JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Feature Selection of Brain Tumor based on Gene Expression using Bootstrapping Spider Wasp Optimization

[1] Y.Tezaaw, [2] Dr.K.Vijaya Lakshmi
[1] Research Scholar, [2] Professor
[1, 2] Dept.of Computer Science,
S.V.University, Tirupati, A.P,India
[1] tezaawyavanaboina@gmail.com, [2] vijayalakshmik4@gmail.com

Abstract : Determining early-stage Brain tumor through gene expression data offers significant challenges for effective treatment. Gene expression datasets typically consist of a multitude of features, each representing specific genes. However, the presence of irrelevant or redundant features often leads to multicollinearity issues, complicating analysis and decision-making processes. To overcome these problems, this research proposes the Bootstrapping Guidance Strategy based on the Spider Wasp Optimization (BGS-SWO) approach for the feature selection of brain cancer gene expression. Initially, the two standard datasets as Curated Microarray Database (CuMiDa) and the Cancer Genome Atlas (TCGA) are used for the estimation of the proposed method. Pre-processing techniques like handling missing values and min-max normalization are performed to enhance the quality of the input data. Then, the BGS-SWO approach is used for the selection of important features and these are classified by proposing the Contextual Bayesian Optimization based Support Vector Machine (CBO-SVM) approach. The experimental results demonstrate that the proposed BGS-SWO with CBO-SVM approach attains a better accuracy of 99.35%, precision of 98.34%, recall of 99.10% and F1-score of 98.71% on CuMiDa dataset as compared to Depth-wise Separable Convolutional Neural Network (DSCNN) method.

IndexTerms **- Bootstrapping Guidance Strategy, Gene Expression Data , Contextual Bayesian Optimization, Spider Wasp Optimization, Support Vector Machine..**

1.INTRODUCTION

Gliomas are the utmost general prime Central Nervous System (CNS), exhibiting the highest incidence and mortality rates among tumors in both paediatric and adult populations [1]. Gliomas are categorized into major types, including glioblastomas, mixed gliomas, oligodendrogliomas and astrocytomas [2]. World Health Organization (WHO) partitions the gliomas into different types: Grade I (pilocytic astrocytomas), grade II (difuse astrocytomas), grade III (anaplastic astrocytomas), and grade IV (glioblastomas), with increasing levels of malignancy and aggressiveness [3]. Glioblastoma Multiforme (GBM) is the most aggressive and severe form of glioma, with a median survival of 12 to 15 months following diagnosis [4]. It accounts for 15% of primary brain tumors, 46% of malignant brain tumors, and 60–75% of astrocytomas [5]. Early detection of cancer cell growth is crucial to improving survival rates. However, the development of new therapies for gliomas is significantly hindered by their high heterogeneity across molecular subgroups [6]. Histology itself is inappropriate for providing precise diagnosis and prognostic delamination required for successful treatment. Integration of histological evaluation through molecular markers enables the most precise patient diagnosis as well as hazard delamination [7]. Nevertheless, the most important prognostic molecular markers are specially required for enhancing the diagnosis and long-term life quality of the patients through long-winded gliomas [8].

The present classification of gliomas incorporates both genomic alterations and histopathological features, providing insights into tumor aggressiveness and patient diagnosis [9]. For lower-grade gliomas (Grades II and III), the mutation status of isocitrate dehydrogenase 1 or 2 (IDH1/2) genes and 1p/19q codeletion are key molecular characteristics used for classification [10]. Microarray provides genome-wide patterns in gene expression, enhancing the clinical utilization of this advancement for illness diagnosis, prognosis and treatment [11]. The identification and classification of gene expression using microarray data is an efficient tool for cancer diagnosis and prognosis for particular cancer subtypes [12]. Artificial Intelligence (AI)-based learning methods are essential tools widely used to extract significant features from gene expression data, playing a crucial role in gene classification [13]. In cancer classification, it is important for the explicit identification of the features due to large data dimensionality [14]. In modern days, Machine Learning (ML)-based approaches have obtained admiration in distinguishing cancer cells because of the significance of the identification of important features as well as the applicability of the data [15].

The primary highlights of this paper are given as follows:

Pre-processing techniques like handling missing values and Min-max normalization are used for improving the quality of the input features of CuMiDa and TCGA. This leads to more reliable, accurate and interpretable results in ML tasks.

The Bootstrapping Guidance Strategy based Spider Wasp Optimization (BGS-SWO) approach is proposed for the selection of important features from the brain cancer gene expression. The BGS-SWO crucially minimizes the dimensionality through the selection of the most informative features, resulting in better accuracy results.

Contextual Bayesian Optimization (CBO) algorithm is utilized optimally tuning the hyperparameters of the Support Vector Machine (SVM), commonly called the CBO-SVM approach. The CBO-SVM approach is utilized for the categorization of Brain tumor gene expression.

This research paper is given as follows: Section 2 describes a literature survey. Section 3 demonstrates a proposed methodology. Section 4 illustrates the results and discussion and Section 5 provides a conclusion of this research.

2.LITERATURE SURVEY

In this section, the existing works related to ML and DL based brain cancer gene expression are discussed, along with their advantages and limitations.

Heba M. Afify et al. [16] introduced the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for various categories of brain cancer gene expression. The 1-dimensional CNN (1D-CNN) approach pursued through the RNN approach with and without Hyperparameter tuning approach of the Bayesian Optimization (BO) approach which is commonly called BO+1D-CNN+RNN was also introduced for the classification process. A classification based on an introduced hybrid approach provided the most precise and effective analysis for patients with various classes of brain cancer. However, BO supported to optimize hyperparameters, and the performance of BO+1D-CNN+RNN hybrid was highly sensitive to initial choices such as the kernel size in CNN.

Anju Das et al. [17] developed the DL approach of the Depth-wise Separable CNN (DSCNN) approach for categorization of large cancerous and non-cancerous classes. In the pre-processing step, data augmentation was utilized to enhance data size and Min-max normalization was utilized for normalize the input brain cancer gene expressions. An Enhanced Optimization (ECO) approach was utilized for the selection of important features while the elimination of irrelevant and redundant features. The effectiveness of the developed DSCNN approach was significantly enhanced through the resolution of dimensionality as well as overfitting issues. However, data augmentation introduced synthetic artifacts that do not accurately model the biological variations potentially minimized the model's ability to generalize effectively.

Amena Mahmoud and Eiko Takaoka [18] presented the various ML approaches for the diagnosis of liver cancer, integrated feature selection approaches with the stacking ensemble learning approach. The feature selection procedure was performed for the identification of the most appropriate gene expressions integrated with liver cancer. The selected features were utilized further to train the stacking ensemble approach, integrated various individual approaches like Multi-Layer Perceptron (MLP), Random Forest (RF), K-Nearest Neighbour (KNN) as well as SVM through meta-learner Extreme Gradient Boosting (XGBoost) approach to predictions. However, the feature selection inadvertently discarded the features that were weakly correlated with the cancer.

Rabia Emhamed Al Mamlook et al. [19] proposed to apply ML methods to the classification of GBM instances from The Cancer Genome Atlas (TCGA) data. Various supervised learning approaches like SVM, Gradient Boosting (GB), KNN, Logistic Regression (LR) and Decision Tree (DT) were performed in an analysis. This finding demonstrated that the DT and GB approaches were the most efficient for this classification task as well as attained better accuracy results. However, the presented ML approach does not inherently provide insights into the underlying biological mechanisms or gene interactions associated with GBM.

Amol Avinash Joshi and Rabia Musheer Aziz [20] introduced the novel optimization algorithm of Particle Swarm Optimization (PSO) through Cuckoo search (CS) commonly called PSCS. Then, the classified gene expression data of brain tumour through the DL approach was utilized for the identification of various groups or classes based on the specific tumour along through the PSCS optimization approach. An introduced PSCS approach with DL attained effective performance. However, the hybrid PSCS optimization algorithm combined with deep learning, lacks interpretability.

From this overall analysis, some of the limitations have been identified: Highly sensitive to the model, affected the model's generalization ability, lack of interpretability and inadvertently discarded the features. Hence, to overcome these problems, this research proposes the BGS-SWO approach for the optimal selection of features from the brain cancer gene expression. Then, the CBO approach is performed to effectively tune the hyperparameters of the SVM approach.

3.METHODOLOGY

This research's aim is to the effective and precise classification of Brain tumor gene expressions by proposing a feature selection algorithm based on the BGS-SWO. This research comprises four important steps data acquisition using Curated Microarray Database (CuMiDa) and The Cancer Genome Atlas (TCGA), pre-processing, feature selection through BGS-SWO and classification by CBO-SVM approach. Figure 1 signifies a systematic diagram of the proposed CBO-SVM.

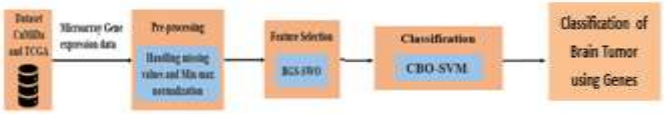

Figure 1: Systematic Diagram of a Proposed Method

3.1 Data Acquisition:

The primary phase of this research is the data acquisition to validate the importance of the BGS-SWO with the CBO-SVM approach. The two standard datasets named CuMiDa [21] and TCGA [22] are considered in this article for the evaluation of the proposed method. The detailed information of these data is described below.

3.1.1 CuMiDa Dataset:

CuMiDa is an open-source gene expression dataset that involves large-scale tumor microarray datasets curated from 30,000 Gene Expression Omnibus (GEO) educations. As demonstrated in CuMiDa, various classes were removed due to their insufficient representation and failure to meet the standards required for reliable ML analysis. The data set is demonstrated in a CSV file through 130 cells (samples) in rows and 54678 features (genes) in columns.

3.1.2 TCGA Dataset:

TCGA is a standard data platform for tumor genome schemes. It gives experimental as well as pathological information on 33 types of tumor which are simple to acquire. TCGA dataset was utilized to identify clinical data and high-throughput RNA sequencing (RNA-seq) data from patients with glioma. Then, the Fragments Per Kilobase Mer million (FPKM) involved in HTseq were utilized for the identification of SLC31A1 transcript expression levels. Moreover, for the subsequent analysis of RNA-Seq gene expression level 3 HTSeqFPKM data of 703 patients through glioma, clinical data are renovated into Transcripts Per Million (TPM) recites.

3.2 Pre-processing:

Here, acquired data is utilized as income to improve the quality of a dataset. Pre-processing of raw data is important for accurately classifying a brain tumor gene expression. In this research, pre-processing techniques like handling missing values and Min-max normalization are performed.

3.3 Feature Selection using Spider Wasp Optimization:

When analysing the microarray data, selecting the important genes based on the gene expression data poses a significant challenge. SWO [23] is a meta heuristic optimization algorithm which is encouraged through the behaviours of the hunting, nesting and important parasitism of offspring of particular wasp species. The SWO algorithm involves hunting replication, nesting and mating behaviour of female spider wasps in nature. A mathematical description of the SWO approach is provided below.

3.3.1 Population Initialization:

In SWO, every spider wasp is a candidate approach for an optimization issue which is formulated in equation (2) as follows:

$$\overrightarrow{SW} = [x_1, x_2, x_3, \dots, x_D,] \quad (2)$$

Where, D demonstrates the dimension; $\overrightarrow{SW}$ signifies an individual female wasp. A cohort of spider wasp population is initialized through equation (3) as follows:

$$\overrightarrow{SW_i} = \vec{L} + \vec{r} \times \left(\vec{H} - \vec{L}\right) \quad (3)$$

Where, i signifies an individual's index; $\vec{L}$ and $\vec{H}$ depicts the lower bound and upper bound of problem solution; $\vec{r}$ denotes a randomly produced D dimensional vector.

3.3.2 Exploration and Exploitation:

This phase imitates how female wasps identify more suitable spiders to forage their children. A female wasp utilizes the continuous phase to arbitrarily look for a search space concerning tracing a spider that is appropriate for their offspring. This searching process is implemented in equations (4) to (8) as follows:

$$\overrightarrow{SW_i}^{t+1} = \overrightarrow{SW_i}^{t} + \mu_1 + \left(\overrightarrow{SW_a}^{t} - \overrightarrow{SW_b}^{t}\right) \quad (4)$$

$$\mu_1 = |rn| * r_1 \quad (5)$$

$$\overrightarrow{SW_i}^{t+1} = \overrightarrow{SW_c}^{t} + \mu_2 * \left(\vec{L} + \vec{r}_2 * \left(\vec{H} - \vec{L}\right)\right) \quad (6)$$

$$\mu_2 = B * \cos(2\pi l) \quad (7)$$

$$B = 1/(1 + e^{l}) \quad (8)$$

Where, r_1 and r_2 denotes the consistently distributed random variables in [0,1]. t signifies the present iteration number. a, b and c denotes the metrices of the intentionally selected wasps. μ_1 is utilized for an identification of the wasp movement's direction; l depicts the randomly produced from -1 to 2 respectively. rn signifies the random number produced through an actual distribution. In pursuing the escaping phase, capturing wasp behaviour is formulated in equations (9) to (11) as follows:

$$\overrightarrow{SW_i}^{t+1} = \begin{cases} \overrightarrow{SW_i}^{t} + C * \left|2 * \vec{r}_5 * \overrightarrow{SW_a}^{t} - \overrightarrow{SW_i}^{t}\right| & r_3 < r_4 \\ \overrightarrow{SW_i}^{t} * \overrightarrow{vc} & otherwise \end{cases} \quad (9)$$

$$C = \left(2 - 2 * (t/t_{max})\right) * r_6 \quad (10)$$

$$k = 1 - (t/t_{max}) \quad (11)$$

Where, r_3, r_4 and r_6 denote the random numbers in the range between 0 and 1; $\vec{r}_5$ denotes the vector where every element is a random number; t_{max} depicts the maximum number of iterations; $\overrightarrow{vc}$ signifies vector generated among $-k$ and k based on an actual distribution. During a nesting phase, a female wasp forces a paralysed spider into the pre-prepared nest. A nesting attitude is formulated in equations (12) to (14) as follows:

$$\overrightarrow{SW_i}^{t+1} = \overrightarrow{SW^*} + \cos(2\pi l) * \left(\overrightarrow{SW^*} - \overrightarrow{SW_i}^{t}\right) \quad (12)$$

$$\overrightarrow{SW_i}^{t+1} = \overrightarrow{SW_a}^{t} + r_3 * |\gamma| * \left(\overrightarrow{SW_a}^{t} - \overrightarrow{SW_1}^{t}\right) + (1 - r_3) * \vec{U} * \left(\overrightarrow{SW_b}^{t} - \overrightarrow{SW_c}^{t}\right) \quad (13)$$

$$\vec{U} = \begin{cases} 1 \; \vec{r}_7 > \vec{r}_8 \\ 0 \; otherwise, \end{cases} \quad (14)$$

Where, $\overrightarrow{SW^*}$ signifies the present wasp through best fitness; $\vec{U}$ denotes the binary vector; $\vec{r}_7$ and $\vec{r}_8$ denotes the two random vectors where every component is an arbitrary number in a range between 0 and 1; γ denotes the number generated based on levy flight.

3.3.3 Mating Behaviour:

In SWO, each spider wasp symbolizes the possible resolution into the current generation, while spider wasp eggs stand for newly created potential solutions into generation. The new solution is generated by using equation (15) to (18) as follows:

$$\overrightarrow{SW_i^{t+1}} = Crossover\left(\overrightarrow{SW_i^t}, \overrightarrow{SW_f^t}, CR\right) \quad (15)$$

$$\overrightarrow{SW_f^{t+1}} = \overrightarrow{SW_i} + e^l * |\beta| + \overrightarrow{v_1} + (1 - e^l) * |\beta_1| * \overrightarrow{v_2} \quad (16)$$

$$\overrightarrow{v_1} = \begin{cases} \overrightarrow{x_a} - \overrightarrow{x_\iota} & f(\overrightarrow{x_a}) < f(\overrightarrow{x_\iota}) \\ \overrightarrow{x_\iota} - \overrightarrow{x_a} & otherwise \end{cases} \quad (17)$$

$$\overrightarrow{v_1} = \begin{cases} \overrightarrow{x_b} - \overrightarrow{x_c} & f(\overrightarrow{x_a}) < f(\overrightarrow{x_\iota}) \\ \overrightarrow{x_c} - \overrightarrow{x_b} & otherwise, \end{cases} \quad (18)$$

Where, $\overrightarrow{SW_i^t}$ signifies the female wasps and $\overrightarrow{SW_f^t}$ signifies the newly generated male wasps; $Crossover$ denotes the crossover tasks that happen with the probability CR. β and β_1 depicts the randomly produced based on an actual distribution.

3.3.4 Population minimization and memory saving:

As a population iterates, some individuals persist for minimize, which is formulated in equation (19). For every individual, if a wasp's performance in a subsequent generation surpasses in the current generation, the unchanged individual will continue into the following generation:

$$N = N_{min} + (N - N_{min}) \times k \quad (19)$$

Where, N_{min} signifies a minimum population size which solves local optima. However, a present global optimal solution is in a local optimal trap, a large number of individuals are directed into a trap, which improves the problem of local optima. Hence, to overcome this, the BGS approach is incorporated into the conventional SWO approach. The detail of BGS is provided in the following.

3.3.5 Bootstrapping Guidance Strategy

In nesting behaviour, Equation (7) directs a candidate solution toward globally optimal individuals, which helps enhance the convergence speed of the approach toward a global optimum. To solve the local optima problem, the BGS is proposed in the traditional SWO approach. The BGS is formulated in equation (20) as follows:

$$\overrightarrow{SW_i^{t+1}} = \overrightarrow{SW_{better}^t} + \cos(2\pi l) \,.\, (SW_{better}^t - SW_i^t) \quad (20)$$

Where, $\overrightarrow{SW_{better}^t}$ signifies the better individual in t iteration. Bootstrapping with a better directing scheme ensures a convergence speed of an approach that will enhance its capability to jump out of local traps.

3.4 Classification using Support Vector Machine:

The utilization of high-dimensional gene expression data to examine the state of a cell has significantly facilitated the application of machine learning in cancer research. SVM [24] is a mathematical approach that is mostly used to solve classification and regression problems. The SVM is mostly related for structural risk reduction idea as contrasting to an experiential error minimization idea in Neural Networks (NN). SVM significantly employs linear as well as non-linear classification. This is performed through the projection of the training dataset to higher dimensional space where an identified hyperplane separates the classes of training data. Hence, the purpose of the SVM learning procedure is to pursue optimal linear support vectors in that dimension.

3.4.1 Contextual Bayesian Optimization:

CBO allows the optimization functions based on an addition and external variables which are known as the contexts. A concept involves a functional dependency on a context in the Gaussian Process (GP) process, however, to deliberate them fixed when selecting further parameters to estimate.

For instance, provided a context $z \in \mathcal{Z}$ defined by a background, this research models how the effectiveness and constraint functions change across various contexts. This is achieved by extending a kernel function k_a over parameters, along with different kernel $k_z: \mathcal{Z} \times \mathcal{Z} \to R$ for the contexts, as formulated in equation (21) as follows:

$$k\left((a,z),(a',z')\right) = k_a\left((a,z),(a',z')\right) k_z\left((a,z),(a',z')\right) \quad (21)$$

The kernel structure suggests that function values are connected when both parameters and contexts are equal. For instance, selecting similar parameters "a" to a control approach leads the similar performance values if the contexts are similar.

Since the contexts lack an optimization condition, an improved type are utilized. This approach estimates the GP criterion provided by the fixed context z_n, which is formulated in equation (22) as follows:

$$a_n = \underset{a \in A}{\operatorname{argmin}}\, \mu_{n-1}(a, z_n) + \beta_n^{1/2} \sigma_{n-1}(a, z_n) \quad (22)$$

Particularly, after observing a particular context sufficiently, Equation (22) will query parameters that are close to the optimal solution.

4. EXPERIMENTAL RESULTS

The importance of the proposed BGS-SWO with CBO-SVM approach is estimated by using different performance metrics such as accuracy, precision, recall and F1-score.

4.1: Performance Analysis:

The effectiveness of the proposed BGS-SWO with CBO-SVM approach method is estimated with existing methods through two different datasets like CuMiDa and TCGA respectively.

Table 1 demonstrates the performance analysis of feature selection approaches based on CuMiDa dataset. The significance of the proposed BGS-SWO method is estimated and compared through other optimization approaches like Ant Colony Optimization (ACO), Pelican Optimization Algorithm (POA), Golden Jackle Optimization (GJO) and SWO. The proposed BGS-SWO approach attains a better accuracy of 99.35%, a precision of 98.34%, a recall of 99.10% and an F1-score of 98.71% on the CuMiDa dataset respectively.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ACO	91.44	91.25	93.78	92.49
POA	92.19	93.28	94.12	93.69
GJO	94.28	95.29	95.32	95.30
SWO	96.28	96.39	97.56	96.97
BGS-SWO	99.35	98.34	99.10	98.71

Table 1: Performance analysis of feature selection methods using CuMiDa dataset

The performance analysis of feature selection approaches based on the TCGA dataset is represented in Table 2. The significance of the proposed BGS-SWO method is estimated and compared with the other optimization approaches like ACO, POA, GJO and SWO. The proposed BGS-SWO approach attains a better accuracy of 99.01%, precision of 99.12%, and recall of 99.13% and 99.12% respectively based on the TCGA dataset.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ACO	91.48	91.12	91.38	91.24
POA	92.13	93.76	93.19	93.47
GJO	94.28	95.29	95.21	95.24
SWO	97.92	98.43	96.29	97.34
BGS-SWO	99.01	99.12	99.13	99.12

Table 2: Performance analysis of feature selection methods using TCGA dataset

Figure 2 demonstrates the performance analysis of classification approaches based on CuMiDa dataset. The significance of the proposed CBO-SVM method is estimated and compared through the other ML approaches like DT, RF, KNN and SVM. The proposed CBO-SVM approach attains a better accuracy of 99.35%, precision of 98.34%, recall of 99.10% and F1-score of 98.71% on CuMiDa dataset respectively.

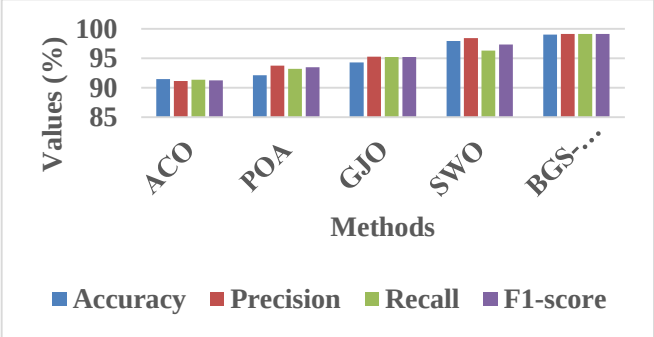

Figure 2: Performance analysis of classification methods using CuMiDa dataset

Figure 3 demonstrates the performance analysis of feature selection methods based on TCGA dataset. The significance of the proposed CBO-SVM approach is estimated and compared through the other ML approaches like DT, RF, KNN and SVM. The proposed CBO-SVM approach attains a better accuracy of 99.01%, precision of 99.12%, and recall of 99.13% and 99.12% respectively based on the TCGA dataset.

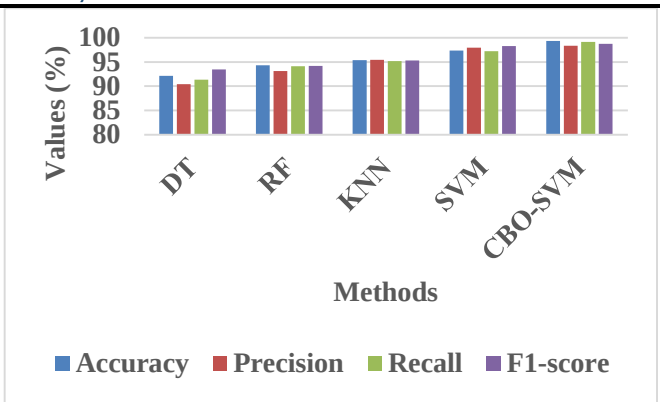

Figure 3: Performance analysis of classification methods using TCGA dataset

4.2 Comparative Analysis:

In this section, the effectiveness of the proposed BGS-SWO with CBO-SVM is estimated and compared with the existing methods based on a CuMiDa dataset. Table 3 signifies a comparative analysis of the proposed method with the existing method. An existing approach like 1D-CNN+RNN without BO [16] and DSCNN [17] are estimated and compared with the proposed BGS-SWO with CBO-SVM approach using various performance metrics based on CuMiDa dataset.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1D-CNN+RNN without BO [16]	90	86	91.52	86.16
DSCNN [17]	99.20	98.10	98.40	98.30
Proposed BGS-SWO with CBO-SVM	99.35	98.34	99.10	98.71

Table 3: Comparative Analysis of proposed with existing methods based on CuMiDa dataset

4.2 Discussion:

In this section, the advantages of the proposed method and the limitations of the existing works are discussed. The limitations of the existing works are: 1D-CNN+RNN without BO approach [16] performance was highly sensitive to the kernel size in CNN, resulting in poor classification results. DSCNN [17] approach with data augmentation introduced the synthetic artifacts which do not faithfully represent real biological variations, potentially impacting the model's generalization ability. In [18], the feature selection inadvertently discarded the features which were weakly correlated with the tumor. To overcome these problems, this research proposes the BGS-SWO approach for the selection of important features for brain cancer gene expression. Gene expression datasets are typically high-dimensional with many irrelevant or redundant features. BGS-SWO effectively reduces dimensionality by selecting only the most informative features, which simplifies the model and speeds up computations. The CBO-SVM approach is utilized for the classification of gene expressions.

5.CONCLUSION

The proposed BGS-SWO approach combined with the CBO-SVM, effectively addresses the challenges associated with brain tumor identification using gene expression data. The research focused on optimizing feature selection by reducing redundancy and mitigating multicollinearity issues in gene expression datasets, which are often characterized by a high dimensionality of features. Primarily, the CuMiDa and TCGA datasets were utilized to evaluate the proposed method. Data pre-processing techniques such as handling missing values and min-max normalization are performed to improve the quality of the datasets, ensuring more reliable and accurate results. The selected features were classified through the CBO-SVM method, which leveraged CBO to fine-tune SVM hyperparameters and improve classification accuracy. The experimental results demonstrate that the proposed BGS-SWO with CBO-SVM approach attains a better accuracy of 99.35% and precision of 98.34% on CuMiDa dataset respectively as compared to DSCNN. The future work will involve the Deep Learning (DL) approach to enhance the overall model performance.

REFERENCES

- Zhang, Z., Xiao, Y., Zhao, S., Liu, J., Zeng, J., Xiao, F., Liao, B., Shan, X., Zhu, H. and Guo, H., 2024. FAM109B plays a tumorigenic role in low-grade gliomas and is associated with tumor-associated macrophages (TAMs). Journal of Translational Medicine, 22(1), p.833.
- Wang, J., Li, S., Guo, Y., Zhao, C., Chen, Y., Ning, W., Yang, J. and Zhang, H., 2023. Cuproptosis-related gene SLC31A1 expression correlates with the prognosis and tumor immune microenvironment in glioma. Functional & Integrative Genomics, 23(3), p.279.
- Mijiti, M., Maimaiti, A., Chen, X., Tuersun, M., Dilixiati, M., Dilixiati, Y., Zhu, G., Wu, H., Li, Y., Turhon, M. and Abulaiti, A., 2023. CRISPR-cas9 screening identified lethal genes enriched in Hippo kinase pathway and of predictive significance in primary low-grade glioma. Molecular Medicine, 29(1), p.64.

- Williams, A.S., Wilk, E.J., Fisher, J.L. and Lasseigne, B.N., 2023. Evaluating cancer cell line and patient-derived xenograft recapitulation of tumor and non-diseased tissue gene expression profiles in silico. Cancer Reports, 6(9), p.e1874.
- Agrawal, K., Chauhan, S. and Kumar, D., 2023. Expression analysis and regulation of GLI and its correlation with stemness and metabolic alteration in human brain tumor. 3 Biotech, 13(1), p.10.
- Bharanidharan, N., Chakravarthy, S.S., Venkatesan, V.K., Abbas, M., Mahesh, T.R., Mohan, E. and Venkatesan, K., 2024. Local entropy based remora optimization and sparse autoencoders for cancer diagnosis through microarray gene expression analysis. IEEE Access.
- El-Gawady, A., Makhlouf, M.A., Tawfik, B.S. and Nassar, H., 2022. Machine learning framework for the prediction of Alzheimer's disease using gene expression data based on efficient gene selection. Symmetry, 14(3), p.491.
- Zhang, J.J., Shen, Y., Chen, X.Y., Jiang, M.L., Yuan, F.H., Xie, S.L., Zhang, J. and Xu, F., 2023. Integrative network-based analysis on multiple Gene Expression Omnibus datasets identifies novel immune molecular markers implicated in non-alcoholic steatohepatitis. Frontiers in Endocrinology, 14, p.1115890.
- Deng, X., Das, S., Kaur, H., Wilson, E., Camphausen, K. and Shankavaram, U., 2023. Glioma-BioDP: database for visualization of molecular profiles to improve prognosis of brain cancer. BMC Medical Genomics, 16(1), p.168.
- Li, L., Leng, W., Chen, J., Li, S., Lei, B., Zhang, H. and Zhao, H., 2023. Identification of a copper metabolism-related gene signature for predicting prognosis and immune response in glioma. Cancer Medicine, 12(8), pp.10123-10137.
- Almarzouki, H.Z., 2022. Deep-learning-based cancer profiles classification using gene expression data profile. Journal of Healthcare Engineering, 2022(1), p.4715998.
- Gupta, S., Gupta, M.K., Shabaz, M. and Sharma, A., 2022. Deep learning techniques for cancer classification using microarray gene expression data. Frontiers in Physiology, 13, p.952709.
- Sankar, D., Kannan, B., Jayaseelan, V.P., Vasagam, J.M. and Arumugam, P., 2024. Alteration in PNMA1 expression is associated with poor prognosis and tumor immune infiltration in head and neck squamous cell carcinoma. Journal of Oral Biology and Craniofacial Research, 14(1), pp.1-7.
- Mahin, K.F., Robiuddin, M., Islam, M., Ashraf, S., Yeasmin, F. and Shatabda, S., 2022. PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. Genomics, 114(2), p.110264.
- Vega-Benedetti, A.F., Loi, E., Moi, L. and Zavattari, P., 2023. DNA methylation alterations at RE1-silencing transcription factor binding sites and their flanking regions in cancer. Clinical Epigenetics, 15(1), p.98.
- Afify, H.M., Mohammed, K.K. and Hassanien, A.E., 2024. Leveraging hybrid 1D-CNN and RNN approach for classification of brain cancer gene expression. Complex & Intelligent Systems, pp.1-13.
- Das, A., Neelima, N., Deepa, K. and Özer, T., 2024. Gene selection based cancer classification with adaptive optimization using deep learning architecture. IEEE Access.
- Mahmoud, A. and Takaoka, E., 2024. An Enhanced Machine Learning Approach with Stacking Ensemble Learner for Accurate Liver Cancer Diagnosis Using Feature Selection and Gene Expression Data. Healthcare Analytics, p.100373.
- Al Mamlook, R.E., Nasayreh, A., Gharaibeh, H. and Shrestha, S., 2023, May. Classification of cancer genome atlas glioblastoma multiform (tcga-gbm) using machine learning method. In 2023 IEEE International Conference on Electro Information Technology (eIT) (pp. 265-270). IEEE.
- Joshi, A.A. and Aziz, R.M., 2024. Deep learning approach for brain tumor classification using metaheuristic optimization with gene expression data. International Journal of Imaging Systems and Technology, 34(2), p.e23007.
- CuMiDa dataset link: https://www.kaggle.com/datasets/brunogrisci/brain-cancer-gene-expression-cumida. (Accessed on January 2025)
- TCGA dataset link: https://portal.gdc.cancer.gov/.
- Fathy, A., 2024. An efficient spider wasp optimizer-based tracker for enhancing the harvested power from thermoelectric generation sources. Case Studies in Thermal Engineering, 61, p.104878.
- Bajaj, N.S., Patange, A.D., Jegadeeshwaran, R., Pardeshi, S.S., Kulkarni, K.A. and Ghatpande, R.S., 2023. Application of metaheuristic optimization based support vector machine for milling cutter health monitoring. Intelligent Systems with Applications, 18, p.200196.