[1]Abdul Wahid,

[1]M.Tariq Banday

ETL-CCP: An Effective Ensemble Transfer Learning approach for Cancer Classification and Prediction using Gene Expression data

Abstract: - Pan-cancer classification refers to classifying and diagnosing different types of cancer. The deep learning approach enables the detection and diagnosis of cancer across multiple organs and tissue types rather than being limited to a specific type of cancer. In deep learning, transfer learning is a technique whereby a neural network model is first trained on a problem similar to the problem that is being solved. Similarly, Ensemble Learning is a strategy where multiple models are combined to perform a particular task. The ensemble transfer learning method can be an effective alternative for Pan-Cancer classification as it can overcome the limitations of base methods by combining the predictions of multiple models to improve performance. In this work, we have presented a novel ensemble method called ETL-CCP for the classification of 33 cancer types based on gene expression data by combining the predictions of multiple pre-train models of transfer learning to improve performance and robustness and increase interpretability. The method combines the efficiency of DenseNet, ResNet, Inception-V3, and Xception models to capture high-level features from structured input data. In addition to this, a class-weighting mechanism is used to overcome data imbalance issues. The experiments were conducted on a gene expression dataset comprising 10,267 cancer samples from 33 cancer types. Our method achieved a test data accuracy of 96.88%, outperforming current baseline methods. This research demonstrates the potential of ETL-CCP as a powerful tool for cancer detection and highlights the importance of ensemble methods in high-dimensional data analysis.

Keywords: Pan-Cancer, TCGA, Ensemble Transfer learning, Gene-Expression

1. INTRODUCTION

Cancer can profoundly impact humans by causing physical, emotional, and social challenges, potentially leading to severe illness, decreased quality of life, and, in some cases, loss of life (Yau E et al. 2017**)**. Genes play a crucial role in the spread of cancer (metastasis) by regulating processes related to cell adhesion, migration, invasion, and angiogenesis, allowing cancer cells to detach from the primary tumour, travel through the bloodstream or lymphatic system, and establish secondary tumours in distant organs. Genetic mutations can cause genomic instability and an increased tendency for DNA mutations to occur. This instability can lead to further genetic alterations and an increased risk of cancer development. Genomic instability is a hallmark of many cancers (Bray F et al. 2018). These mutations in genes that control mitosis result in the uncontrollable division of cells and, therefore,tumour formation. If the tumours are malignant, they can metastasize or spread to other parts of the body and become life-threatening. Each type of cancer has its specific variance in the structure of the genetic aberrations that arise, including somatic mutations (SM), copy number variations (CNV), profiles, and various epigenetic alterations. Thus, changes in gene expression might arise from environmental factors, such as cell division or genetic inheritance (Zuo S et al. 2019; Cruz-RoaA et al. 2017). For example, mutations in the BRCA1 and BRCA2 genes increase the risk of breast and ovarian cancer, while mutations in the Lynch syndrome genes increase the risk of colorectal and other types of cancer (Mukherjee A et al. 2022). Gene expression changes can occasionally alter howspecific proteins are produced, which impacts how normally behaving cells behave. These damaged cells begin to divide more quickly than usual, and as they continue to do so, they eventually fill the afflicted area with tumour-like growths (Podolsky M D et al. 2016). Therefore, understanding the genetic changes that lead to cancer is essential because it can help researchers develop new therapies that target specific genetic mutations or pathways. Existing algorithms for feature engineering of gene expression data can be categorized into three major groups: filter, wrapper, and embedded methods. The information flow in the three groups of feature engineering methods is depicted in Figure 1.

[1] Department of Electronics and Instrumentation Technology,University of Kashmir, Srinagar, Jammu & Kashmir

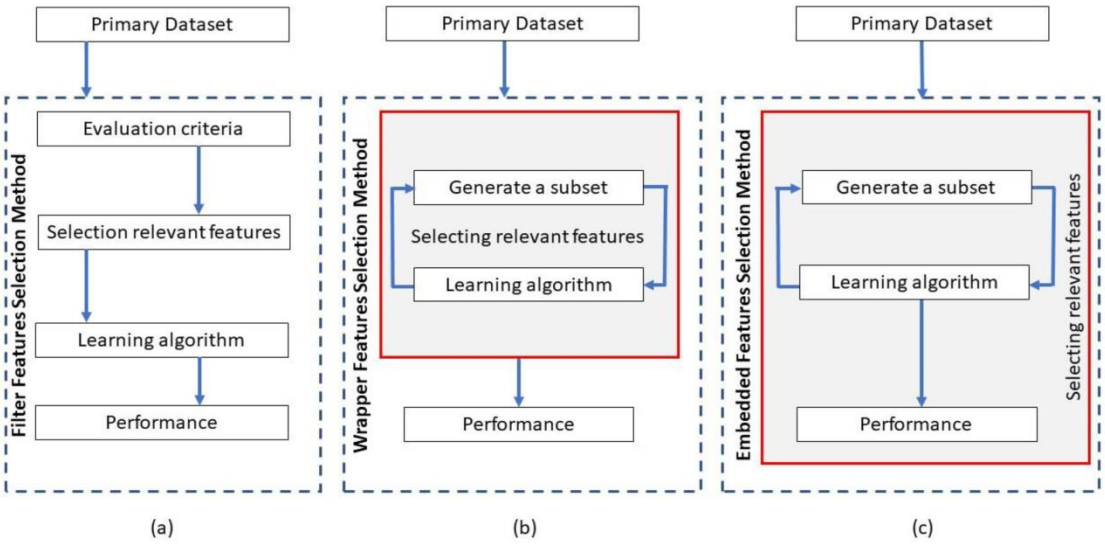

Figure1. Information flow in the three primary categories of feature engineering methods: (**a**) filter, (**b**) wrapper, and (**c**) embedded methods

1.1 PAN-CANCER CLASSIFICATION

Pan-Cancer classification refers to simultaneously classifying and analyzing genomic data across multiple types of cancer. Pan-Cancer classification aims to identify common patterns, mutations, and molecular characteristics that cut across different cancer types, providing insights into shared mechanisms of cancer development and potential treatment strategies. It is a relatively new approach to cancer classification that aims to identify commonalities and shared features among different types of cancer, regardless of where they originate in the body (Crosby D et al. 2022). Cancer classification has been the subject of extensive research because it allows researchers to identify new therapeutic targets and develop more effective therapies. It can also help improve the accuracy of cancer diagnosis and aid in developing personalized medicine (Sakri S. B et al. 2018; Wang J et al. 2021).

To explore and understand the similarities and variations across various cancer types, the Pan-Cancer Atlas project was introduced by The Cancer Genome Atlas(TCGA)(Weinstein J N et al.2013). Pan-Cancer research has progressively increased worldwide, and experts are trying to identify the tumor-related genes to identify the specific type of cancer precisely. The Pan-Cancer Atlas offers substantial information on 33 common cancer types, which we may utilize as base material to acquire cancer-specific biomarkers. The invention of next-generation sequencing methods has advanced the analysis of human genomics due to its efficiency and accuracy. It has been observed that a massivevolume of tumor tissue has been sequenced and managed by TCGA. By these samples, TCGA further analyzed over 11,000 tumors from the 33 most dominant forms of cancer, which fostered the accomplishment of the Pan-Cancer Atlas (Wang Z et al. 2016). However, classification can be difficult and time-consuming to perform manually and experimentally due to the large amount of data involved. For example, whole-genome sequencing generates large amounts of data, and it's impossible to make sense of the data only by manual inspection. Cancer is also a complex disease that can arise from genetic and environmental factors, making it difficult to classify based on traditional methods (Zhu W et al. 2020). Additionally, environmental and human factors make manual observation prone to error. Given the ongoing advancement of cancer research, it may become more and more challenging as identification techniques advance (Sharma A& Rani R. 2021).

Therefore, prediction methods can be a better alternative for Pan-Cancer classification because they allow researchers to analyze large amounts of data quickly and accurately. For example, machine learning algorithms can be trained on large genetic and molecular data datasets to classify new samples. These methods can identify patterns and trends in the data that would not be apparent through manual inspection and can help improve the accuracy of cancer classification (Sharma A& Rani R. 2021).

1.2 MOTIVATION AND CONTRIBUTION

Using computational methods to classify Pan-Cancer can improve clinical outcomes by providing more accurate and precise diagnosis and treatment, identifying new therapeutic targets, and developing more effective therapies (Kourou K et al. 2015; Okamoto O K. 2005). Therefore, numerous machine learning (ML) techniques have been used in recent years to cope with cancer detection and prediction based on gene-expression data (Zhu W et al. 2020; Sharma A& Rani R. 2021; Kourou K et al. 2015). However, most traditional cancer classification methods primarily concentrate on a few cancer types while ignoring the variability among various cancer types (Lawrence M S et al. 2013; Lyu B &Haque A. 2018). Classical ML methods for Pan-Cancer classification (Way G P et al. 2018) generally could not handle high-dimensional data efficiently, resulting in less accurate outputs (Lyu B &Haque A. 2018). Remarkably, these models suffer from several limitations, such as data scarcity, imbalance, overfitting, and limited interpretability. These limitations can affect the performance and robustness of these models, making it challenging to achieve accurate and reliable results in Pan-Cancer classification. On the other hand, inspired by the success of deep learning methods in image classification (Mansour R et al. 2022; Wu H et al. 2019) text classification, and sequence-to-sequence classification(Deng L & Li X. 2013; Alsolami B et al. 2020) different methods such as Convolutional neural networks (CNN) and Recurrent Neural networks (RNN) have been popularly used for improving cancer prediction from gene expression data. Ensemble methods can be an effective alternative for Pan-Cancer classification using deep learning, as they can overcome the limitations of deep learning methods by combining the predictions of multiple models to improve performance and robustness and increase interpretability.

In this work, we propose the ensemble method (ETL-CCP) for the classification of all 33 types of cancer based on high-dimensional gene-expression data. The main contribution of this work includes the construction of Ensemble of transfer learning models(Zhuang F et al. 2021) such as DenseNet(Huang G et al. 2017), ResNet(He K et al. 2016), Inception-V3(Szegedy C et al. 2016) and Xception(Chollet F. 2017)which were based on advanced deep learning architectures We then used a Class-weighting strategy to handle the imbalanced nature of the data and finally various evaluation metrics were used for Extensive experiments and analysis of the Pan-Cancer data.

The remaining paper is organized as follows: Section 2 discusses the review of related work. Section 3 discussed the ETL-CCP (a proposed framework) and the materials and methods used in this work. The experimental setup, implementation results, and comparative analysis are discussed in subsequent sections.

2.RELATED WORK

Over the years, research has been conducted to develop reliable and accurate methods for classifying different tumor types. Machine learning (ML), a branch of artificial intelligence, has been widely used for image and speech recognition, traffic prediction, disease classification , medical diagnosis, and online fraud detection (Alsolami B et al. 2020; Fatima M & Pasha M. 2017; Eltanbouly S et al. 2020). Meanwhile, these methods have been used in Pan-Cancer classification. Li et al. used the K-nearest-neighbor (KNN) and genetic algorithm (GE) methods to classify 31 cancer types, achieving an accuracy of 90% (Li Y et al 2017). (Kang et al. 2019)proposed a support vector machine (SVM) method combined with relaxed Lasso selection for cancer classification. As reported, this method outperformed other ML methods, such as KNN and Logistic regression. Hsu et al. extended the cancer classification from 31 to 33 types. They used various machine-learning methods on the Pan-Cancer dataset(Hsu Y H & Si D. 2018) which include decision trees (DT), KNN, SVM (linear and polynomial), and sophisticated neural networks. These methods often have limitations when dealing with high-throughput genomic data with high dimensionality and heterogeneity, resulting in less reliable outcomes. In addition, ML methods demand considerable vigilant feature extraction and selection for better performance besides fine-tuning.

On the other hand, deep learning has become the fastest-growing technology in the last decade, where massive data storage and computational resources are required. Deep learning has made phenomenal progress in the previous five years, and it is comprised of various algorithms such as Convolutional neural networks (CNN), Recurrent neural networks (RNN), etc. Inspired by the success of deep learning methods in image, text, and sequence-to-sequence classification, deep learning methods have been popularly used for improving cancer prediction for gene expression data. As a result, researchers have utilized deep neural networks for gene-

expression public datasets for cancer prediction. Deep learning techniques outperform conventional machine learning techniques for faster convergence and data representation.

(Danaee P et al. 2017) have proposed an Artificial neural network (ANN)-based method for improving cancer classification. This work uses auto-encoders to efficiently capture and extract features fed to the subsequent ANN method for classification. (Cristovao et al. 2020)conducted an extensive study on predicting clinical outcomes and cancer sub-type classification from gene expressions using semi-supervised methods and feed-forward neural networks.(Lyu B &Haque A. 2018) provided one of the earliest discussions of convolving over gene expression data to extract local patterns efficiently. They outlined a CNN-based model for enhanced cancer sub-type classification.In their work, the authors transformed the raw expression data into two-dimensional shapes, which were then forwarded to the CNN method. Although the techniquesignificantly improved the classification of 33 types of cancer, it could not perform well for classes with fewer samples. (Khalifa N E M et al. 2020) attempted to use CNN with particle swarm optimization to classify five cancer subtypes, achieving an accuracy of 96.6%. However, it did not predict all types of classes. (Divate M et al 2020) proposed another deep neural network for classifying cancer types, consisting of five dense layers with 50 neurons at each layer. However, this method could not perform well on gene expression data. Therefore, a grid search was conducted to find the optimum hyper-parameters. The study confirms that fine-tuning deep learning algorithms can considerably improve performance. Joseph et al. developed a two-dimensional CNN-based algorithm consisting of three CNN layers for the Pan-Cancer classification of 33 cancer types, and the method achieved an accuracy of 95.43% on the test dataset (Joseph M et al. 2019). Wang et al. extended cancer classification work by filtering noise and duplicate data from the dataset and then using a popular DenseNet model for classification (Wang J et al. 2021). Although deep learning has significantly aided in identifying and categorizing many malignancies, there is still an opportunity for advancement. (Cava et al. 2023)proposed a pan-cancer classification on 16 cancer types using deep learning model on gene expression data. The three classifiers used by the authors are neural network, random forest and XGBoost and the mean accuracy achieved by each of these classifiers are 0.84, 0.86 and 0.90 respectively. (Li et al. 2020) proposed a method to classify cancer types by using the self-normalizing neural network (SNN) for analysing Pan-Cancer copy number variation data. The MCFS and IFS were used for feature selection. The author selected 3654 features for the prediction model in which the accuracy value of 0.798 and macro F1 of 0.789 were yielded.

3. ETL-CCP (ENSEMBLE TRANSFER LEARNING FOR CANCER CLASSIFICATION & PREDICTION)

This section will discuss the dataset, pre-processing, ensemble method, and evaluation metrics.

3.1 DATASET

The proposed work used the normalized gene-expression data of 33 cancer types extracted from the Pan-Cancer Atlas available at the TCGA data portal and the Cancer Genomic Hub web portal. The dataset consists of 10267 cancer samples with 20531 genes. The TCGA research network, known as the “process of genomic discovery,” consists of four essential components: sample collection and processing, genome characterization, genome sequencing, and data management and analysis (Wang Z et al. 2016). The first step in our approach is to pre-process the Pan-Cancer dataset and transform the raw gene expression data into well-formed shapes compatible with deep learning algorithms. As the dataset consists of trivial information, data was collated from 33 distinct cancer types and then curated and filtered out with less divergent genes. After pre-processing, the number of genes from 20531 was reduced to 10381. Based on the approach of (Lyu B &Haque A. 2018)the high-dimension expression data of (10381x1) was converted into into a 2-D image (102x102) data and same was used as input to the proposed deep learning architecture. In the next step, the pre-processed data is fed to the proposed ETL-CCP method for classifying cancer types.

3.2 PROPOSED TRANSFER LEARNING FRAMEWORK

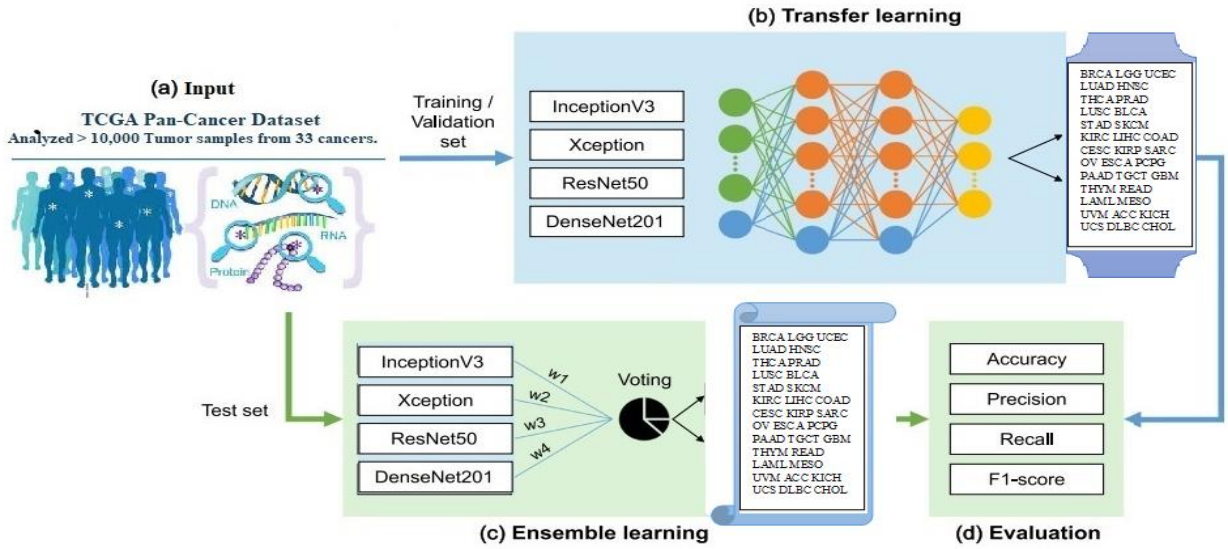

Figure 2.The overall framework of the ETL-CCP method for Pan-Cancer

Transfer learning allows a pre-trained model to transfer knowledge across domains, making it an essential technique for various applications, including image recognition, natural language processing, and more. It accelerates training, improves performance, and reduces the need for massive data. Ensemble learning is a powerful technique combining multiple models to achieve better predictive performance than any individual base model. It helps mitigate overfitting, reduce bias, and improve generalization. The pre-trained models blends the insights from base models and create a more robust and accurate ensemble model that captures a broader range of features and patterns. The proposed ensemble method consists of advanced deep learning architectures such as DenseNet, ResNet, Inception-V3, and Xception(Huang G et al. 2017; He K et al. 2016; Szegedy C et al. 2016; Chollet F. 2017) were used for the classification of 33 cancer types. The general framework of ETL-CCP is shown in Figure. 1.

DenseNet: DenseNet, or a densely connected convolutional neural network, is an image classification algorithm developed by (Huang G et al. 2017)to improve the performanceand accuracy of model by handling the problem of vanishing gradients.It is known that deep neural networks have many hidden layers between the input and output layers, and due to this long path between the input and output layers, information vanishes before reaching its destination. DenseNet overcomes the vanishing gradient problem and achieves higher accuracy than other baseline CNNs by connecting every layer.

ResNet: Although DenseNetshas improved performance and accuracy, these densely connected neural networks require substantial computational resources to achieve the best results. Therefore, Residual networks gained much popularity as these networks comprise a novel pathway called skip connections. These connections provide an alternative path for data and gradients to flow, thus making Training more efficient. Previous research confirms that the performance of ResNets is comparable with that of DenseNets, while the computational cost of ResNets is less than that of DenseNets(He K et al. 2016).

Inception-V3: Inception-V3 is an updated version of Inception-v1 in which 1x1 convolutional layers have been added to reduce the number of dimensions, making it a more accurate and computationally efficient model.The other improvements in Inception-V3 include the factorization of large convolutions into smaller and asymmetric convolutions and the reductionof the number of parameters (Szegedy C et al. 2016).

Xception: The Inception model from Google inspired the 71-layer deep CNN known as theXception model, which is based on an extreme interpretation of the Inception model (Chollet F. 2017). Convolutional layers that can be separated based on depth make up its architecture. These Depth-wise separable convolutions have a few valuable features in how they “factorize” the convolution layer, which allows a convolution layer to be emulated with fewer parameters. Additionally, the separable convolution handles the cross-depth features and

2D features separately. This way, cross-depth features would not be destroyed by normal convolution. This is known as the “Extreme Inception Hypothesis”.

Although these advanced deep learning architectures have gained overwhelming popularity due to their competitive performance, they also have inherent drawbacks. For ResNet, the identity shortcut stabilizing training may limit its representation capacity, and DenseNet mitigates it with multi-layer feature concatenation. However, the dense concatenation causes the new problem of requiring high GPU memory and more training time. Moreover, inception networks using CNNs as their core are translation-invariant and are generally bad at handling rotation and scale-invariance without explicit data augmentation. The Xception model can be challenging to interpret and analyze, making it difficult to understand how they arrive at their predictions. Therefore, combining these models in an ensemble increases diversity through collections of different models and reduces bias across the full collection of models. This work is focused on combining DenseNet-201, ResNet50, Inception-V3, and Xception models in an ensemble, as shown in Figure 1 above.

3.3 EVALUATION METRICS

In this work, various evaluation metrics have been utilized to measure the quality of the ETL-CCP method for classifying cancer sub-types.

a) Accuracy: Accuracy is one of the most popular and simplest measures to assess the model's quality.It is the percentage at which our model predicted or classified our data points correctly and is calculated as formula 1:

$$\text{Accuracy} = \frac{\text{No.of correctly classified points}}{\text{Total no.of points}} \times 100 \qquad (1)$$

b) Precision: Since accuracy does not fit well for models trained on imbalanced data, Precision-Recall metrics give a more generalized assessment of the model.Precision is the ratio of True-positives tothe total number of data points predicted as Positive by the model and is calculated as given in equation 2.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

c) Recall: Recall is the ratio of True positives toa Total number of Positive data points in the given dataset and is calculated as equation 3.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

d) F1-score: The F1-score is the harmonic mean of Precision and Recall.It is an important evaluation metric, as accuracycannot be a basis for a better-generalized model.While Precision and Recall depict the model's generalizability, the F1-score captures this important property and is calculated as equation 4.

$$F_1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

4. RESULTS AND DISCUSSION

Different evaluation metrics have been used to evaluate the performance and generalization of the proposed ETL-CCP.This section discussesthe experimental training setup and the method evaluation against the test data.In addition, performance analysis anda comparative analysis of our proposed approach with cutting-edge methods are discussed in subsequent sub-sections.

4.1 EXPERIMENTAL SETUP AND IMPLEMENTATION

TensorFlow(Pang B et al. 2020)and Keras(Chicho B T &Bibo Sallow A 2021) are the frameworks of deep learning (Yu Li et al. 2019) that have been used to develop and train the proposed models for Pan-Cancer classification, respectively. The experiment was done on a High-Performance Computing System whose configuration details included 9 CPUs distributed as 2 Head nodes, 2 Compute nodes, 3 GPU nodes, and 2 Storage nodes, respectively. Each CPU is an Intel Xeon Gold 3.1G with 240 compute Cores, 15360 CUDA cores, and 160 TB of storage with a performance of 2.976 TFLOPS. We used Python on a Jupyter Notebook in the Anaconda Environment to train and test the model. ETL-CCP was fed with default initial TensorFlow weights. To avoid overfitting the model, L2 regularizes(Xue Ying. 2019) have been used, and an early stopping

approach with a patience level of 7 has been exploited to monitor the performance and overfitting. The learning rate of 0.0001 and decay of 0.5 were used to reduce the learning rate after every 40 epochs. The total number of epochs for training was limited to 300 because the change in accuracy started converging to 300 epochs. The optimum batch size for this work after conducting several experiments was 128. The final output layer consists of a SoftMax activation function. For the overfitting issue, we introduced the dropout (Salehin et al. 2023) whose value was set as 0.5, which means 50% of input units will drop out randomly during each training epoch, and then it randomly goes through drop nodes.

Transfer learning is a powerful approach that allows a pre-trained model, which has been trained on a large dataset for a specific task, to be fine-tuned and adapted to perform well on a new, related task with a smaller dataset. On the other hand, Ensemble learning involves combining the predictions of multiple individual models (base models) to create a single, stronger model. By aggregating the predictions of diverse models, the overall performance can be improved compared to using a single model. It can be combined with ensemble learning to enhance the performance of model transfer learning. It works by starting with a pre-trained model trained on a source task. The model has been trained to learn useful features from the source domain and then fine-tune the pre-trained model on the target task's data, adapting it to the specifics of the target task. Once done, multiple instances of the fine-tuned model are created by introducing minor variations, like random initialization of specific layers or dropout. Finally, these variants are ensembled by combining their predictions using techniques like bagging or boosting.

In our work, the output of each base classifier forpredicting cancer types is ensembled using the average method to compute the final desired output. As the Pan-Cancer data is highly imbalanced, as shown in Figure 2, we used the class weighting strategy to address this problem. Class weighting is the process of weight balancing. It is implemented by defining a dictionary with keys as labels and values as weights and then feeding the dictionary as a parameter. For example, if class j has a greater class weight than class j′, the gradients computed from samples of class j will be greater than those of j′, which affects the neural network training more than j′. The weights assigned to different classes in this work, as shown in Figure 2, are higher for minority classes and lower for majority classes. As depicted in the Figure, if class j′ has 100 times more samples than class j, then one sample of a class is as important as 100 samples of class j′; therefore, the class weight of j needs to be 100 times greater than that of j′. The magnitude of the gradients computed from j samples will be 100 timesmore extensive than that of j′.

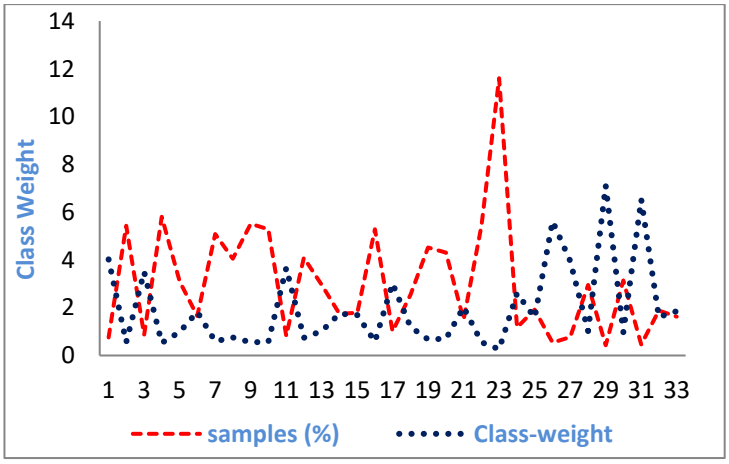

Figure 3.The magnitude of class weight values on major and minor classes in the ETL-CCP method

4.2 PERFORMANCE ANALYSIS

In this work, we conducted experiments on the Pan-Cancer dataset using advanced transfer-learning architectures such as DenseNet, ResNet50, Inception-V3, and Xception. Each model was trained and tested separately to assess its performance against the Pan-Cancer dataset. The experimental results are shown in Table 1.

Table1. Performance of the proposed ETL-CCP method and other deep learning methods for the classification of cancer types

Method	Accuracy	Precision	Recall	F-score
DenseNet	87.32	87.32	87.67	87.44
ResNet50	90.45	91.05	90.77	91.22
Inception-V3	93.10	93.53	93.05	93.15
Xception	92.54	93.76	93.45	93.71
ETL-CCP (ensembled method)	**96.88**	**96.94**	**96.90**	**96.93**

*Numbers in bold indicate the best performance

As shown in the Table-1, DenseNet achieved an accuracy of 87.32, while ResNet50, Inception-V3, and Xception methods attained 90.45%, 93.10%, and 92.54%, respectively. Considering the individual model performance, the Xception method performed better than other methods by achieving a precision of 93.76%, a recall of 93.45%, and an F1-score of 93.71%. All methods were combined in an ensemble to improve the classification performance further, while the average method calculated the final output. As shown in Table 1, the ensemble method (ETL-CCP) achieved the best accuracy of 96.89%. The Precision, recall, and F1-score achieved by the ETL-CCP method are 96.94%, 96.90%, and 96.93%, respectively.

While training the model, we conducted several experiments to find the best batch size for this problem. As shown in Figure 3, we achieved the maximum accuracy when the batch size was equal to 128. However, when batch size was further increased, it was observed that accuracy was reduced.

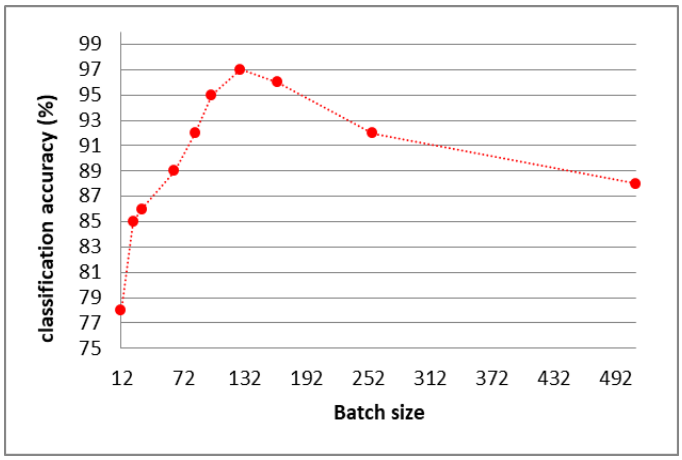

Figure 4. Performance (Accuracy) of our method on different batch sizes.

We used a confusion matrix to evaluate the method further and calculate a cross-tabulation of observed (true) and predicted classes (model). This gives us a holistic view of how well our classification model performs and what errors it makes. The confusion matrix of our ETL-CCP on the test dataset of Pan-Cancer data is shown in the Figure-4, where each column of the confusion matrix represents instances of the predicted class, and each row of the confusion matrix represents an instance of the actual class. It provides insight not only into the errors made by a classifier but also into the errors themselves.

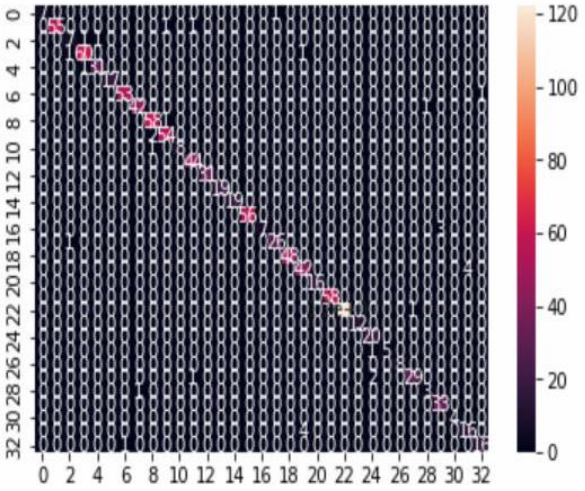

Figure 5.Confusion matrix of ETL-CCP method against the test dataset for Pan-Cancerclassification.

The class-wise performance of the ETL-CCP method is demonstrated in Table 2.

Table 2. Class-wise performance of the ETL-CCP method for Pan-Cancer Classification

Class	Precision	Recall	F1-Score
BRCA	1	0.88	0.93
LGG	1	0.96	0.98
UCEC	0.78	0.78	0.78
LUAD	0.97	0.97	0.97
HNSC	0.97	0.97	0.97
THCA	1	1	1
PRAD	0.98	0.98	0.98
LUSC	0.98	0.98	0.98
BLCA	0.95	0.98	0.97
STAD	0.96	0.96	0.96
SKCM	1	0.89	0.94
KIRC	0.96	1	0.98
LIHC	1	1	1
COAD	1	1	1
CESC	1	1	1
KIRP	1	1	1
SARC	1	0.7	0.82
OV	0.96	0.96	0.96
ESCA	1	1	1
PCPG	0.89	0.91	0.9
PAAD	1	1	1
TGCT	1	1	1
GBM	1	0.99	1
THYM	1	1	1
READ	0.87	1	0.93
LAML	1	0.83	0.91
MESO	1	1	1
UVM	0.97	0.91	0.94
ACC	0.75	0.75	0.75
KICH	0.92	1	0.96
UCS	1	1	1
DLBC	0.8	0.8	0.8

Class	Precision	Recall	F1-Score
CHOL	0.94	0.94	0.94

4.3 PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH EXISTING METHODS.

We compared the performance of our method with current baseline methods in terms of accuracy, precision, and recall. The classification performance of these methods and our proposed method is shown in Table 3.

Table 3. Performance comparison of the proposed method with existing baseline methods

Authors	Method	Accuracy	Precision	Recall
Karim et al. 2019	VGG16	96.25	96.25	95.42
Li et al. 2020	SNN	79.80	78.90	78.90
Kang et al. 2019	rL-GenSVM	87.29	87.73	87.29
Cava et al.2023	XGBoost	90.0	-	-
Hsu et al. 2018	ET-SVM	90.73	90.22	90.73
Y.Li et al. 2017	GA/KNN	95.60	-	-
Carson et al. 2019	CNN	95.65	95.55	95.69
Proposed method	**Ensemble method**	**96.88**	**96.94**	**96.90**

*Numbers in bold indicate the best performance

These baseline methods are based on machine learning or deep learning approaches and have used different techniques to handle unbalancing. Among all these methods, (Karim et al. 2019) have a higher performance result than other methods. It uses the VGG16 network to train the model for the classification of cancers. As depicted in the model, our model used transfer learning architectures in an ensemble and achieved better results than the baseline methods. The results indicate better predictions and the significance of the class weighting strategy for handling imbalanced data without data augmentation. We also compared the performance of our method in terms of class-wise precision and recall with one of the baseline methods proposed by (Karim et al. 2019) using CNN, as shown in Table 4. It can be seen that our model outperformed the baseline method in most of the classes, while the overall accuracy was reported to be better than CNN, as well as the VGG16 method.

Table 4.Class-wise Precision and Recall comparison between the proposed method and baseline methods

Cancer Type	Code	Proposed ETL-CCP		Existing method (CNN)	
		Precision	Recall	Precision	Recall
Breast invasive carcinoma	BRCA	**1**	0.88	0.8785	0.8612
Brain Lower Grade Glioma	LGG	**1**	**0.96**	0.9254	0.8926
Uterine Corpus Endometrial Carcinoma	UCEC	0.78	0.78	**0.8753**	**0.8819**
Lung adenocarcinoma	LUAD	**0.97**	**0.97**	0.8235	0.8354
Head and Neck squamous cell carcinoma	HNSC	**0.97**	**0.97**	0.852	0.8743
Thyroid carcinoma	THCA	**1**	**1**	0.8528	0.8323
Prostate adenocarcinoma	PRAD	**0.98**	**0.98**	0.8827	0.8778
Lung squamous cell carcinoma	LUSC	**0.98**	**0.98**	0.8726	0.8634
Bladder urothelial carcinoma	BLCA	**0.95**	**0.98**	0.8956	0.9037
Stomach adenocarcinoma	STAD	**0.96**	**0.96**	0.8253	0.8156
Skin Cutaneous Melanoma	SKCM	**1**	**0.89**	0.8853	0.8711
Kidney renal clear cell carcinoma	KIRC	**0.96**	**1**	0.8967	0.9123
Liver hepatocellular carcinoma	LIHC	**1**	**1**	0.8194	0.8085

Colon adenocarcinoma	COAD	**1**	**1**	0.8368	0.8245
Cervical and endocervical cancers	CESC	**1**	**1**	0.8785	0.8743
Kidney renal papillary cell carcinoma	KIRP	**1**	**1**	0.8254	0.8032
Sarcoma	SARC	**1**	0.7	0.8753	**0.867**1
Ovarian serous cystadenocarcinoma	OV	**0.96**	**0.96**	0.8825	0.8733
Esophagealcarcinoma	ESCA	**1**	**1**	0.8913	0.8719
Pheochromocytoma and Paraganglioma	PCPG	**0.89**	**0.91**	0.8537	0.8611
Pancreatic adenocarcinoma	PAAD	**1**	**1**	0.9629	0.9567
Testicular Germ Cell Tumors	TGCT	**1**	**1**	0.8736	0.8722
Glioblastoma multiforme	GBM	**1**	**0.99**	0.8952	0.8845
Thymoma	THYM	**1**	**1**	0.9255	0.9123
Rectum adenocarcinoma	READ	**0.87**	**1**	0.6795	0.6857
Acute Myeloid Leukemia	LAML	**1**	0.83	0.8697	**0.8567**
Mesothelioma	MESO	**1**	**1**	0.8991	0.9028
Uveal Melanoma	UVM	**0.97**	**0.91**	0.8765	0.8623
Adrenocortical carcinoma	ACC	0.75	0.75	**0.9217**	**0.9345**
Kidney Chromophobe	KICH	0.92	**1**	**0.9335**	0.9475
Uterine Carcinosarcoma	UCS	**1**	**1**	0.9157	0.9064
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	0.8	0.8	**0.8678**	**0.8729**
Cholangiocarcinoma	CHOL	**0.94**	**0.94**	0.8838	0.8975

*Numbers in bold indicate the best performance

5. CONCLUSION AND FUTURE SCOPE

The proposed work presented an Ensemble method based on advanced deep-learning architectures to improve the classification of 33 cancer types using high-dimensional Pan-Cancer data.The proposed approach combines the efficiency of DenseNet, ResNet, Inception, and Xception models to capture high-level features from structured input data. In addition to this, a class-weighting mechanism was used to overcome data imbalance issues. The experiments were conducted on a gene expression dataset comprising 10,267 cancer samples from 33 cancer types. Our method achieved a test data accuracy of 96.88%, outperforming current baseline methods.The ensemble method proposed in this work can be applied to other sequence-to-sequence and image classification problems. In the future, given the encouraging outcomes of using genomic data for cancer classification, we will analyse the underlying rules in our image-based deep-learning modelsfor critical gene identification and Pan-Cancer classification. For example, gene arrangement in the mutation map and labelledcolour may potentially be significant features that the deeplearning algorithm uses for categorization. Therefore, further research is required to fully understand those elements and develop an integrated model for precise cancer profiling.

DATA AVAILABILITY AND ACCESS

The dataset used and analyzed during the current study is available in The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, https://portal.gdc.cancer.gov. The data is publicly available and can be used by the deep learning community for the empirical analysis of deep learning algorithms.

REFERENCES

[1] Alsolami, B., Mehmood, R., &Albeshri, A. (2020). Hybrid statistical and machine learning methods for road traffic prediction: A review and tutorial. *Smart Infrastructure and Applications*, 115-133. https://doi.org/10.1007/978-3-030-13705-2_5.

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., &Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A cancer journal for clinicians*, *68*(6), 394-424. https://doi.org/10.3322/caac.21492.

[3] Cava, Claudia & Salvatore, Christian & Castiglioni, Isabella.(2023). Pan-Cancer Classification of Gene Expression Data Based on Artificial Neural Network Model.Appl Sciences.https://doi.org/10.3390/app13137355.

[4] Chicho, B. T., &Bibo Sallow, A. (2021).A Comprehensive Survey of Deep Learning Models Based on KerasFramework. *Journal of Soft Computing and Data Mining*, *2*(2),https://doi.org/10.30880/jscdm.2021.02.02.005

[5] Chollet, F. (2017).Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).https://doi.org/10.48550/arXiv.1610.02357.

[6] Cristovao, F., Cascianelli, S., Canakoglu, A., Carman, M., Nanni, L., Pinoli, P., &Masseroli, M. (2020).Investigating deep learning-based breast cancer subtyping using Pan-Cancer and multi-omic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.https://doi.org/10.1109/TCBB.2020.3042309.

[7] Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., &Balasubramanian, S. (2022).Early detection of cancer. *Science*, *375*(6586), eaay9040.https://doi.org/10.1126/science.aay9040.

[8] Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N. N.,&Madabhushi, A. (2017). Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Scientific reports*, *7*(1), 1-14. https://doi.org/10.1038/srep46450.

[9] Danaee, P., Ghaeini, R., & Hendrix, D. A. (2017).A deep learning approach for cancer detection and relevant gene identification.In *Pacific Symposium on Biocomputing 2017*. https://doi.org/10.1142/9789813207813_0022.

[10] Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060-1089. https://doi.org/10.1109/TASL.2013.2244083.

[11] Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., & Nagaraj, S. H. (2022). Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. *Cancers*, *14*(5), 1185.https://doi.org/10.3390/cancers14051185.

[12] Eltanbouly, S., Bashendy, M., AlNaimi, N., Chkirbene, Z., &Erbad, A. (2020, February). Machine learning techniques for network anomaly detection: A survey. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)* (pp. 156-162).IEEE.https://doi.org/10.1109/ICIoT48696.2020.9089465.

[13] Fatima, M., & Pasha, M. (2017).Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *9*(01) https://doi.org/ 10.4236/jilsa.2017.91001.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016).Deep residual learning for image recognition.In *Proceedings of the IEEE conference on computer vision and pattern recognition*.https://doi.org/10.48550/arXiv.1512.03385.

[15] Hsu, Y. H., & Si, D. (2018, July). Cancer type prediction and classification based on RNA-sequencing data. In *2018, the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5374-5377). IEEE.https://doi.org/10.1109/EMBC.2018.8513521.

[16] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017).Densely connected convolutional networks.In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).https://doi.org/10.1109/CVPR.2017.243.

[17] Joseph, M., Devaraj, M., & Leung, C. K. (2019, August).DeepGx: Deep learning using gene expression for cancer classification. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 913-920). IEEE.https://doi.org/10.1145/3341161.3343516.

[18] Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of theoretical biology*, *463*, 77-91.https://doi.org/10.1016/j.jtbi.2018.12.010

[19] Karim, M. R., Cochez, M., Beyan, O., Decker, S., & Lange, C. (2019, October).OncoNetExplainer: explainable predictions of cancer types based on gene expression data. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 415-422). IEEE.https://doi.org/10.1109/BIBE.2019.00081.

[20] Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A., &Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach. *IEEE Access*, *8*, 22874-22883. https://doi.org/10.1109/ACCESS.2020.2970210

[21] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015).Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.https://doi.org/10.1016/j.csbj.2014.11.005

[22] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A.,& Getz, G. (2013).Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214-218. https://doi.org/10.1038/nature12213.

[23] Li J, Xu Q, Wu M, Huang T and Wang Y (2020) Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection.Front. Bioeng.Biotechnol. 8:766. https://doi.org/10.3389/fbioe.2020.00766.

[24] Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., & Li, L. (2017).A comprehensive genomic Pan-Cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics*, *18*(1), 1-13. https://doi.org/10.1186/s12864-017-3906-0.

[25] Lyu, B., &Haque, A. (2018, August).Deep learning-based tumor type classification using gene expression data.In *Proceedings of the 2018 ACM International Conference on Bioinformatics, computational biology, and health informatics* (pp. 89-96).https://doi.org/10.1145/3233547.3233588.

[26] Mansour, R. F., Alfar, N. M., Abdel-Khalek, S., Abdelhaq, M., Saeed, R. A., &Alsaqour, R. (2022). Optimal deep learning-based fusion model for biomedical image classification. *Expert Systems*, *39*(3), e12764.https://doi.org/10.1111/exsy.12764.

[27] Mukherjee, A., Bisht, B., Dutta, S., & Paul, M. K. (2022). Current advances in the use of exosomes, liposomes, and bioengineered hybrid nanovesicles in cancer detection and therapy. *ActaPharmacologicaSinica*, 1-18. https://doi.org/10.1038/s41401-022-00902-w

[28] Okamoto, O. K. (2005). DNA microarrays in cancer diagnosis and prognosis. *Einstein*, *3*(1), 31-34.

[29] Pang, B., Nijkamp, E., & Wu, Y. N. (2020). Deep Learning with TensorFlow: A Review. Journal of Educational and Behavioral Statistics, 45(2), 227–248. https://doi.org/10.3102/1076998619872761.

[30] Podolsky, M. D., Barchuk, A. A., Kuznetcov, V. I., Gusarova, N. F., Gaidukov, V. S., &Tarakanov, S. A. (2016). Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific journal of cancer prevention*, *17*(2), 835-838.https://doi.org/10.7314/APJCP.2016.17.2.835.

[31] Sakri, S. B., Rashid, N. B. A., & Zain, Z. M. (2018). Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, *6*, 29637-29647. https://doi.org/10.1109/ACCESS.2018.2843443.

[32] Salehin, Imrus,&Dae-Ki Kang. 2023. "A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain" Electronics 12, no. 14: 3106. https://doi.org/10.3390/electronics12143106

[33] Sharma, A., & Rani, R. (2021).A systematic review of applications of machine learning in cancer prediction and diagnosis. *Archives of Computational Methods in Engineering*, *28*(7),https://doi.org/10.1007/s11831-021-09556-z

[34] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016). Rethinking the inception architecture for computer vision.In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).https://doi.org/10.48550/arXiv.1512.00567

[35] Wang, J., Dai, X., Luo, H., Yan, C., Zhang, G., & Luo, J. (2021).MI_DenseNetCAM: A Novel Pan-Cancer Classification and Prediction Method Based on Mutual Information and Deep Learning Model. *Frontiers in Genetics*, *12*, 670232.https://doi.org/10.3389/fgene.2021.670232

[36] Wang, Z., Jensen, M. A., &Zenklusen, J. C. (2016).A practical guide to the cancer genome atlas (TCGA).In *Statistical Genomics* (pp. 111-141). Humana Press, New York, NY. https://10.1007/978-1-4939-3578-9_6

[37] Way, G. P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W. K., Luna, A.,&Bodenheimer, T. (2018). Machine learning detects Pan-Cancer Ras Pathway activation in the Cancer Genome Atlas. *Cell Reports*, *23*(1), 172-180. https://doi.org/10.1016/j.celrep.2018.03.046

[38] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., & Stuart, J. M. (2013). Cancer Genome Atlas Research Network. *Nat Genet*, *45*(10), 1113-1120. https://doi.org/10.1038/ng.2764

[39] Wu, H., Liu, Q., & Liu, X. (2019). A review on deep learning approaches to image classification and object segmentation. *Comput.Mater.Contin*, *60*, 575-597. https://doi.org/10.32604/cmc.2019.03595

[40] Xue Ying.2019." An Overview of Overfitting and its Solutions", Journal of Physics, Conf. Series.https://doi.org/10.1088/1742-6596/1168/2/022022

[41] Yau, E. H., Kummetha, I. R., Lichinchi, G., Tang, R., Zhang, Y., & Rana, T. M. (2017). Genome-Wide CRISPR Screen for Essential Cell Growth Mediators in Mutant KRAS Colorectal CancersGenome-Wide CRISPR Screen of KRAS-Mutant Tumor Xenografts Cancer Research, 77(22), 6330-6339. https://doi.org/10.1158/0008-5472.can-17-2043.

[42] Yu Li, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, Xin Gao, Deep learning in bioinformatics: Introduction, application, and perspective in the big data era, Methods, Volume 166, 2019, Pages 4-21,https://doi.org/10.1016/j.ymeth.2019.04.008.

[43] Zhu, W., Xie, L., Han, J., &Guo, X. (2020).The application of deep learning in cancer prognosisprediction. *Cancers*, *12*(3), 603.https://doi.org/10.3390/cancers12030603

[44] Zhuang F et al., "A Comprehensive Survey on Transfer Learning," in Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, Jan. 2021, https://doi.org/10.1109/JPROC.2020.3004555

[45] Zuo, S., Dai, G., & Ren, X. (2019).Identification of a 6-gene signature predicting prognosis for colorectal cancer. *Cancer Cell International*, *19*(1), 1-15. https://doi.org/10.1186/s12935-018-0724-7.