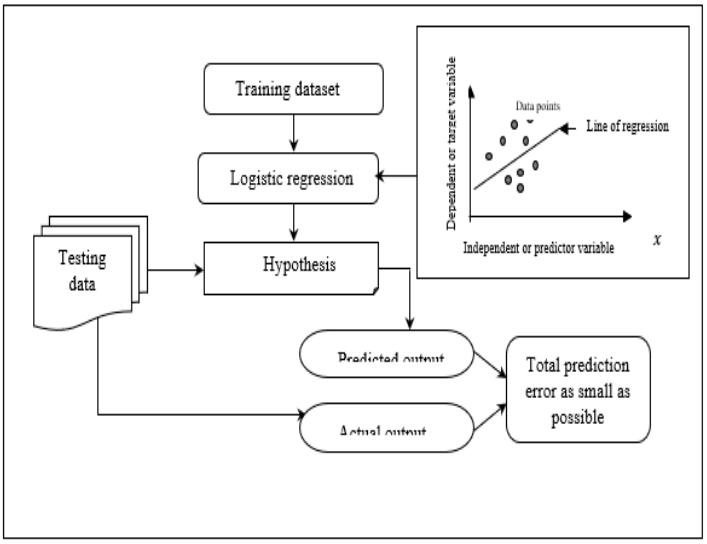


Fig. 3.4: Logistic Regression model

Figure 3.3 illustrates the classification results of DeSVM. DeSVM generates the multiple separating hyperplanes such that the data space is divided into multiple classes and each segment contains only one type of data. The DeSVM works based on finding a decision boundary (i.e hyperplane) between binary classes and maximizing the margin i.e. the distance between the decision boundary and the closest data points.

3.2. Distributed Extend Logistic Regression Model (Modified Linear Regression). The Distributed Extend Logistic Regression Model (DeLR) [22] is a machine learning technique employed to analyze the relationship between one or more independent variables (x) on the training data set and a dependent variable (y). By applying a Linear Regression, that employs a linear function called a hypothesis. To solve the parameters of this hypothesis equation, the cost function also known as the squared error function is minimized for regression problems. These parameters are determined using the gradient descent approach.

Figure 3.4 illustrates the process of Logistic Regression where the modeling technique provides the rela-

tionship or correlation between the dependent and independent variables. Logistic Regression is a type of prescriptive modeling technique that measures the relationship or correlation between the dependent and independent variables. In the two-dimensional space, 'x' axis represents the independent variable and 'y' represents the dependent variable.

Linear regression with only one variable is represented as

$$y \ = \ AX \ + \ B \tag{3.1}$$

Multiple linear regression is expressed as

$$y \ = A_0 + A_1 * X_1 \ + \ A_2 * X_2 + A_3 * X_3 + \ldots . A_n * X_n \tag{3.2}$$

where A denotes a coefficient. The goal of machine learning is to establish a mapping between data

$$f : X \rightarrow Y \tag{3.3}$$

The mapping (f) between input variables 'X' and output variables 'Y' is represented by a regression model. The Logistic Regression specifies a line that fits the relationship between the input variable (X) and the output variable (Y) by determining the specific weight of the input variable coefficient. The hypothesis analysis the training and testing results and provides the predicted output. Finally, the classification result with minimum error is obtained.

3.3. Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System. A Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System (DHEGQDRLCS) is an adaptive system that integrates machine learning, evolutionary computing, and other heuristics to deal with a Pancreatic Cancer diagnosis. The conventional DeSVM and DeLR single optimal models not having the ability to perform accurate multiclass classification while considering a large volume of input samples. Contrary to the conventional method, the purpose of DHEGQDRLCS is to accomplish the issue of a single optimal model by constructing a hybridized model. In other words, the DHEGQDRLCS have the ability to solve the multi-class classification while handling the large volume of input samples by integrating three different techniques namely Elitism gradient gene optimization, Reinforcement algorithm, and kernel Quadratic discriminant analysis.

Figure 3.5 depicts the architecture diagram of the proposed DHEGQDRLCS to perform accurate Pancreatic Diagnosis. Let us consider the Pancreatic dataset D'and collect the number of 590 data samples $S_i \in S_1, S_2, S_3, ..S_n$. The 590 urine samples are collected for the four biomarker panels, including 183 control samples, 208 benign hepatobiliary disease, and 199 pancreatic ductal adenocarcinomas (PDAC) samples. The four biomarker panels are creatinine, LYVE1, REG1B, and TFF1. The dataset includes 14 columns (i.e. attributes) listed in table 3.1.

After the data collection, training and testing are identified for pancreatic cancer diagnosis. Then the proposed technique performs samples are separated into training and testing subsets. The separation of training and testing is a technique used for evaluating the performance of a machine learning algorithm. It is also used for solving classification problems for any supervised learning algorithms. The process involves taking a dataset and dividing it into two subsets. The first subset is called as training dataset set used to fit the machine learning model. The second subset called the testing dataset is used to train the model rather than the input provided to the model. In the proposed technique, each subset has half of the samples. After the separation, the pancreatic cancer diagnosis is performed using Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System (DHEGQDRLCS). The proposed DHEGQDRLCS technique integrates three methods as Elitism gradient gene optimization, Reinforcement algorithm, and kernel Quadratic discriminant analysis to diagnose pancreatic disease.

The proposed DHEGQDRLCS first uses the kernel Quadratic discriminant analysis for analyzing the testing and training samples. Kernel Quadratic discriminant analysis is a supervised machine learning classifier used to classify the samples into two or more classes based on likelihood estimation with help of Gaussian kernel functions. A likelihood method is a measure of the relationship between the training and testing data samples.

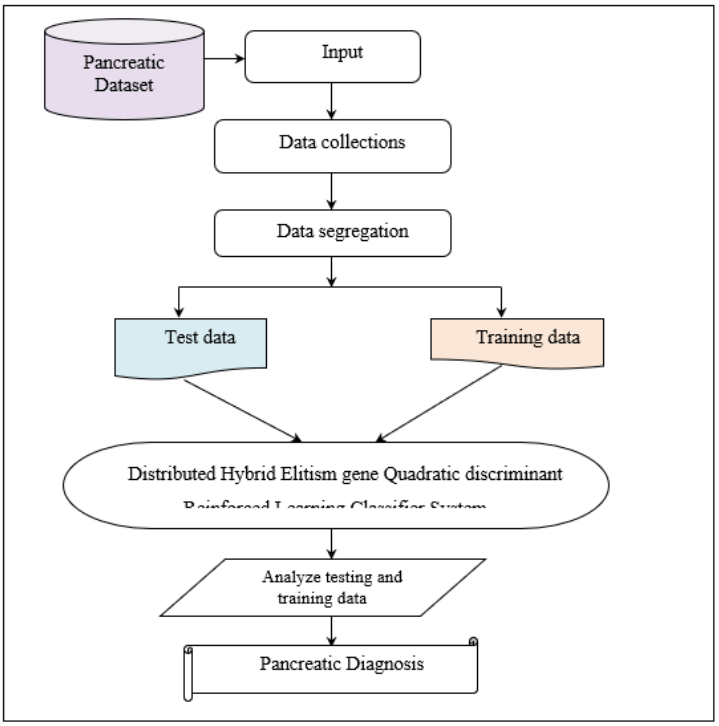


Fig. 3.5: Architecture diagram of proposed DHEGQDRLCS

Let us consider the number of training samples in the dataset. First, the number of classes 'C_j = 'C_1,'C_2,..'C_n are defined. The mean value of the class is calculated as given below:

$$M_c = \frac{1}{n} \sum_{i=1}^{n} S_i \tag{3.4}$$

From (3.4), M_c denotes a mean value of class and n denotes the number of training samples 'S_i'. Therefore, the correlation between the mean and samples is measured by applying a Kernel Quadratic discriminant analysis as given below:

$$R\left(S_i, M_{cj}\right) = \frac{1}{\sqrt{2\pi}\ d} \exp\left[-0.5\left[\frac{\left|S_i - M_{cj}\right|}{d}\right]^2\right] \tag{3.5}$$

where $R\ \left(S_i, M_{cj}\right)$ represents a Kernel Quadratic discriminant analysis output, d represents the deviation, M_c represents the mean of the class, S_i indicates a training sample. After that, the maximum correlation for the training samples being classified into the particular class is identified through the Elitism gene optimization technique.

Elitism gradient gene optimization is a metaheuristic evolutionary algorithm to generate high-quality solutions in the search problems (i.e. finding maximum correlated results. By applying the Elitism gradient gene optimization technique, first, the population of genes (i.e. correlations between the samples and mean of classes) are initialized in search space.

$$W = R_1\ \left(S_i, M_{cj}\right), R_1\ \left(S_i, M_{cj}\right)..R_n\ \left(S_i, M_{cj}\right) \tag{3.6}$$

Table 3.1: Attribute descriptions

S.no	Attributes	Description
1	sample_id	id of the sample
2	patient_cohort	Any group of individuals affected by common diseases,
3	sample_origin	origin of the patient samples (BPTB: Barts Pancreas Tissue Bank)
4	Age	Patient age
5	sex	M- Male F-Female
6	diagnosis	1. non-cancerous pancreas 2.benign hepatobiliary disease i.e., pancreatic cancer, 3. Pancreatic ductal adenocarcinoma
7	stage	stages of pancreatic cancer IA, IB, IIA, IIIB, III, IV
8	benign_sample_diagnosis	benign hepatobiliary disease samples
9	plasma_CA19_9	Monoclonal antibody levels in blood plasma, which are frequently elevated in pancreatic cancer patients.
10	creatinine	Creatinine is a protein that commonly utilized as a kidney function indicator
11	LYVE1	Lymphatic vessel endothelial hyaluronan receptor 1 is a protein discovered in the urine that may have a role in tumour spread
12	REG1B	Urinary levels of a protein linked to pancreatic regeneration,
13	TFF1	Urinary Trefoil Factor 1 levels evaluated, which could be linked to urinary tract regeneration and repair
14	REG1A	REG1A: Urinary levels of a protein connected to pancreatic regeneration,

After the initialization, the fitness is measured based on the gradient ascent function also often called steepest descent for finding a local maximum of a function (i.e. arg max).

$$F\ (W) = arg\ \max\ R\ (S_i, M_{cj}) \tag{3.7}$$

where $F\ (W)$ indicates the fitness function, $arg\ max$ denotes an argument of the maximum function. After that, the Elitism selection operation is applied to a gene optimization for finding the fittest individual among the populations by setting the threshold 't'. Therefore, the optimal selection procedure is obtained as follows,

$$Z_p = \begin{cases} F\ (W) > t\ ; & \text{select individuals} \\ \text{Otherwise;} & \text{Reject the individuals} \end{cases} \tag{3.8}$$

where Z_p denotes an elitism selection outcome (i.e. predicted classification results), t denotes a threshold, $f\ (X)$ indicates fitness. In this way, optimum correlated results are identified and classified the sample into a particular class. As a result, the optimum correlated results are used to provide the final predicted output classification results.

Then applying the model-free reinforcement algorithm considers the state-action pair at each time step to predict the accurate classification results by minimizing the loss function. The state-action pair in the reinforcement algorithm finds the total amount of expected rewards for taking an action from the state. The state describes the current situation of the classifier whereas the action describes the set of possible decisions made after observing the results space (i.e. loss function) from the states.

During the learning process, loss of the classification is estimated based on the target class and the predicted classification results as given below,

$$l = \ (Z_t - Z_p)^2 \tag{3.9}$$

$$Z_t = r + \beta \; max \; Z'(s,a) \tag{3.10}$$

where l denotes a loss function, Z_t indicates a target classification result, 'Z_p' predicted classification results, r denotes a reward, β discount factor that denotes between 0 and 1, $max \; Z'(s,a)$ denotes a maximum predicted reward from all possible actions. In order to minimize the loss function, the output results get updated based on the learning rate.

$$Z_{p_{(t+1)}} = Z_p + \quad \tau \; (r + \beta \; max \; Z'(s,a) - Z_p)^2 \tag{3.11}$$

where $Z_{p_{(t+1)}}$ denotes an updated value of the output functions, Z_p indicates the previous classification value, τ denotes a learning rate $(0 < \tau < 1)$, r denotes rewards,β indicates a discount factor between 0 and 1. This process is iterated until finding the minimum loss. Finally, accurate classified results are obtained. Based on classification results, non-cancerous pancreas, benign hepatobiliary disease i.e., pancreatic cancer, and Pancreatic ductal adenocarcinoma are correctly diagnosed with minimum error. The DHEGQDRLCS algorithmic process is explained as follows:

// Algorithm 1: Distributed Hybrid Elitism gene Quadratic discriminant Reinforced Learning Classifier System

Input: Dataset**,** Number of data samples $\boldsymbol{S_i = S_1, S_2, \ldots S_n}$
Output: Improve the pancreatic diagnosing accuracy

Begin
Collect the data samples ' $S_i = S_1, \; S_2, \ldots S_n$ from the dataset
Separate testing and training samples
For each training samples ' S_i'
Define number of classes 'C_j'
for each class 'C_j
Compute mean value 'M_c'
end for
Measure the correlation between training samples and mean value '$R \; (S_i, M_{cj})$'
Obtain multiple correlation results
Generate the population of correlation results 'X '
For each individual '$R \; (S_i, M_{cj})$' in population
Measure the fitness '$f \; (X)$'
Apply elitism selection **'Z_p'**
if $(f \; (X) > t)$ **then**
Select the individuals
Classify the data into particular class
else
Reject the individuals
End if
For each classification results 'Z_p'
Compute the loss function 'l'
Update the **results** '$Z_{p_{(t+1)}}$ '
Obtain the classification results with minimum loss
End for
End

Algorithm 1 describes the step-by-step process of the proposed DHEGQDRLCS for accurate pancreatic disease diagnosis. First, the number of data samples is collected from the dataset. The collected data are separated into testing and training. The input training data samples are given to the hybrid learning classifier system. The proposed classifier analyzes the training data samples with the mean of kernel Quadratic discriminant function. Then the maximum correlation results are identified by applying the Elitism gradient gene optimization. Initialize the population of correlation results. Then gradient ascent function is applied to

measure the fitness of the individual. Followed by, the elitism selection procedure is applied for finding the maximum correlated results. Finally, the samples that highly correlated to the mean value are classified into that particular class. In this way, classification results are obtained. Finally, model-free reinforcement learning is applied for minimizing the loss function by updating the learning rate. In this way, accurate classification results are observed. Based on classification results, pancreatic cancers are correctly identified.

4. Experimental settings. In this section, experimental evaluation of proposed DHEGQDRLCS and existing DeSVM [21] and DeLR [22] are implemented in Java language with Hadoop distributed system by using a pancreatic dataset. In this article, the Hadoop Distributed System is applied that provides high-performance data access while handling a large volume of datasets. The pancreatic dataset includes 590 data samples and 14 attributes. The 590 urine samples are collected for the four biomarker panels, including 183 control samples, 208 benign hepatobiliary disease, and 199 pancreatic ductal adenocarcinomas (PDAC) samples. The attributes are sample_id, patient_cohort, sample_origin, Age, sex, diagnosis, stage, benign_sample_diagnosis, plasma_CA19_9, creatinine, LYVE1, REG1B, TFF1, and REG1A. The original dataset is divided into three subsets such as dataset 1, dataset 2, and dataset 3 to compare the accuracy in this paper.

5. Performance evaluation under various metrics. In this section, the performance analysis of the proposed DHEGQDRLCS and existing DeSVM and DeLR are discussed with different metrics such as accuracy, precision, recall, F1_Score, Specificity, true negative (TN), true positive (TP), False negative (FN), False positive (FP), ROC_AUC, PRC_AUC, and PRC_APS.

Accuracy: It is defined as the number of data samples that are correctly diagnosed into different classes. The accuracy is calculated as given below:

$$Accuracy = \left[\frac{TP + TN}{TP + TN + FP + FN}\right] * 100 \tag{5.1}$$

where TP indicates a true positive i.e. number of data samples correctly diagnosed as the non-cancerous pancreas or benign hepatobiliary disease i.e., pancreatic cancer, or Pancreatic ductal adenocarcinoma, TN indicates a number of true negatives i.e. normal samples correctly diagnosed as normal, FP represents the false positive i.e. normal samples incorrectly identified as pancreatic cancer, FN indicates a false negative i.e. cancer samples incorrectly identified as normal.

Precision: The precision is computed as given below:

$$Precision \ = \left[\frac{TP}{TP + FP}\right] * 100 \tag{5.2}$$

where, TP denotes the true positive, FP symbolizes the false positive.

Recall: The formula for calculating the recall is given below:

$$Recall = \left[\frac{TP}{TP + FN}\right] * 100 \tag{5.3}$$

where TP represents the true positive, FN symbolizes the false negative.

F1_Score: It is estimated based on the average mean of precision as well as recall. The F1_Score is formulated as given below:

$$F1\_Score = \left[2 * \frac{Precision \ * \ Recall}{Precision \ + \ Recall}\right] * 100$$

Specificity: It is the test's ability to correctly reject healthy samples without a condition. The Specificity is measured as follows:

$$Specificity = \left[\frac{TN}{TN + FP}\right] * 100 \tag{5.4}$$

where TN represents the true negative, FP symbolizes the false positive.

Table 5.1: Comparison of Statistic

Parameters	Statistic		
	Existing DeSVM	**Existing DeLR**	**Proposed DHEGQ-DRLCS**
Accuracy	23.71	18.60	25.45
Balanced accuracy	15.27	17.69	18.98
F1_Score	13.27	2.14	14.47
Precision	13.24	15.70	16.21
Recall	13.12	19.54	20.58
Specificity	24.81	5.72	25.24
TN	25.57	22.21	19.78
TP	20.02	21.52	22.45
FN	23.89	20.74	19.44
FP	25.57	22.22	19.86
ROC_AUC	15.1	17.31	18.91
PRC_AUC	19.47	5.70	20.51
PRC_APS	21.18	9.07	22.22

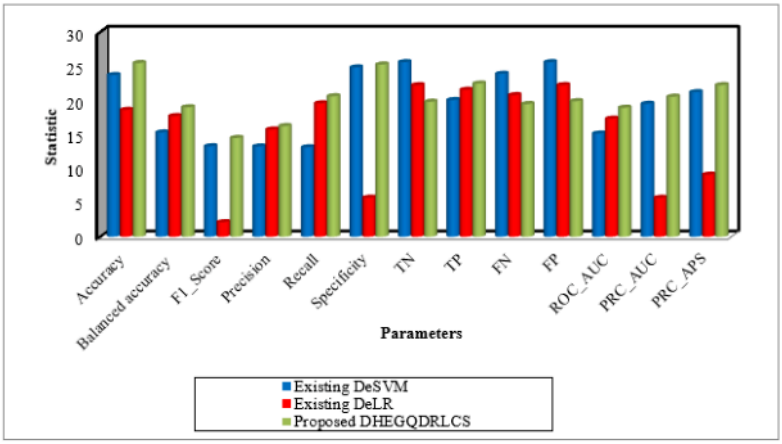


Fig. 5.1: Performance results of statistic

P- value: p-value is the particular statistical measure, such as the mean or standard deviation for the null hypothesis test.

$$P - value = \frac{X - \mu}{\sigma}$$

where X denotes input samples, μ mean and 'σ' denotes a deviation.

The performance analysis of the DHEGQDRLCS technique is conducted and results are depicted in table 5.1 and graph 5.1.

Figure 5.1 depicts the performance results of statistics with respect to a number of parameters such as Accuracy, balanced accuracy, F1-score, precision, recall, specificity, TN, TP, FN, FP, ROC_AUC, PRC_AUC, and PRC_APS. In figure 5.1, three different colors green, blue and red indicate the performance outcomes of the statistic using three methods namely Proposed DHEGQDRLCS, Existing DeSVM, and Existing DeLR respectively. Figure 5.1 clearly shows that performance results of statistics are significantly improved using

Table 5.2: Comparison of the p-value

	p-value		
	Existing DeSVM	**Existing DeLR**	**Proposed DHEGQ-DRLCS**
Accuracy	7.00E-06	0.00	0.00
Balanced accuracy	0.000483	0.00	0.00
F1_Score	0.001312	0.34	0.00010
Precision	0.001333	0.00	0.00010
Recall	0.001417	0.00	0.00
Specificity	4.00E-06	0.06	0.00
TN	3.00E-06	0.00	0.00
TP	4.50E-05	0.00	0.00
FN	6.00E-06	0.00	0.00
FP	3.00E-06	0.00	0.00
ROC_AUC	0.000526	0.00	0.00
PRC_AUC	5.90E-05	0.06	0.00001
PRC_APS	2.50E-05	0.01	0.00001

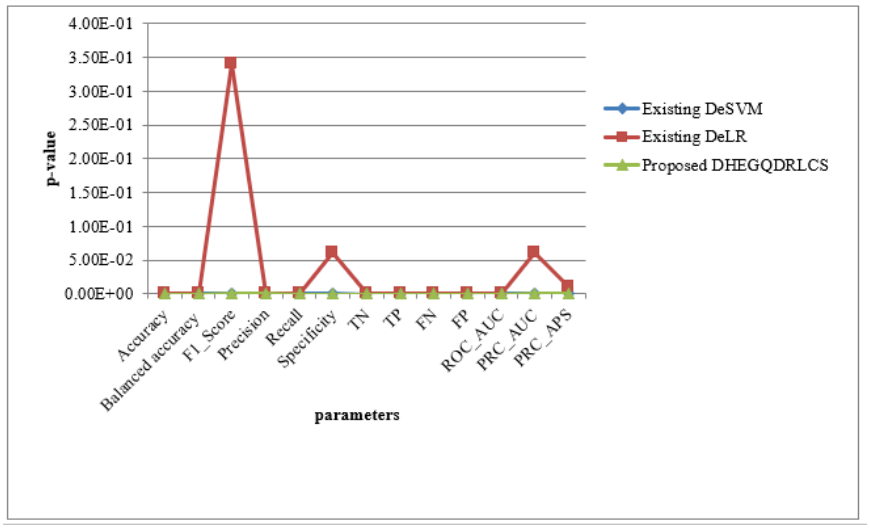


Fig. 5.2: Performance results of the p-value

proposed DHEGQDRLCS when compared to existing DeSVM and DeLR. This is because the DHEGQDRLCS uses the kernel Quadratic discriminant function to analyze the training samples and identify the maximum correlated results by applying the Elitism gradient gene optimization. Based on the optimization results, the accuracy of pancreatic cancer diagnosis is increased by improving the true positive and minimizing the true negative, true positive, and false negative.

The experimental results of p-value using three methods Proposed DHEGQDRLCS, existing DeSVM, and existing DeLR are shown in table 5.2 and Figure 5.6. The p-value is a statistical test that measures the probability of extreme results of the statistical hypothesis test. Among three different methods, the proposed DHEGQDRLCS provides an improved performance than the other two existing methods. As shown in figure 5.6, the DHEGQDRLCS obtains the p-value equal to 0. 000 are lesser than which is less than. 05. Then, the

Table 5.3: Comparison of Mean and Standard deviation using dataset 1

Parameters	Mean			Standard deviation		
	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS
Accuracy	0.77	0.75	0.79	0.02	0.02	0.01
Balanced accuracy	0.68	0.68	0.69	0.02	0.03	0.01
F1_Score	0.42	0.42	0.43	0.03	0.04	0.02
Precision	0.35	0.33	0.37	0.03	0.04	0.02
Recall	0.54	0.57	0.58	0.04	0.05	0.03
Specificity	0.81	0.79	0.85	0.02	0.02	0.01
TN	347.5	338.60	364.20	7.6	8.78	5.63
TP	43.5	45.70	46.90	2.84	3.86	2.33
FN	36.5	34.30	33.10	2.84	3.86	2.21
FP	82.3	91.20	65.60	7.42	8.60	5.58
ROC_AUC	0.71	0.72	0.73	0.03	0.04	0.03
PRC_AUC	0.34	0.37	0.38	0.03	0.02	0.01
PRC_APS	0.34	0.38	0.39	0.03	0.02	0.01

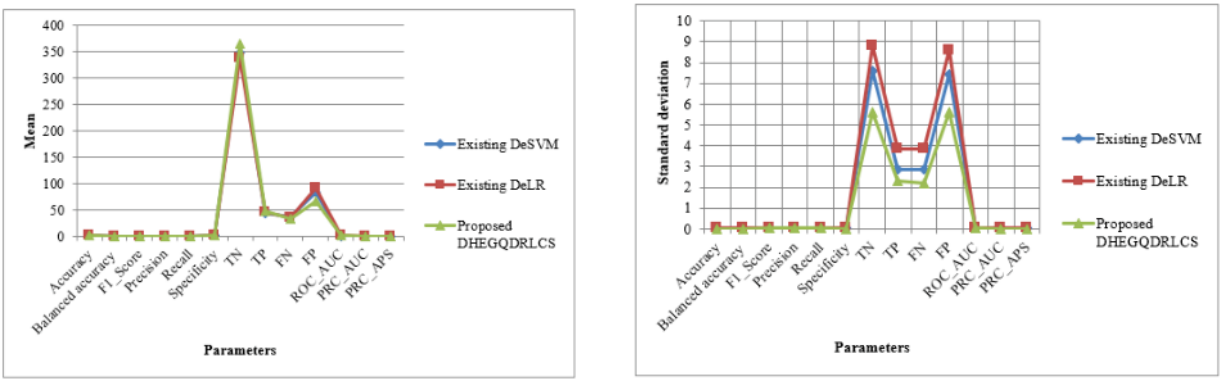


Fig. 5.3: (a) Performance results of mean deviation using dataset 1; (b) Performance results of standard deviation using dataset 1

results are statistically significant.

Table 5.3 and figure 5.3 reveal the experimental results of the mean and standard deviation using dataset 1. The performance of the mean is estimated using three methods DHEGQDRLCS, existing DeSVM, and existing DeLR depending on a number of parameters. From the observed results, it is demonstrated that the proposed DHEGQDRLCS technique provides improved mean value results and minimizes the deviation when compared to existing techniques. Let us consider the input data samples, the mean of the DHEGQDRLCS technique by computing the accuracy is 0.79, and the accuracy of existing DeSVM and DeLR are observed as 0.77 and 0.75 respectively. Similarly, the various parameters are estimated to get final results.

Figure 5.4 illustrates the experimental results of mean and standard deviation with the different numbers of parameters. As shown in the figure, the mean and standard deviation is estimated by using three methods DHEGQDRLCS, existing DeSVM, and existing DeLR. Among the three methods, the performance of mean using DHEGQDRLCS is improved and the standard deviation is minimized when compared to DeSVM and

Table 5.4: Comparison of Mean and Standard deviation using dataset 2

Parameters	Mean			Standard deviation		
	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS
Accuracy	0.79	0.74	0.80	0.01	0.02	0.01
Balanced accuracy	0.59	0.69	0.70	0.04	0.04	0.03
F1_Score	0.3	0.43	0.44	0.07	0.04	0.03
Precision	0.31	0.33	0.36	0.04	0.03	0.02
Recall	0.3	0.62	0.63	0.11	0.06	0.05
Specificity	0.88	0.77	0.89	0.02	0.02	0.01
TN	378	330.40	319.70	9.4	6.77	5.71
TP	24.2	49.20	50.00	8.47	5.12	4.52
FN	55.8	30.80	28.00	8.47	5.12	4.52
FP	51.8	99.40	46.10	9.39	6.48	5.42
ROC_AUC	0.65	0.73	0.74	0.04	0.04	0.03
PRC_AUC	0.28	0.39	0.42	0.04	0.02	0.01
PRC_APS	0.29	0.39	0.43	0.04	0.02	0.01

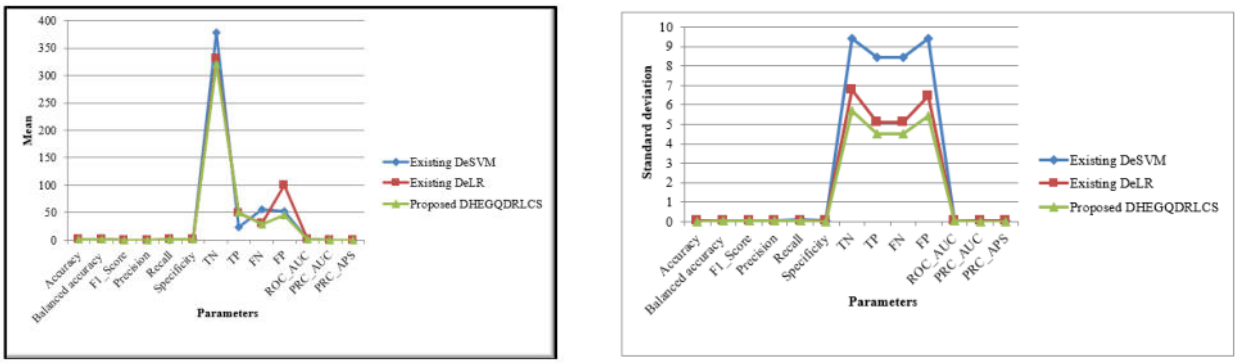


Fig. 5.4: (a) Performance results of mean using dataset 2; (b):Performance results of standard deviation using dataset 2

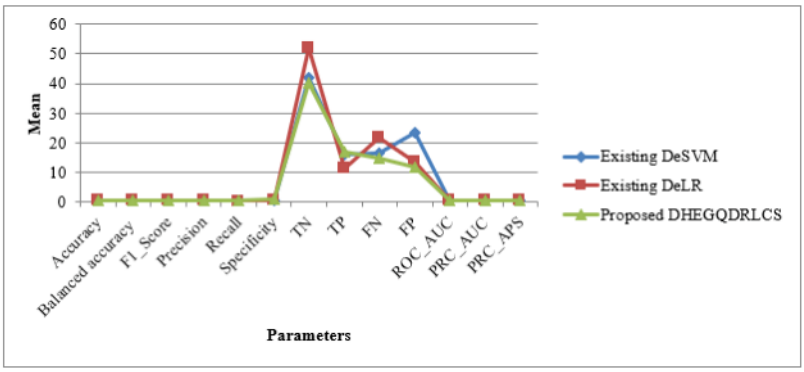
DeLR. This is due to the maximum correlated function used for early identification of pancreatic cancer. The false positive rate is minimized by applying reinforcement learning to accurately detect pancreatic cancer and minimize the loss.

Table 5.5 and figure 5.5 illustrate the performance analysis of the mean and standard deviation using dataset 3 with respect to three different methods DHEGQDRLCS, existing DeSVM, and existing DeLR. The observed results indicate that the DHEGQDRLCS provides better performance when compared to conventional methods. Let us consider the accuracy parameter in the mean estimation. The observed performance of accuracy using DHEGQDRLCS is 0.66 and the accuracy of existing methods is 0.59 and 0.64. The results indicate that the accuracy using DHEGQDRLCS is significantly improved by improving cancer detection and minimizing the loss function.

Table 5.7 indicates that the specificity cutoff with mean for four different stages stage I, stage II, stage III, and stage (IV). The machine learning-based Hadoop distributed eSVM, Distributed eLR, and DHEGQDRLCS

Table 5.5: Comparison of Mean and Standard deviation using dataset 3

Parameters	Mean			Standard deviation		
	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS	Existing DeSVM	Existing DeLR	Proposed DHEGQ-DRLCS
Accuracy	0.59	0.64	0.66	0.07	0.07	0.03
Balanced accuracy	0.57	0.57	0.59	0.06	0.05	0.03
F1_Score	0.44	0.39	0.46	0.07	0.07	0.05
Precision	0.41	0.49	0.52	0.06	0.12	0.05
Recall	0.49	0.34	0.51	0.12	0.11	0.10
Specificity	0.64	0.79	0.82	0.11	0.15	0.07
TN	41.9	51.80	40.20	7.4	10.14	6.44
TP	16.2	11.30	17.00	4.42	3.83	2.67
FN	16.6	21.50	14.80	3.5	3.41	2.34
FP	23.3	13.40	12.00	7.53	9.66	3.80
ROC_AUC	0.58	0.58	0.59	0.07	0.07	0.06
PRC_AUC	0.43	0.45	0.47	0.07	0.08	0.06
PRC_APS	0.45	0.46	0.47	0.07	0.07	0.05



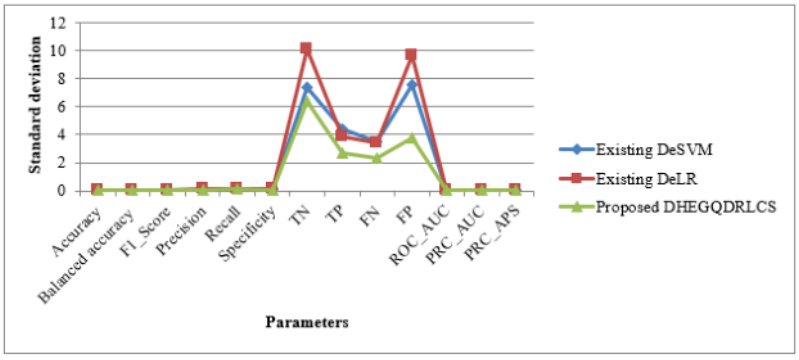


Fig. 5.5: (a) Performance results of mean using dataset 3; (b) Performance results of standard deviation using dataset 3

Table 5.6: Model accuracy with comparisons

Methods		Accuracy	F1_Score	Precision	Recall	TN	TP	FN	FP	ROC_AUC
Existing DeSVM	statistic	23.71	13.27	13.24	13.12	25.57	20.02	23.89	25.57	15.10
	p-value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	avg_mean	0.72	0.39	0.36	0.45	255.80	27.97	36.30	52.47	0.65
	avg_std	0.03	0.06	0.04	0.09	8.14	5.24	4.94	8.11	0.05
Existing DeLR	statistic	18.60	2.14	15.70	19.54	22.21	21.52	20.74	22.22	17.31
	p-value	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	avg_mean	0.71	0.41	0.38	0.51	240.27	35.4	28.87	68	0.68
	avg_std	0.04	0.04	0.05	0.06	8.56	4.27	4.13	8.25	0.05
Proposed DHEGQ-DRLCS	statistic	25.45	14.47	16.21	20.58	19.78	22.45	19.44	19.86	18.91
	p-value	0.00	0.11	0.03	0.00	0.00	0.00	0.00	0.00	0.00
	avg_mean	0.75	0.44	0.41	0.57	238.36	37.96	25.3	41.23	0.68
	avg_std	0.01	0.03	0.03	0.05	5.92	3.17	3.02	4.93	0.04

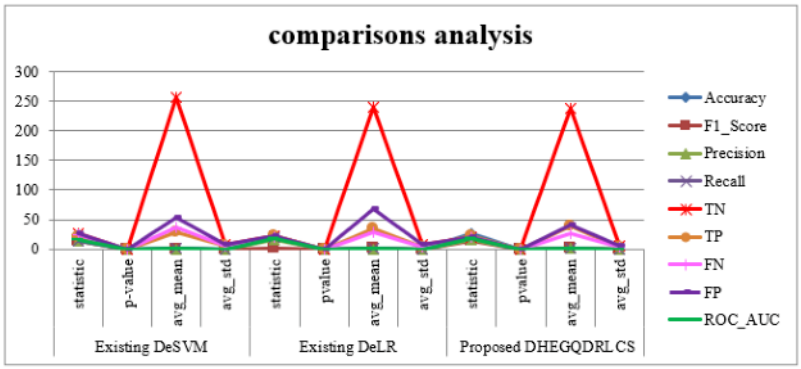


Fig. 5.6: Model accuracy with comparisons

are applied for pancreatic cancer detection. The experiments demonstrated that the proposed DHEGQDRLCS improves the performance of the classifier for early identification of pancreatic cancer with stage values ranging from 0.6 to 0.10 and a mean specificity cutoff. The algorithms in the Hadoop distributed file system of DHEGQDRLCS provide better accuracy than other existing approaches in this comparison research.

5.1. Advantages and Limitation of DHEGQDRLC. The architecture of DHEGQDRLCS is used to increase the accuracy of early pancreatic cancer diagnosis based on the hybridization of three techniques namely kernel quadratic discriminant function, Elitism gradient gene optimization, and reinforcement learning algorithm. The hybridization process minimizes the loss function of disease diagnosis and increases the classification results. The limitation of the proposed DHEGQDRLCS is not having the ability to analyze time consumption and space complexity in the present state when applying the large volume of the dataset.

6. Conclusion. In this section, early detection of pancreatic cancer is very significant before cancer swells to other organs in the body. However, early detection of pancreatic cancer is difficult because this cancer has non-specific symptoms. The conventional distributed eSVM, Distributed eLR approaches perform pancreatic

Table 5.7: Cutoff values with a mean of proposed models

specificity cutoff WITH MEAN	**stages**
0.6 to 0.7	I
0.7 to 0.8	II
0.8 to 0.9	II
0.9 to 0.10	IV

cancer of pancreatic cancer. But the higher accuracy and loss rate was not minimized. Therefore, a novel DHEGQDRLCS is developed to increase the classifier's performance for early pancreatic cancer diagnosis with four different stages based on p-value and statics, according to preliminary data samples. In the proposed DHEGQDRLCS, the kernel quadratic discriminant function analyzes the testing and training data samples using the kernel function. Then the Elitism gradient gene optimization is applied to provide the final disease diagnosis results. In order to minimize the loss function of disease diagnosis, the reinforcement learning algorithm is applied to find the optimal learning rate. Finally, the hybridized classifier provides precise classification results with a minimum loss function. Based on the classification results, pancreatic cancer is correctly identified. A comprehensive experimental evaluation is carried out using three datasets with different parameters such as Accuracy, balanced accuracy, F1-score, precision, recall, specificity, $TN, TP, FN, FP, ROC_{AUC}, PRC_{AUC}$, and PRC_{APS}. The quantitative performance result indicates that the presented DHEGQDRLCS achieves higher accuracy in a pancreatic cancer diagnosis than the conventional methods. The proposed future work could increase the number of applications for pancreatic disease detection and also improve accuracy.

REFERENCES

[1] Stephen P Pereira, Lucy Oldfield, Alexander Ney, Phil A Hart et al. "Early detection of pancreatic cancer" , The Lancet Gastroenterology & Hepatology, 2020

[2] Santosh Reddy P, Chandrasekar M. "PAD: A Pancreatic Cancer Detection based on Extracted Medical Data through Ensemble Methods in Machine Learning" , International Journal of Advanced Computer Science and Applications, 2022

[3] Md Manjurul Ahsan, ShahanaAkter Luna, Zahed Siddique. "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review" , Healthcare, 2022

[4] Stefanus Tao Hwa Kieu, Abdullah Bade, Mohd Hanafi Ahmad Hijazi, HoshangKolivand. "A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions" , Journal of Imaging, 2020

[5] Rahaman, M.M.;Li, C.;Yao, Y.;Kulwa, F.;Rahman, M.A.;Wang, Q.;Qi, S.;Kong, F.;Zhu, X.;Zhao, X.Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transferlearningapproaches.J.X-RaySci.Technol. 2020, 28,821–839.

[6] Ma,J.;Song, Y.;Tian,X.; Hua,Y. Zhang, R.;Wu, J.Survey on deep learning for pulmonary medicalimaging.Front.Med. 2019,14,450–469.

[7] Ardila,D.;Kiraly,A.P.;Bharadwaj,S.;Choi,B.;Reicher,J.J.;Peng,L.;Tse, D.; Etemadi, M.; Ye, W. Corrado, G.; et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography.Nat.Med. 2019, 25,954–961.[CrossRef]

[8] Kieu, S.T.H.;Hijazi, M.H.A.;Bade, A.;Yaakob, R.;Jeffree, S. Ensemble deep learning for tuberculosis detection using chest X-Ray and canny edge detected images. IAES Int. J. Artif. Intell.**2019**, 8, 429–435.

[9] Ayan,E.;Ünver,H.M.Diagnosis of Pneumonia from Chest XRay Images using DeepLearning .Sci.Meet.Electr.-Electron.Biomed.Eng.Comput.Sci. 2019.

[10] Salman, F.M.; Abu-naser, S.S.; Alajrami, E.; Abu-nasser, B.S.; Ashqar, B.A.M.COVID-19 Detection using Artificial Intelligence Int.J.Acad.Eng.Res. 2020,4,18–25.

[11] Gao, X.W.; James-reynolds, C.; Currie, E.Analysis of tuberculosis severity levels from CT pulmonary images based on enhanced residual deep learning architecture .Neurocomputing 2019,392,233–244.[CrossRef]

[12] Gozes,O.;Frid,M.;Greenspan,H.;Patrick,D. Rapid AIDevelopment Cycle for the Coronavirs (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis Article . arXiv 2020 ,arXiv : 2003 .05037.

[13] Mithra, K.S.; Emmanuel, W.R.S. Automated identification of mycobacterium bacillus from sputum images for tuberculosis diagnosis. Signal Image Video Process. 2019.

[14] Samuel, R.D.J.;Kanna, B.R.Tuberculosis (TB) detection system using deep neural networks. Neural Comput.Appl. 2019, 31,1533–1545.

[15] O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.;Walsh, J. Deep

Learning vs .Traditional Computer Vision.Adv.Intell.Syst.Comput.**2020**, 128–144.
[16] Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop,Swinoujscie,Poland.
[17] Shorten,C.;Khoshgoftaar,T.M.A survey on Image Data Augmentation for DeepLearning . J.BigData 2019,6.
[18] Ker, J.; Wang, L. Deep Learning Applications in Medical Image Analysis. IEEE Access 2018, 6, 9375–9389.
[19] Wang, C.; Chen, D.; Hao, L.; Liu, X.; Zeng, Y.; Chen, J.; Zhang, G. Pulmonary Image Classificatin Based on Inception-v3 Transfer Learning Model .IEEE Access 2019 ,7, 146533 –146541
[20] Sadewo W, Rustam Z, Hamidah H and Chusmarsyah A.R, "Pancreatic Cancer Early Detection Using Twin Support Vector Machine Based on Kernel" Symmetry, 2020, 12, 1-8.
[21] Jacobson S, Dahlqvist P, Johansson M, Svensson J, Billing O, Sund M, Franklin O, "Hyperglycemia as a risk factor in pancreatic cancer: A nested casecontrol study using prediagnostic blood glucose levels", Pancreatology, Elsevier, 2021, 21, 1112-1118
[22] Jacobson S, Dahlqvist P, Johansson M, Svensson J, Billing O, Sund M, Franklin O, Alves N, Schuurmans M, Litjens G , S. Bosma J, Hermans .J and Huisman H., "Fully Automatic Deep Learning Framework for Pancreatic Ductal Adenocarcinoma Detection on Computed Tomography. Cancers, 2020, 14(2), 1-10.
[23] Ju J, V. Wismans L, A.M. Mustafa D, J.T. Reinders M, H.J. van Eijck C, A. P. Stubbs A, and Li Y, "Robust deep learning model for prognostic stratification of pancreatic ductal adenocarcinoma patients", iScience, Elsevier, 2021, 24 (12), 1-18
[24] Yana Z, Ma C, Mo J, Han W, Lv X, Chen C, Chen C, Nie X, "Rapid identification of benign and malignant pancreatic tumors using serum Raman spectroscopy combined with classification algorithms", Optik, Elsevier, 2020, 208, 1-5
[25] N. Al-Shaheri F, S.S. Alhamdani M, S. Bauer A, Nathalia Giese, Markus W. Büchler, Thilo Hackert, J rg D. Hoheisel, "Blood biomarkers for differential diagnosis and early detection of pancreatic cancer", Cancer Treatment Reviews, Elsevier, 2021, 96, 1-16
[26] Park J, Artin M G, Lee K E., Pumpalova Y S, Ingram M A, May B L, Park M, Hur C, Tatonetti N P., "Deep learning on time series laboratory test results from electronic health records for early detection of pancreatic cancer", Journal of Biomedical Informatics, Elsevier, 2022, 131, 1-11
[27] Savareh B A, Aghdaie H A, Behmanesh A, Bashiri A, Sadeghi A, Zali M, Shams R, "A machine learning approach identified a diagnostic model for pancreatic cancer through using circulating microRNA signatures", Pancreatology, Elsevier, 2020, 20, 1195-1204

Edited by: Vinoth Kumar
Received: Jun 16, 2022
Accepted: Oct 19, 2022