A Novel Method of Resampling and Support Vector Machine for Brain Tumor Classification

R. Jayanthi[1], A. Hepzibah Christinal[1], R. Hephzibah[1], D.Abraham Chandy[2], T. Shekinah[3]

[1]Department of Mathematics, Karunya University, Coimbatore, India

[2]Department of Electronics and Communication Engineering, Karunya University, Coimbatore, India

[3]Department of Computer Science and Engineering, Loyola-ICAM College of Engineering and Technology, Chennai, India

Abstract: Brain tumors are life-threatening conditions that require accurate diagnosis for effective treatment, and magnetic Resonance Imaging plays a significant role in the diagnosis of brain tumors. Categorization of the tumor type is essential for making necessary medical decisions. Brain Tumor is commonly classified as Normal, Benign, or Malignant. There is a publicly available dataset in Kaggle for brain tumor classification with classes such as meningioma, pituitary gland, glioma, and no tumor. In our work, we proposed a novel method, Smote Tomek with Support Vector Machine (SVM), for brain tumor classification.The Combined sampling technique of smote from oversampling and Tomek from Undersampling was applied to compensate for the imbalance in the data.First, we implemented the combined technique of SMOTETomek to clear this data imbalance, leading to an improvement in the results.We then fitted the balanced data to the SVM classifier. Hence, our proposed method produces the best result with an accuracy of 95% for categorizing the data as pituitary tumor or no tumor. It also provides better results in terms of other metrics such as sensitivity and specificity. This method was also compared with other competent classifiers and was found to be an effective method for the classification of brain tumor data.

Keywords: Brain tumor, SmoteTomek, Support Vector Machine, Classification, Machine Learni

1. Introduction:

Brain tumor classification is important because it may lead to an increase in mortality rates.Brain tumors are cancerous or noncancerous masses of abnormal cells.Brain tumors are serious, depending on the type; they can be cancerous (malignant) or not (benign).When benign or malignant tumors grow, the pressure inside the skull increases.They can cause the pressure inside the skull to increase when benign or malignant tumors grow.In addition to causing brain damage, it is potentially fatal.It can be life-threatening if it causes damage to the brain.The types of tumors are

meningiomas, pituitary glands, and gliomas.It is frequently applied to medical image processing for the early identification of brain cancer, which has led to the development of more effective treatments and a rise in cancer survival rates.There are various diagnosed tumor samples from patients with gliomas and glioneuronal tumors.Glioma and glioneuronal tumors are further classified as glioblastoma multiforme (GBM), gliosarcoma, anaplastic astrocytoma, anaplastic oligodendroglioma, anaplastic oligoastrocytoma, diffuse astrocytoma, oligodendroglioma, oligoastrocytoma, gliomatosis cerebri (GC), astroblastoma, pilomyxoid astrocytoma, pilocytic astrocytoma pleomorphic, xanthoastrocytoma, ganglioglioma, dysembryoplastic neuroepithelial tumor, and glioneuronal tumor [1].Central nervous system tumors are rare, but they are an important cause of death in approximately 20% of young adults and 30% of children.The most common tumors affecting children are pilocytic astrocytoma, malignant gliomas, and embryonal tumors. The most common tumors in adults are pituitary tumors, meningiomas, and malignant gliomas [2]. A survey by the American Cancer Society stated that nearly 16,800 new intracranial tumors were diagnosed in 1999.In the same year, nearly 13,100 people died from primary cancer of the central nervous system [3].Various imaging techniques for detecting tumors in the central nervous system include Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET).The various causes of tumors are based on differences in the characteristics of tissue attenuation, blood-brain barrier breakdown to contrast agents, variations in amino acid or glucose transport, and mass effect [4].Our study proposes a novel method for brain tumor classification using resampling and support vector machine classifiers.The classification is performed using a two-class classification.Here, the brain image is classified as data with no tumor or pituitary tumor.We implemented the sampling technique of SMOTETomek to solve the problem of data imbalance and fitted it to the support vector machine classifier for classification.

2. Related works:

Segmenting brain abnormalities is still a challenging task because of the high irregularity of tumors. In addition, their boundaries are unclear. As the segmentation process is based on the separation of abnormal and normal brain tissues, it has the potential to detect tumors from images. This provides the images. This will provide valuable information for upcoming stages of diagnosis and therapy. In addition, recognizing and diagnosing tumors using MRI at an early stage is extremely important. Glial cells are the source of most primary brain tumors, and are called gliomas [5].The transfer learning method outperformed the other methods and achieved a classification accuracy of 98.71% [6].The author of this paper used the random forest algorithm for autism spectrum disorder to identify and characterize cognitive subtypes and achieved an accuracy of 72.7%, specificity of 80.7%, and sensitivity of 63.1% [7].The authors proposed the construction of classification models, along with hybrid fusion. The

following techniques are utilized to promote classification: low-level features based on the redundant discrete wavelet transform (RDWT), empirical color features, and the gray-level co-occurrence matrix (GLCM) [8]. Wavelet characteristics are then employed to classify the input MRI images using a CNN. This method outperformed other regularly used approaches and achieved an overall accuracy of 99.3% [9].Currently, modern and efficient automated computer-assisted diagnostic (CAD) systems can address difficulties with high accuracy. The overall classification resulted in an accuracy of 98.04% compared to the proposed methods [10]. Machine learning-driven data preparation (MLDP) is used to achieve optimal data preparation (DP) before building cancer prediction models [11].They achieved the best results using an Ensemble Classifier and sampling technique (SMOTE + Tomek). The best-performing model across all evaluation metrics correctly identified 84% of all stroke cases (sensitivity = 0.84) and 75% of all healthy cases (specificity = 0.75) [12].We presented a novel CNN architecture for classifying brain tumors with an accuracy of 96.56% [13]**.** The authors proposed a hybrid model that combines CNN and support vector machine (SVM) classification with threshold-based segmentation detection. The hybrid CNN-SVM produced an overall accuracy of 98.49% [14].The author used a back propagation neural network (BPN) and a Radial Basis Function Neural network (RBFN) to automatically classify brain MRI images as cancerous or non-cancerous tumors. This proves that the RBFN algorithm outperforms the BPN algorithm, with a classification accuracy of 85.71% [15].The authors proposed extracting texture features from brain MR images. We achieved a classification accuracy greater than 99% [16]. Theyused textural Features and Supervised learning methods (FSVM). Based on this classifier, FSVM has been used to classify subjects' brain MR images as normal or abnormal. The proposed algorithm achieved 95.80 a classification accuracy [17].

3. Material And Methodology:

3.1 Data Description:

There is a publicly available dataset in Kaggle for brain tumor classification with classes such as Meningioma, Glioma, Pituitary tumor, and no tumors. This dataset contains 3264 files separated into training and testing folders. Brain Tumors are classified as benign, malignant, or pituitary. In our method, we classified our data into two classes: no tumor and pituitary tumors.Magnetic Resonance Imaging (MRI) is a reliable method for detecting brain tumors (MRI). These scans produced a large amount of image data. The radiologist inspected the images. Owing to the intricacy of brain tumors and their traits, a manual assessment can lead to errors.

3.2 Application of sampling technique

In this study, we proposed a novel resampling classification technique implemented in an SVM classifier. We used MRI brain tumor data publicly available for our classification process. As there was an uneven distribution of data, in the case of pituitary and no tumor data, we applied our sampling technique. There are two types of sampling techniques, Oversampling and Undersampling.The predominance class sample count was determined using the undersampling technique. Included among the numerous undersampling methods are the edited nearest neighbor rule, Tomek link undersampling, and random undersampling [25]. By using oversampling techniques, which increase the number of training samples for minority classes, an unbalanced dataset can be corrected. Numerous techniques for oversampling are available, including random oversampling, synthetic minority oversampling, borderline SMOTE, ADASYN, safe-level SMOTE, and others [26]. In our study, we used the combined technique of sampling SMOTETOMEK. SmoteTomek is the integration of SMOTE from Oversampling and Tomeklinks from Undersampling.

3.2.1 Synthetic minority oversampling technique (SMOTE)

This is a method of oversampling, known as the synthetic minority oversampling technique (Smote). Inserting minority-class examples that are close together is the main idea behind this. Through this expansion of the minority class's decision boundaries, the minority class avoids the overfitting problem[18].An integrating technique of oversampling the minority class and undersampling the majority class can help improve the performance of classification. Creating synthetic minority class instances entails minority class over-sampling [19].

3.2.2 Tomek-Links:

The Tomek link is defined as follows: Consider S_a and S_c belong to different classes, and d (S_a, S_c) is the distance between S_a and S_c. The (S_a, S_c) pair is considered as Tomelink if there is no example of S_b, such that d (S_a, S_b) < d (S_a, S_c) or d (S_c, S_b) < d (S_a, S_c).Whenever two examples form a Tomek link, they may be noise or borderline.They can use a Tomek link to undersample or clean the data.The examples from the majority class are only eliminated using an under-sampling method, whereas examples from both classes are removed using a data-cleaning method [20].

3.2.3 Smote+Tomek:

Although over-sampling minority class examples helps to balance class distributions, they do not solve some other issues that are common in datasets with skewed class distributions. Class clusters are frequently not properly defined because some examples from the majority class may invade the minority class space. Improving minority class examples can extend minority class clusters, paving the way for artificial minority class examples too deeply within the majority class space.

Prompting a classifier in such a situation may cause overfitting. We propose using Tomek links as a data-cleaning method on an oversampled training set to create well-defined class clusters. As a result, instead of removing the examples that form Tomek links, examples from both classes were removed [21].

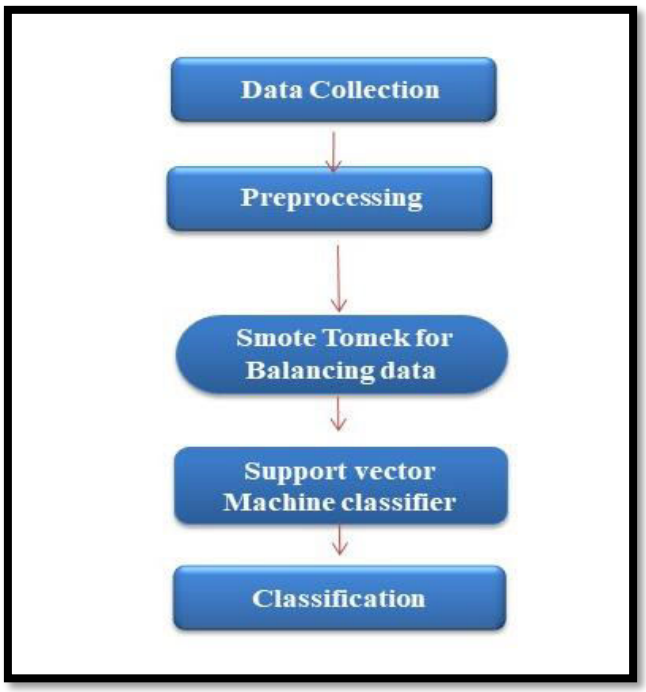

Fig:1 Flow chart of our proposed method

3.3 SVM classifier:

SVM is a classifier method that constructs hyperplanes in a multidimensional space that splits cases with various labels. Multivariate SVMs can hold both categorical and continuous variables and can perform regression and classification tasks. The black line that divides the two clouds of the class runs straight down the center of the channel. Segregation occurs in two dimensions: a line, in three dimensions, a plane, four or more dimensions, and a hyperplane. We can find something numerically in the separation by taking two critical members, one for each class. These are known as support vectors (SVs). The critical points (members) defining the channel are listed below. The perpendicular bisector of the line connecting these two support vectors was then separated. This is the concept of a support vector machine.

$$\frac{1}{2}f^{n}f + Q\sum_{t=1}^{M}\sigma_{t}$$

Subject to Constraints

$$y_{t}(f^{n}\,\phi(x_{t}) + b) \geq 1 - \sigma_{t} \text{ and } \sigma_{t} \geq 0, t = 1, \ldots M$$

where Q and b areconstants, f is the coefficient vector, and σ_t represents the parameters for non-separable data (inputs). where t denotes M training instances. Note that $y \in \pm 1$ indicates the labels, and x_t denotes the independent variables. The kernel ϕ is used to change the data from the input (independent) space to the feature space. It should be noted that as Q increased, the error decreased.Thus, Q should be carefully selected to avoid overfitting [22].

4. Experimental Results:

4.1 Performance Metrics

Specificity:

They defined specificity as the extent to which they correctly identified negatives. They related this to a test's ability to detect negative results.

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})}$$

where TN refers to the True Negative and FP refers to the False Positive.

Accuracy:

It is calculated as the ratio of all true results to the total number of cases checked

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

Sensitivity:

The percentage of true positives correctly identified is referred to as the sensitivity. This is related to the ability of the test to detect positive results.

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad [23]$$

Where TP and TN represent True Positive and Negative and FP and FN represent False Positive and Negative

Precision:

Precision measures the accuracy of the model predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall:

Model's ability to detect classes.

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

F1-score:

It combines precision and recall to evaluate a classifier.

$$\text{F1} - \text{Score} = 2 \times \frac{(\text{precision} * \text{Recall})}{\text{precision} + \text{Recall}}$$

4.2 Results and Discussion

Brain MRI data used for experimentation are publicly available on the Kaggle website. The dataset contains various types of tumor data. In this study, we implemented our proposed method to classify tumors with no tumor data. Experiments were performed using Google Colab with a GPU backend. The experimental results of the proposed method are presented in Table 1.

Technique	class	Precision	Recall	F1-score	Specificity	Sensitivity	Accuracy
Proposed SMOTETOMEK with SVM	0(no tumor)	0.99	0.86	0.92	0.99	0.86	0.95
	1(Tumor)	0.94	0.99	0.96			

Table:1 Results of our proposed method

The results of our method are clearly described in Table 1. Our method achieved an accuracy of 95% for classifying the no-tumor and pituitary data. Here, class 0 represents no tumor data, and class 1 represents pituitary tumor data. It also achieved a specificity of 99% and sensitivity of 86%. The Precision, Recall, and F1-score results of the data for classes 0 and 1 are listed in Table 1. The Precision was 99% in classifying the no-tumor data and 94% in classifying the tumor data. The F1 score was also found to be 96% in classifying the tumor data and 92% in classifying the no-

tumor data. Overall, our proposed method, which is a combination of SMOTE from oversampling and Tomeklinks from Undersampling implemented in the SVM classifier, was found to achieve promising results in terms of Accuracy, Sensitivity, and Specificity in classifying the tumor and no tumor data.

Technique	**Specificity**	**Sensitivity/recall**	**Accuracy**
Naïve Bayes	81.85	80.83	81.33
Decision Tree	93.51	92.80	93.157
Neural Network	92.76	93.27	93
KNN [27]	95.23	94.30	94.765
Proposed Method	**99.0**	**86.0**	**95**

Table:2 Comparison of other methods with the proposed method

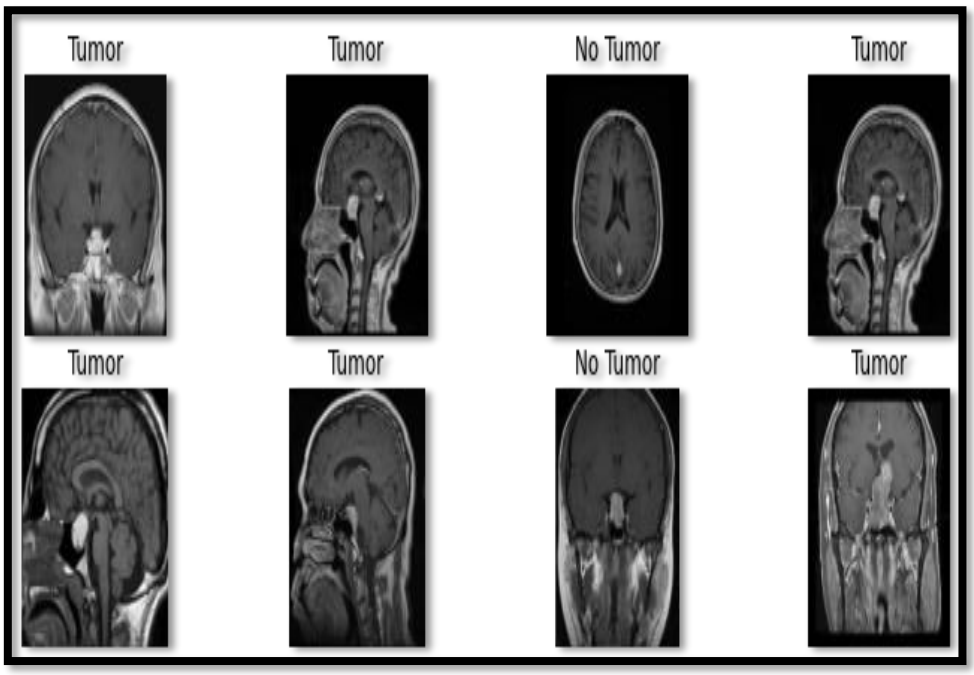

Fig2 Classification of Brain Tumor Data using SMOTETOMEK and SVM classifiers.

Table 2 presents the comparative results of our proposed method with other basic classifiers such as Naïve Bayes, Decision Tree, Neural Network, and KNN. The results of our method were compared with those of these methods in terms of Accuracy, Sensitivity, and Specificity. As shown in Table 2, the results of the Naive Bayes

classifier are found to be less than those of the other classifiers. The Decision Tree, Neural Network, and KNN were found to have an accuracy closer to that of our proposed method, although our method has a greater accuracy than other methods. Although the sensitivity is less than that of some classifiers, our method produces a specificity of 99%, which is an astonishing performance of our proposed method. Hence, our method is compared with other methods, and it is concluded that our method has produced remarkable results. Fig 2 depicts the Classified Brain tumor dataset into tumor and no-tumor images with our proposed method.

Conclusion:

Brain tumor classification is important because it may lead to a greater increase in deaths. Magnetic Resonance Imaging (MRI) has the foremost role in the diagnosis of brain tumors. In our study, we used the Brain Tumor MRI Kaggle dataset for classification. We propose a novel resampling method using an SVM classifier for classification. Because there is a problem of data imbalance, we implemented sampling techniques to balance the data. The resampling method involves a combination of the oversampling technique SMOTE and the Undersampling technique Tomeklinks. Resampling was used for data imbalance. We then used a support vector machine for classification. In this study, we implemented a two-class classification for classifying tumor and no tumor data. We then evaluated the performance of our proposed classifier in terms of Accuracy, Sensitivity, and Specificity, and compared it with other basic classifiers such as Naïve Bayes, KNN, Neural Network, and Decision Tree. By analyzing the results, we found that our method achieved excellent results and was an effective classifier. Further improvements can be achieved by implementing a bagging or boosting technique with the classifier.

References:

1. 赤木洋二郎. (2020). Reclassification of 400 consecutive glioma cases based on the revised 2016 WHO classification (Doctoral dissertation, 九州大学).
2. Kuwahara, K., Ohba, S., Nakae, S., Hattori, N., Pareira, E. S., Yamada, S., ... & Hirose, Y. (2019). Clinical, histopathological, and molecular analyses of IDH-wild-type WHO grade II–III gliomas to establish genetic predictors of poor prognosis. Brain tumor pathology, 36, 135-143.
3. Afshar, P., Mohammadi, A., & Plataniotis, K. N. (2018, October). Brain tumor type classification via capsule networks. In 2018 25th IEEE international conference on image processing (ICIP) (pp. 3129-3133). IEEE.

4. Kuraparthi, S., Reddy, M. K., Sujatha, C. N., Valiveti, H., Duggineni, C., Kollati, M., & Kora, P. (2021). Brain Tumor Classification of MRI Images Using Deep Convolutional Neural Network. Traitement du Signal, 38(4).
5. Mourad, A., & Afifi, A. (2020, November). Automated Brain Tumor Segmentation in MRI using Superpixel Over-segmentation and Classification. In 2020 21st International Arab Conference on Information Technology (ACIT) (pp. 1-8). IEEE.
6. Chelghoum, R., Ikhlef, A., Hameurlaine, A., & Jacquir, S. (2020). Transfer learning using convolutional neural network architectures for brain tumor classification from MRI images. In Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16 (pp. 189-200). Springer International Publishing.
7. Feczko, E., Balba, N. M., Miranda-Dominguez, O., Cordova, M., Karalunas, S. L., Irwin, L., Demeter, D. V., Hill, A. P., Langhorst, B. H., Grieser Painter, J., Van Santen, J., Fombonne, E. J., Nigg, J. T., & Fair, D. A. (2018). Subtyping cognitive profiles in Autism Spectrum Disorder using a Functional Random Forest algorithm. NeuroImage, 172, 674–688.
8. Yadav, A. S., Kumar, S., Karetla, G. R., Cotrina-Aliaga, J. C., Arias-Gonzáles, J. L., Kumar, V., Srivastava, S., Gupta, R., Ibrahim, S., Paul, R., Naik, N., Singla, B., & Tatkar, N. S. (2022). A Feature Extraction Using Probabilistic Neural Network and BTFSC-Net Model with Deep Learning for Brain Tumor Classification. Journal of imaging, 9(1), 10.
9. Sarhan, A. M. (2020). Brain tumor classification in magnetic resonance images using deep learning and wavelet transform. Journal of Biomedical Science and Engineering, 13(06), 102.
10. Guan, Y., Aamir, M., Rahman, Z., Ali, A., Abro, W. A., Dayo, Z. A., ... & Hu, Z. (2021). A framework for efficient brain tumor classification using MRI images.
11. Krajnc, D., Spielvogel, C. P., Grahovac, M., Ecsedi, B., Rasul, S., Poetsch, N., ... & Papp, L. (2022). Automated data preparation for in vivo tumor characterization with machine learning. Frontiers in Oncology, 12.
12. Jagwani, G. (2019). Identifying the Patients at Risk of Stroke Using Anomaly Detection Based Classification Approach (Doctoral dissertation, Dublin, National College of Ireland).
13. Badža, M. M., & Barjaktarović, M. Č. (2020). Classification of brain tumors from MRI images using a convolutional neural network. Applied Sciences, 10(6), 1999.
14. Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., & Jhanjhi, N. Z. (2022). A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. Irbm, 43(4), 290-299.

15. Deepa, S. N., & Devi, B. A. (2012, January). Artificial neural networks design for classification of brain tumour. In 2012 International Conference on Computer Communication and Informatics (pp. 1-6). IEEE.
16. Qurat-Ul-Ain, G. L., Kazmi, S. B., Jaffar, M. A., & Mirza, A. M. (2010). Classification and segmentation of brain tumor using texture analysis. Recent advances in artificial intelligence, knowledge engineering and data bases, 147-155.
17. Jayachandran, A., Dhanasekaran, R., & Ammal, S.G. (2013). Brain Tumor Detection and Classification of MR Images Using Texture Features and Fuzzy SVM Classifier. Research Journal of Applied Sciences, Engineering and Technology, 6, 2264-2269.
18. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
19. Kumar, V., Lalotra, G. S., Sasikala, P., Rajput, D. S., Kaluri, R., Lakshmanna, K., ... & Uddin, M. (2022, July). Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. In Healthcare (Vol. 10, No. 7, p. 1293). MDPI.
20. Tomek, I. (1976). Two modifications of CNN.
21. Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing training data for automated annotation of keywords: a case study. In WOB (pp. 10-18).
22. Jagannathan, D., & Phil, M. (2017). Cardiotocography-a comparative study between support vector machine and decision tree algorithms. International Journal of Trend in Research and Development, 4(1).
23. Bangare, S. L. (2022). Classification of optimal brain tissue using dynamic region growing and fuzzy min-max neural network in brain magnetic resonance images. Neuroscience Informatics, 2(3), 100019.
24. Demir, F., & Akbulut, Y. (2022). A new deep technique using R-CNN model and L1NSR feature selection for brain MRI classification. Biomedical Signal Processing and Control, 75, 103625.
25. Yen, S. J., & Lee, Y. S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In Intelligent Control and Automation (pp. 731-740). Springer, Berlin, Heidelberg.
26. Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. Applied Soft Computing, 83, 105662.
27. Garg, G., & Garg, R. (2021). Brain tumor detection and classification based on hybrid ensemble classifier. arXiv preprint arXiv:2101.00216.