



**IJIRCCCE**

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH


IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 3, March 2021

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

**Impact Factor: 7.488**

 9940 572 462

 6381 907 438

 [ijircce@gmail.com](mailto:ijircce@gmail.com)

 [www.ijircce.com](http://www.ijircce.com)

# Gene Expression Data based tumor Classification using Machine Learning Techniques

**Deepak S, Lakshmanakumar K, Pravin P, Nithin Venkatesh B**

Assistant Professor, Department of ECE, Sri Eshwar College of Engineering, Coimbatore, India

UG Student, Department of ECE, Sri Eshwar College of Engineering, Coimbatore, India

UG Student, Department of ECE, Sri Eshwar College of Engineering, Coimbatore, India

UG Student, Department of ECE, Sri Eshwar College of Engineering, Coimbatore, India

**ABSTRACT:** Cancer classification is important in the field of medical imaging. There are many successful outcomes in the field of classification based on the Diagnostic and treatment facilities. There are many existing algorithms based on RNA profiling and supervised machine learning to analyze the molecular-based classification techniques of various tumor classes. The major problem in the classification process is training a large number of data collected from the various database. To reduce the time consumption and increase the accuracy rate of the classification process, the proposed technique is based on the gene expression collected using microarray data which is trained using the deep learning techniques with the extracted features, and feature selection is carried out using Latent Feature Selection Technique. The deep learning technique such as Convolutional Neural Network used to classify the various tumor classes without labeling the classes. The dataset used for the training and testing process includes the data of lung cancer, renal cancer, and brain cancer data sets. The Convolutional Neural Network using the k-fold cross-validation technique achieves the accuracy rate of 96.4 3%. The proposed work includes the preprocessing and tuning techniques which increase the accuracy rate based on the gene expression in the multi-type cancer detection.

**KEYWORDS:** Tumor database, Bioinformatics, Deep Learning, Cancer Classification, Microarray gene expression, Latent Feature Selection Technique

## I. INTRODUCTION

The growth of the unwanted or abnormal tissue in the human body is represented as a cancer growth cell. Cancer is one of the toxic diseases in the present world which caused various abnormalities in the human body with different characteristics [1]. Classification of Cancer type can protect the patient survival rate many molecules or genetic analysis on how to identify the genetic alteration which various biological characters that help in the several treatments. Based on the analysis early diagnosis and can be improved by reducing the side effects under treatment. The characteristic of carcinogens is based on the analysis of impaired gene expression cell. Microarray is used to collect the gene expression data from the tumor cells which provide a large quantity of data for the research process. Dataset consists of thousands of different gene expression which has accurate feature values and provide an efficient way of analyzing machine learning and deep learning algorithms [2]. This dataset also includes the rules for analyzing the gene expression. They are various previous studies to demonstrate the deep learning and machine learning techniques in microarray gene expression to classify cancer type based on the feature selection and normative subset of gene expression. The proposed work includes the pre-processing stage which is a normalization that helps to save time conception during the training process. Feature Selection methods important in the gene expression data for cancer classification to reduce the dimensional representation. The reduced dimension can direct the biomarker strategy. The feature selection process can eliminate noisy features during the initial process [3]. By reducing the feature during the feature selection process can reduce the model complexity and model the biological features. In unsupervised learning, the selected gene features may improve the tumor profiles during the segmentation. The proposed work consists of the preprocessing stage which eliminates the noisy feature from the data set. The feature selection process is carried out using a Latent Feature Selection Technique that will eliminate noise data from the input data. The classifier which is Convolutional Neural Network consists of multiple kernels learning that depends on the subset of gene feature. This feature is selected from the auto-encoder process that is termed to be latent [4]. The feature extraction process includes the separation of tumor type and its subtypes based

on the training process. The main contribution of the proposed work includes the unsupervised method that can able to select genes from the clinical data with the high dimensional gene expression data without having tumor labels. Proposed unsupervised feature selection is applied to the gene expression data of lung cancer, renal cancer, and brain cancer data sets [5]. The Paper organized as follows. Section 2 demonstrates the related work whereas section 3 includes the material and methodology. Section 4 study of the proposed work section 4 includes the results and discussion of the proposed work. Section 6 includes the conclusion of the proposed work.

## II. STUDY OF RELATED WORK

There are several surveys related to the proposed work which is carried out by various researches. Ang et al concluded that the feature selection process can reduce the dimensionality of gene expression data [6]. This feature selection can improve the accuracy rate of the classifier whereas decreases the processing speed. Hinton and salakhutdinov enter the feature extraction technique which is carried out by a linear and nonlinear combination of the initial set of features [7]. This combined feature from the subset which is subjected to non-linear transforms. The data loss will be obtained by combining the linear and nonlinear processes. Good fellow et al proposed that feature extraction process to reduce dimensional space which is known as Latent space and Latent features respectively [8]. The classifiers are used in the process of the binary classifier that can classify two types of tumors. Chaudhry et al propose the auto-encoder process [9]. The auto-encoder process is widely used in many biomedical applications. The main use of auto-encoder is to make meaningful learning of gene expression data from various cancer patients to identify the stages and types of tumors. The main disadvantage of auto-encoders is the false negative rate tends to be higher during the training process. Way and Greene face members semantic mutation data in pancreas cancer is based on the auto-encoders. This will increase the segmentation performance of the cancer stage [10]. Moreover, Wang and Wang trained the DNA sequence of lung cancer to learn the meaningful characteristic of the supervised and unsupervised task [11]. Du et al proposed that multiple kernel learning for gene selection has supervised problems without improving the tumor type classification [12]. Zhang et al that collection of gene data can be applied to high dimensional RNA sequences while training by the support vector machine selection process [13]. Moon and Nakai proposed the classification of tumor cells on the renal layer [14]. Alelyani proposed the various unsupervised feature selection method for the clustering process to classify the tumor region [15]. The proposed technique is linked to the triple-negative breast cancer subtype based on the feature selection of molecular data. However, this technique is not accurate due to the unsupervised task during the training process. The proposed work describes the overcome of unsupervised learning during the feature selection. The multi-selection feature extraction process is described as an unsupervised training based on the microarray data. This technique is based on the weight of each feature obtained during the partition of data and the subset of features formed. However, the demerits of includes weight loss during the L1 normalization. In this work we propose – strategies of unsupervised learning during classification and feature selection based on the normalization of gene expression. The input dimension can be reduced based on the Latent feature extraction under the selection of gene expression features with the latent representation.

## III. MATERIALS AND METHODS

Machine learning techniques give the solution to the problem of a certain amount of data sets. Data can be subdivided into two groups namely the training data set and validation data set. The training data set is used to calibrate the parameters of some models created. The testing data set is used to test the trained features. This work proposes that the unsupervised learning method is known as the Latin structure method obtained from autoencoders. The proposed work is divided into three stages namely feature selection, training, and testing. Target kernel matrix ring samples based on the latent space. The selected features perform their classification process to classify the types of cancer.

## IV. DATASET

The data set consists of a collection of various patient data which is collected using microarray. The data set has a high dimensional gene expression data from various tumor profiles. The type of cancers includes lung cancer, renal cancer, and brain cancer. Data collected from the international cancer genome quotation portal. The lung cancer data set consists of squamous cell subtype and adenocarcinoma subtype cells with 480 and 430 tumor samples [16]. The brain cancer dataset consists of benign and malignant types with 160 and 439 samples respectively. The renal cancer data set consists of a papillary and clear cell with 230 and 620 tumor samples respectively. The data matrix of each data set is represented in  $U_n, d_m$  where  $n$  is the cancer samples  $d_m$  is equal to 18000 protein-coding genes. To select the subset,  $p$  should be lesser than  $d_m$  where  $p$  represents the gene normalization.

**Table 1 Datasets of Tumor Profiles**

| Type          | Subtype        | Patient samples |
|---------------|----------------|-----------------|
| Lung Dataset  | Squamous cell  | 480             |
|               | Adenocarcinoma | 430             |
| Renal Dataset | Papillary      | 230             |
|               | Clear cell     | 620             |
| Brain Dataset | Benign         | 160             |
|               | Malignant      | 439             |

Table 1 describes the dataset obtained from the various databases which consist of a collection of various patients recordings. The sample dataset consists of various patient samples with a normal stage as well as the infected stage.

## V. PREPARATION OF DATASET

The data can be divided into two groups for the experimental procedures. The first group includes the feature  $X_1$  group include sir classes  $X_2$ . The Matrix of the feature is organized as  $m \times n$  underclass  $n \times 1$  where the M represents the number of samples and n represents the number of genes in each class. The data set containing various samples is subdivided into training and validation data sets. The training samples contain 139 samples and the validation sample contains 36 samples. The initial calibration is processed using the training data set based on a deep learning algorithm. The hyperparameter was performed to the validation set and measures the accuracy of the algorithms. The accuracy of each algorithm is characterized by tuning the hyperparameter with the fold cross-validation to avoid overfitting values [17]. The data sets used in the proposed work is dimensionally higher. The direct use of data set can affect the accuracy and results based on reliability hand stability. The preprocessing stage includes the normalization technique to tackle this problem. The normalization can improve noise reduction and imperfect relevant characteristics during model training. The normalization which is carried using PCA technique is used to reduce the dimension from 12633 to 133 feature components.

### 1. Processing of Data

#### 1. Pre-processing: The pre-processing stages consists of the following stages

- The initial stage begins with the database selection process based on the response to biological question and hypothesis
- The second stage includes the Microarray Matrix experiment gene.
- The third stage includes the transformation of gene expression based on the data transformation with log base 2.

The data set contains the read count value of gene expression. The range of the gene expression vary and smaller than 1. So it is necessary to apply that transform using logarithmic equations.

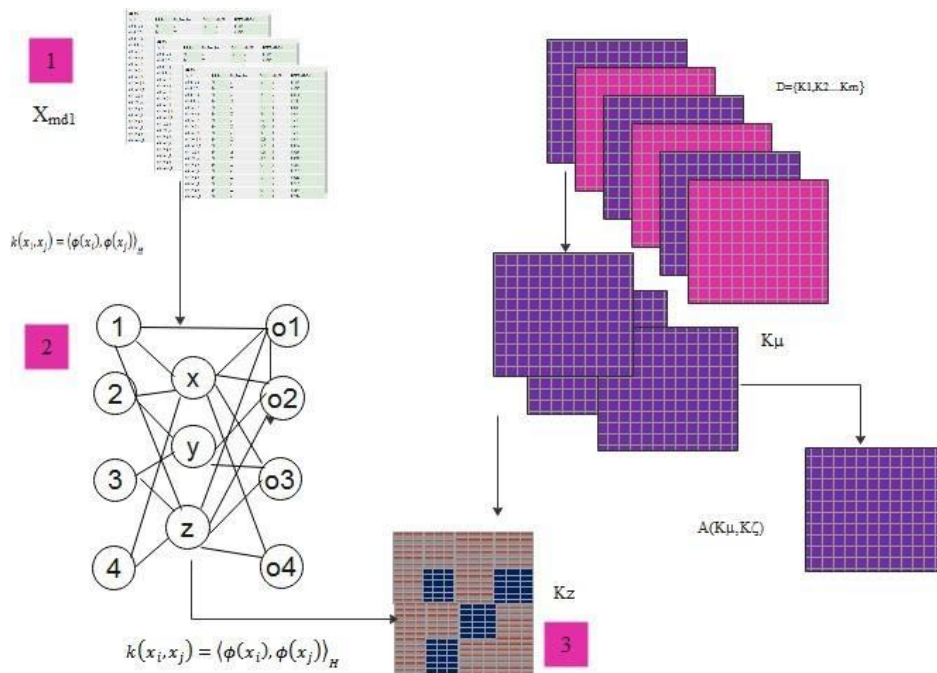
$$y = \log_2(x + 1) \quad (1)$$

Equation 1 will reduce the scaling expression based on the values which are lower than 1 to 0 to reduce noise. By using the normalization technique 1000 genetic expression which is found as noise added gene values is removed.  $x$  in equation 1 represents the gene features gathered from the microarray dataset. The threshold value will be equals to 1.19 using filtering techniques in the gene expression across the sample. The classification accuracy and integrity depends upon the threshold value. The lower threshold value retained more genes so that specific biomarkers value will not be eliminated. The normalized value is altered in the form of gene expression using microarray data. The third stage starts with the transformation to estimate the performance of the proposed work. The data is asymmetrical transformation is based on the features of their data sets. The feature selection process plays a major role in the



proposed work which is passed by the Latent Feature Selection Technique (LFST).

The proposed feature selection process is carried out using the latent feature selection technique which consists of m tumor samples the characteristic of d1 gene expression features in the matrix range of  $A_{md1}$ . Normalized features with lower range are represented in the form of p in the form of d1. The feature selection process starts with their trained feature using the autoencoder model with PCA feature extraction technique. The features extracted using the PCA is trained using Convolutional Neural Network to carry out the feature selection process. The latent space which is represented in the form of z dimension satisfies the condition  $l \ll d$ . The  $A_{md1}$  samples projected using a Latent space with the Gaussian kernel. The Gaussian kernel is represented as  $k_z$  and used as a target kernel.  $A_{md1}$  which is the matrix with the D1 set of features is built producing the kernel per feature. Finally, latent model how to reduce the subset of p kernels which is alternatively selected and combined to built  $K_\mu$  with increase arrangement in the matrix.



**Figure 1. A pipeline of the proposed method. First starting from the raw data (1) an autoencoder is trained (2) and a latent space learned. Then a  $K_z$  kernel built (3) using the sample set projected on the latent space.**

Figure 1 indicates a diagram of the proposed method. This approach results in a sparse solution in which the non-0 values of the  $\mu$  vector suggests the characteristic importance at the result. Features are selected via an unmanaged method that greatly aligns the representation found out from the autoencoder. We name this approach the Latent feature selection Technique (LFST).

## 2. Feature Selection Process

The Latent Feature Selection Technique consists of an optimization process with the normalized score features obtained. The feature selection process to select the best features is organized by an unsupervised approach. Given m dimensional space function value which is symmetric and semi-positive data to meets the condition.

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (2)$$



Where the function represents the  $\phi$  of the high dimensional data with the Hilbert space transform H. Hilbert space transform is defined as Head of the matrix dimension the function H operations in a pair of vectors in Hilbert space. The labeled samples can be illustrated in equation 3.

$$\phi: Y \mapsto \phi(Y) \in \mathcal{H} \quad (3)$$

Equation 3 represents the kernel Matrix obtained is defined as  $Y \times \phi$  the kernel function which can be thought of as information to the training phase.

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (4)$$

here n represents the number of features included in the gene samples. The kernel can be built with function obtained from the kernel matrix which gives the output.

$$A_{m \times n} = \langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j) \quad (6)$$

Equation 3 represents the matrix form of the collected features. The output represented as binary form when close to zero means, the samples are dissimilar whereas output close to 1 means Hilbert space area function. This kernel will be based on the radial basis function demonstrated below.

$$(6)$$

Various kernels K1 and K2 have a various number of samples along with the alignment of the target. To sample set consists of various features that can be training which consists of three normalized dataset features. Equation 7 represents the two sampled dataset values with normalized features. The alignment gives the similarities between two kernels with sample set M.

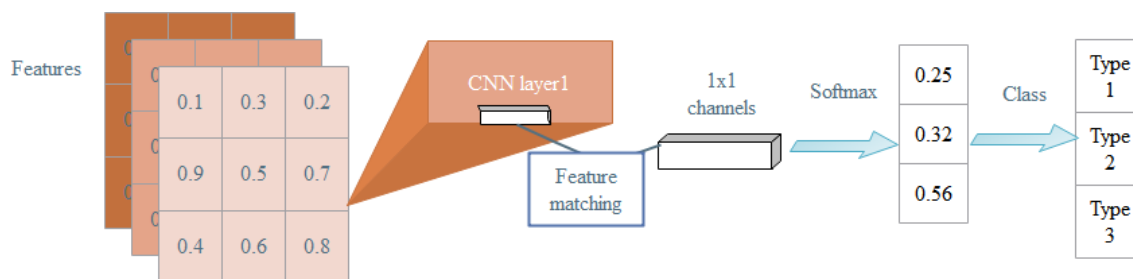
$$\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F} \quad (7)$$

The feature selection process is carried out using equation 8 which gives sort out the features that can be used for the training and testing phase of classifiers to analyze the various classes.

$$K_\mu(x, x') = \sum_{i=1}^n \mu_i K_i(x, x'), \mu_i \geq 0 \quad (8)$$

#### 4. Training and Testing Results

The training and testing phase includes the Convolution Neural Network Classifier which is based on the deep learning approach. CNN model proposed for the specific cancer type prediction. Each CNN model aims to specify the gene expression data models. The selected features or arranged and given as input to predict the results of classes. The design of CNN architecture includes the one layer Convolutional matrix with the Limited samples related to the number of parameters. CNN model consists of a one-dimensional kernel with two input vector [18].



**Figure 2 The Designed Architecture of CNN Model**

Figure 2 represents the designed architecture of the CNN model. The difference between the propose CNN model and the common CNN model is based on the time series prediction which represents the length of the size of the kernel. Initial kernel size is associated with the Global features with Convolutional value. The proposed design can capture selected features from gene expression. The activation function and max-pooling method are connected in a cascaded manner. The Convolutional module consists of a softmax layer to predict the various cancer types.

**Table 1 Input Parameters of the CNN model**

| Parameter                       | Best value |
|---------------------------------|------------|
|                                 | CNN        |
| Batch size                      | 10         |
| Epochs                          | 10         |
| Training optimization algorithm | SGD        |
| Learn rate                      | 0.1        |

|                               |               |
|-------------------------------|---------------|
| Momentum                      | 0             |
| Network weight initialization | Glorot_normal |
| Neuron activation function    | Linear        |
| Weight constraint             | 1             |
| Dropout regularization        | 0.4           |

The visualization of the CNN model is based on the Keras visualization package [19]. Gradient classes concerning the small changes in the gene expression give the result with higher accuracy. Perfect scores range from 0 to 1 where the maximum effect has 1 and no effect has 0 values. The gene features 2.6 and is marked as a cancer gene. The epoch and batch size as 60 and 128. All CNN models were trained with the various tumor samples initially. The training procedure is evaluated based on the over-fitting and the loss function [20]. The model achieves some losses which are tabulated below.



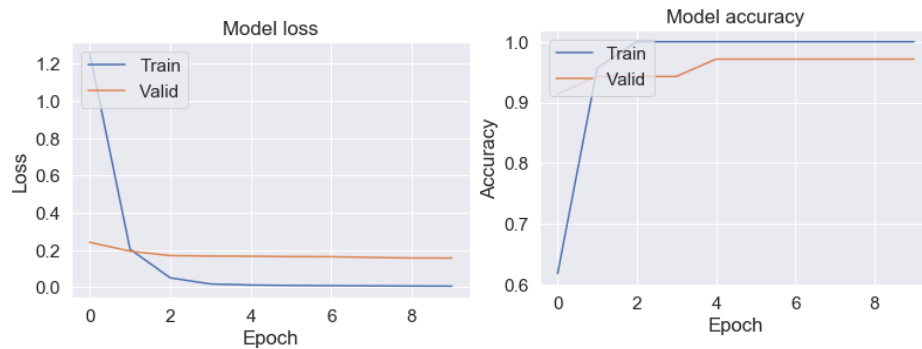
**Table 2 Different samples trained and testing from the datasets**

| Hyperparameters |         |    | Loss  |       |       |       |
|-----------------|---------|----|-------|-------|-------|-------|
| 64              | (1, 60) | 64 | 0.017 | 0.001 | 0.138 | 0.003 |
| 138             | (1, 30) | 8  | 0.030 | 0.013 | 0.147 | 0.003 |
| 138             | (1, 30) | 13 | 0.017 | 0.004 | 0.138 | 0.014 |
| 138             | (1, 30) | 33 | 0.019 | 0.010 | 0.131 | 0.009 |
| 138             | (1, 30) | 34 | 0.001 | 0.001 | 0.133 | 0.013 |
| 313             | (1, 30) | 13 | 0.010 | 0.001 | 0.137 | 0.003 |

| Hyperparameters |         |    | Loss  |       |       |       |
|-----------------|---------|----|-------|-------|-------|-------|
| 313             | (1, 30) | 33 | 0.138 | 0.009 | 0.333 | 0.130 |
| 313             | (1, 30) | 34 | 0.035 | 0.001 | 0.133 | 0.008 |

As the neural network nature depends on the training and testing phase with two-fold cross-validation that can be repeated for 3 times with mean and median location accuracy. The classification accuracy can proceed with 93.3,93.8, and 97.1 with an increase or decrease of 0.3%



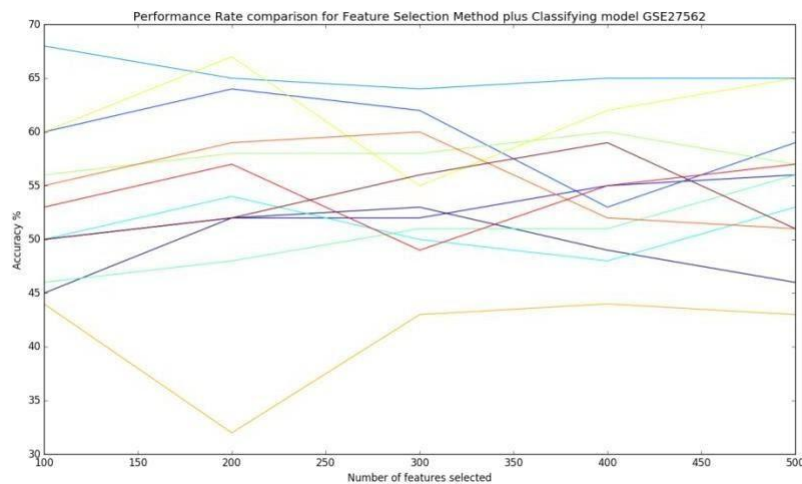


**Figure 3** Results obtained by CNN architecture in training using 10 epochs. (A) Lost value and (B) Accuracy. Lost function and accuracy is plotted on both training and validation datasets to observe behavior. When both datasets show very distant results, the architecture may be overfitting. As we have compared various algorithms to justify the accuracy rate and computational time of our proposed work which is illustrated in Table 3

**Table 3 Comparison of various algorithms**

| METHOD       | ACCURACY  | COMPUTATION TIME |
|--------------|-----------|------------------|
| Naive Bayes  | 91.176471 | 0.015888         |
| SVM          | 97.058824 | 0.024315         |
| ANN          | 98.352411 | 0.019689         |
| KNN          | 98.567814 | 0.018574         |
| PROPOSED CNN | 98.971596 | 0.0163581        |

The comparative study of features selection is shown in the figure 4. The features selection process decides the performance rate of classification model [19] [20].



**Figure 4 performance comparison based on feature selection**

## VI. CONCLUSION

The proposed work consists of feature selection which is known as the Latent feature selection technique that can reduce the subject features, unnecessary features from a large dimensional feature of gene expression data. The experimental results give the best feature selection algorithm output to achieve precise classification based on the

selected features. The classification feature gives an accuracy rate of 93.4 3% minimal loss. By the early classification process, cancer can be easily detected and diagnosed. This helps in the Discovery of selective drugs for the treatment of early diagnosis. The future work includes the consideration of genomic and metabolic features of various data sets.

## REFERENCES

1. Cox, T. R., & Erler, J. T. (2011). Remodeling and homeostasis of the extracellular matrix: implications for fibrotic diseases and cancer. *Disease models & mechanisms*, 4(2), 165-178.
2. Li, T., Zhang, C., & Ogiwara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15), 2429-2437.
3. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015.
4. Zhou, T., Thung, K. H., Zhu, X., & Shen, D. (2019). Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping*, 40(3), 1001-1016.
5. Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*, 2(4), e108.
6. Ang, A., Hodrick, R. J., Xing, Y., & Zhang, X. (2006). The cross-section of volatility and expected returns. *The Journal of Finance*, 61(1), 259-299.
7. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504-507.
8. Heaton, J. (2018). Ian goodfellow, yoshuabengio, and aaroncourville: Deep learning.
9. Chaudhry, B., Wang, J., Wu, S., Maglione, M., Mojica, W., Roth, E., ... & Shekelle, P. G. (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*, 144(10), 742-752.
10. Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., ... & Chakravarty, D. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2), 321-337.
11. S.Deepak,H.Annadakumar "AODV Route Discovery and route Maintenance in MANETs" 5<sup>th</sup> International Conference on Advanced Computing and Communication Systems (ICACCS),2019
12. R Pradeep R Kanimozhi "Brain Abnormalities Detection Using Improved Machine Learning Based Radial Basis Function Neural Network",2020
13. Wang, W., Wang, S., Ma, X., & Gong, J. (2011). Recent advances in catalytic hydrogenation of carbon dioxide. *Chemical Society Reviews*, 40(7), 3703-3727.
14. Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., ... & Yang, Z. (2019). Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes*, 10(3), 200.
15. Xu, H., Zhang, S., Yi, X., Plewczynski, D., & Li, M. J. (2020). Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction. *Computational and structural biotechnology journal*.
16. Moon, M., & Nakai, K. (2018). Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. *Journal of bioinformatics and computational biology*, 16(02), 1850006.
17. Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, 37.
18. Fukuoka, J., Dracheva, T., Shih, J. H., Hewitt, S. M., Fujii, T., Kishor, A., ... & Jen, J. (2007). Desmoglein 3 as a prognostic factor in lung cancer. *Human pathology*, 38(2), 276-283.
19. Tsamardinos, I., Rakhshani, A., & Lagani, V. (2015). Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. *International Journal on Artificial Intelligence Tools*, 24(05), 1540023.
20. Zeng, T., & Ji, S. (2015, November). Deep convolutional neural networks for multi-instance multi-task learning. In *2015 IEEE International Conference on Data Mining* (pp. 579-588). IEEE.
21. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
22. Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11, 2079-2107.



INNO  SPACE  
SJIF Scientific Journal Impact Factor

Impact Factor:  
7.488

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  [ijircce@gmail.com](mailto:ijircce@gmail.com)



[www.ijircce.com](http://www.ijircce.com)

Scan to save the contact details