The Seybold Report

ISSN 1533-9211

DOI 10.17605/OSF.IO/KVUB7

CLASSIFICATION OF KIDNEY CANCER DATA USING DEPTH AWARE GENERATIVE ADVERSARIAL NETWORKS APPROACH

N K SAKTHIVEL

Dean, Computing, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

S SUBASREE*

Professor & Head, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India. *Corresponding Author Email: drssubasree@gmail.com

S SIVAKUMAR

Assistant Professor (SG), Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

M MADAN MOHAN

Assistant Professor, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

Abstract

Advanced Biotechnology methods have led the generation to Large-Scale Bioinformatics and Gene Data and makes it important to analyze this data in Bioinformatics. This study analyses Gene Expression Data from 1157 kidney cancer patients to identify particular genes for prognosis. To overcome data instability, an end-to-end, depth-aware generative adversarial networks (DAGAN) approach including a loss function for the tasks of classification is proposed. The proposed approach combines the empirical wavelet transform (EWT) to rebuild the loss in non-linear Feature Extraction and neutral network for neutral categorization loss. Medical information and genome data are utilized to define the optimum classification method and to analyze the accuracy of classification through sample category, primary detection, tumor level, vital stage as risk factors. The result of this examination shows that the DAGAN is very effectual than the typical machine learning and the data mining strategies to predict kidney cancer prognosis using gene expression data. These findings have important implications for feature extraction from gene biomarkers to predict, prevent and early detect kidney cancer prognosis.

Keywords: Genomic Data, Deep Learning; Kidney Cancer; Bioinformatics

1. INTRODUCTION

Identification of gene is useful for cancer detection and prognosis prediction using bioinformatics approaches and to facilitate treatment by Jena, L[5]. The availability of large amount of gene expression data makes difficult to analyze the cancer data by Rukhsar, L.[1]. Some classification methods depends on extracted genes are developed, they can help in early identification along prognosis prediction. Generally, Gene modifications can cause cancer through enabling cells to proliferate exponentially, permeating normal surrounding cells, and spread all over the body. Deep learning methods are used in previous studies to predict patients' disease condition through the analysis of gene sequence mutations at Spinal Muscular Atrophy, heredity non-polyposis colon cancer & autism by Shao, D.[2]. It combines gene expression

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/KVUB7

with medical data from kidney cancer patients via Cancer Genome Atlas utilizing proposed approach by Gong, P.[3]. Here, Typical Data Mining and Machine Learning is comparing with proposed approach involves two steps; Feature Engineering and over-and under-sampling. Extraction of deep features in Gene Biomarkers Kakati, T., Bhattacharyya [4] through detection, separates disease data and enhance end-to-end prediction mode through the comparison and analysis of classification approaches depends upon extracted genes. This paper has three main contributions:

- An end-to-end approach has proposed to predict kidney cancer like sample type samples, primary detection, tumor level, and significant status.
- Non-linear Transformation Technique and Empirical Wavelet Transform are introduced for the extraction of deep features in Gene Biomarkers.
- Mixed loss function of deep learning method is proposed that consider compression of knowledge representation as well as data imbalance issues.

This study is arranged as follows: the review of literature is presented in Section 2. Proposed method explained under Segment 3. Outcomes demonstrated in Segment 4. Segment 5 presents the conclusion.

2. LITERATURE REVIEW

Shon et al. [6] presented a paper in 2020 titled "Classification of Kidney Cancer Data Using Cost-Sensitive Hybrid Deep Learning Approach". An end-to-end Cost-Sensitive Hybrid Deep Learning (COST-HDL) used Cost-Sensitive Loss to divide the work on unstable cancer data. The author added deep symmetric auto-encoder and decoder which was symmetric to encoder on the basis of framework of layer, with loss in reconstruction of non-linear feature extraction, and Neural Network balanced categorization loss of diagnosis point out data imbalance issues. This approach referred as CKCD-COST-HDL.

Jena et al. [7] presented a paper in 2021 titled "Risk Prediction of Kidney Disease Using Machine Learning Strategies." That paper highlights the usage of classification techniques in the field of bioinformatics, for the prediction of chronic diseases, which was a significant challenge for medical experts. The authors developed a disease prediction method carry out several Machine Learning Classification strategies, which was referred as CKCD-GSA.

Kim et al. [8] presented a paper in 2020 titled "Cancer Classification of Single-Cell Gene Expression Data by Neural Network". Research described development of cancer classifiers and identification of twenty one type cancerous and normal tissues on the basis of bulk RNA-sequences and scRNA-sequences data. The authors trained the classifiers using seven thousand three hundred and ninety eight cancer specimen and six hundred and forty normal specimens through twenty one tumors and normal tissues in TCGA. The training was done according to three hundred most remarked genes expressed in every type of cancer.

A paper by Ahsan et al. [9], published in 2021, introduced a cancer classification approach that used miRNA genome data and deep learning. The authors presented two new architectures, a

ISSN 1533-9211

DOI 10.17605/OSF.IO/KVUB7

basic ANN and a novel architecture on the basis of ResNet called CResNet, to classify all types of cancer by Subasree, [13]. The study have trained four different types of models, including LSTM, Artificial Neural Network, CResNet, and Ensemble models used model averaging.

3. PROPOSED METHODOLOGY

The Gene Expression data has collected from people having kidney cancer and DL method is proposed. Figure 1 illustrates the total workflow of the suggested approach.

3.1 Dataset

Cancer Genome Atlas (TCGA) [10] is a database contains various gene information includes Single-Nucleotide Polymorphism (SNP) and Genome Expressions of numerous people. TCGA data of 1157 kidney cancer patients is extracted, with medical information like sample category, primary detection, tumor level, and vital stage. For the prognosis prediction task, these medical details were served as class labels. After assigning transaction IDs, RNA-level gene expression was measured, and the expressions were digitized. Here, 60,483 Gene Expression data is used, and values denoted in Fragments /Kilo base /Million mapped. This data was utilized for extracting big design gene biomarkers to receive exactness in classification and issues on the basis of type sample, primary detection, tumor level, and significant status indicates level of kidney cancer. Eliminating non- variance Gene Expression data and disturbing specimen are done at preprocessing step. For the prognoses, we have used various samples and gene expression data, to classify them to 80% for training and 20% for testing. The dataset are unbalanced, specifically sample category prognosis, shows primary tumor specimen at 87.9% and normal tissue specimens at12.1%. Figure one illustrates overall workflow of proposed methodology, which is discussed in detail below.

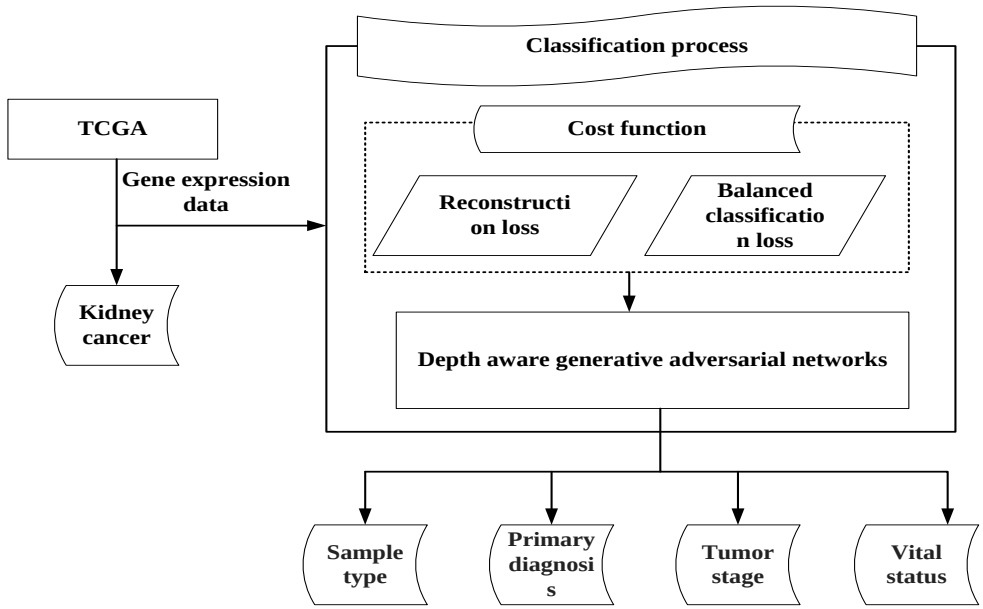

Figure 1: Block diagram for proposed CKCD-DWGAN Methodology

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/KVUB7

3.2 The depth aware generative adversarial networks

The analysis conducted to evaluate the performance of classification analysis on extracted target genes only. The CKCD-DWGAN approach, as shown in Figure 1, Gene Expression data at TCGA site and the result includes type sample, primary detection, tumor level, and significant status. The approach uses DWGAN models [11]. The traditional classification analysis is technically difficult due to the availability of higher number of variables than samples in RNA sequencing data, when working with more than sixty thousand variables, it is really difficult to introduce Data Mining and Machine Learning approaches directly to data-set. The work utilized a preprocessing step to reduce noise and eliminate non-variance gene expression data, followed by the application of the proposed CKCD-DWGAN approach.

A generator and a discriminator were included in our proposed DAGAN approach. The design of our differentiator is directly attracted by FOMM. This approach involves optimization at the network utilizing existing training face videos. mainly, we have started with two continues video frames, I_i, I_{i+1} from face video, I_{i+1} the original image and I_i goal image, Our goal is understanding various geometric features, such as the depth map D_{I_i} for goal image I_i , a camera intrinsic matrix K_n, n indicate n^{th} input video of training, and close camera pose $D_{I_i} \to I_{i+1}$ with translation $t_{I_i} \to I_{i+1}$ between the two images. It is necessary to highlight that utilizing provided camera intrinsic K_n, the camera intrinsic unavailable in training face video dataset. K_n, learns the input-video-clip-specific camera intrinsic in order to account for the fact that each face video can potentially captured through a different camcoder. So, only video frames are needed as input for our approach.

The depth network is responsible for the generation of depth map D_{I_i}. On the other hand, the same pose network $f_d(.)$ predicts pose $D_{I_i} \to I_{i+1}$, translation $t_{I_i} \to I_{i+1}$, and camera intrinsic matrix K_n is given below:

$$D_{I_i} = F_d(I_i) \tag{1}$$

$$[R_{I_i} \to I_{i+1}, t_{I_i} \to I_{i+1}], K_n = F_p(I_i \parallel I_{t+1}) \tag{2}$$

Target image I_i can cover look of original image I_{i+1} using concatenation denoted by || symbol. The DWGAN approach shows loss of reconstruction and balanced categorization loss as Cost function. Next paragraph explains empirical wavelet transform (EWT) [12] model for the extraction of features in Gene Biomarkers and NN model to construct prognosis models.

3.3 Extracting Deep Features from Gene Biomarkers

We have employed EWT non-linear feature transformation strategies for the extraction of Gene Expression data from training dataset compared Principal Component Analysis, Linear Feature Transformation & Least Absolute Shrinkage and Selection Operator (LASSO) extraction

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/KVUB7

approaches. The aim of PCA is to reduce the number of linearly correlated variables in correlated multivariate data through explaining small number of linearly unrelated variables combined to actual variable. Because of its linear constraints, EWT model with non-linear activation functions developed and it provides higher accuracy in the reconstruction of data. Therefore, during the extraction of a complex structure of cancer data, PCA and LASSO may indicates some important information, which leads to the loss of valuable insights.

The extraction of different manner by creating adaptive wavelets is the main goal of EWT. The process involves the following steps. First, the signal $f(t)$ is subjected to FFT where $f(t)$ is a discrete signal, $t=\{t_i\}\ i=1,2,...M$ and M depicts count of sample. This produces frequency spectrum $X(W)$ is analyzed to get set of maxima $M=\{M_i\}\ i=1,2,...N$ and their similar frequencies $w=\{w_i\}\ i=1,2,...N$. The count of maxima N, and the count of filter banks implemented at this stage. Next, exact division of the Fourier spectrum is obtained, boundaries Ω_i of segment, denoted midpoint between 2 continuous maxima.

$$\Omega_i = \frac{w_i + w_{i+1}}{2} \tag{3}$$

Where, a set of boundaries, denoted by $\Omega=\{\Omega_i\}\ i=1,2,...N-1$, is given and two frequencies, w_i and w_{i+1}, are defined. Bank of N wavelet filters is constructed and comprising less-pass filter and multi band-pass filters according to boundaries in B.EWT uses a loss function for handling data reconstruction errors, which quantifies the difference among original data and its reconstructed counterpart. Specifically, Mean Squared Error (MSE) worked as loss function.

3.4 Constructing Prognosis Prediction Models

The developed prognosis prediction models by us contains 1-input layer, 1-hidden level with 100 nodes &1-output layer. For extracting Deep Features served as input for the NN model, the hidden EWT model was used. The NN worked as a loss function to clear classification errors and address class imbalance by measuring the difference among true class and predicted class. Focal loss used as the NN's loss function, remolding the loss of standard cross-entropy to down-weight loss allocated to categorize examples and notes the class balance problem.

4. RESULTS AND DISCUSSION

We employed the EWT models on training set and compared its performance to that of PCA dimension reduction methods, to assess and extract deep features in Gene Biomarkers. We have selected100 features per classification task using the EWT model for further analysis. And we selected100 features per classification using PCA model to ensure a fair comparison. Subsequently, evaluated feature extraction of gene biomarkers on testing set.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/KVUB7

4.1 Evaluation of Prognose Prediction Models

The performance of CKCD-DWGAN approach in classification, we have used four indices, namely accuracy, computational time, and ROC. The definitions of these indices are given below.

4.1.1 Accuracy

The accuracy index shows the proportion of samples that are exactly classified as normal, and it can calculate by eqn (4)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

4.2.2 AUC

Equation (5) provides the expression for the AUC index.

$$AUC = 0.5 \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{5}$$

The simulation results of our proposed CKCD-DWGAN method is shown in Figure 2-4. Comparing the performance of suggested method with two approaches: The Cost-Sensitive Hybrid Deep Learning algorithm to classify kidney cancer data (CKCD-COST-HDL) [6] and machine learning strategies for predicting kidney disease risk (CKCD-GSA) [7].

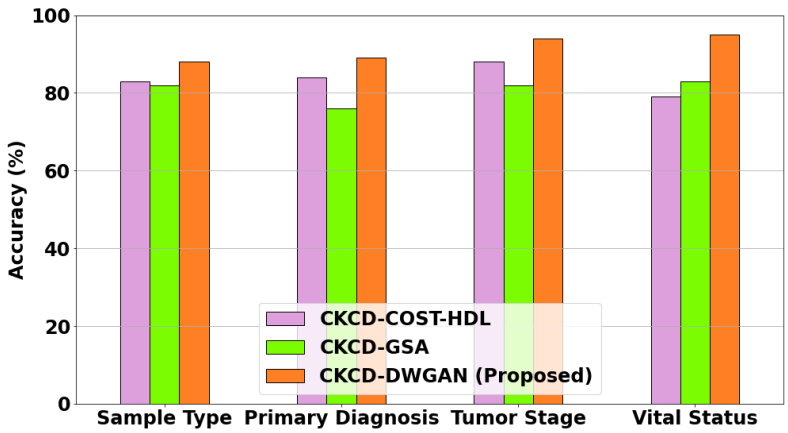

Figure 2: Performance of accuracy analysis

Accurate results through the analysis are presented in Figure 2, shows that our proposed CKCD-DWGAN method achieved higher accuracy when we compared to existing methods like CKCD-COST-HDL and CKCD-GSA. Especially for Sample Type, the proposed method attains 4.94% and 6.16% higher accuracy; for Primary Diagnosis, it achieves 10.51% and 7.23% higher accuracy; for tumor Stage, it attains 2.31% and 8.47% higher accuracy; and for Vital Status, it gain 3.88% and 5.75% higher accuracy compared to CKCD-COST-HDL and CKCD-GSA, respectively.

ISSN 1533-9211

DOI 10.17605/OSF.IO/KVUB7

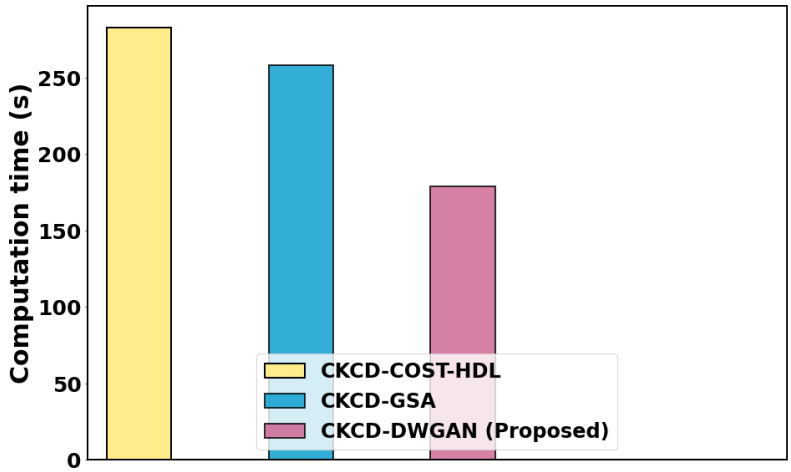

Figure 3: Performance of computational time analysis

Figure 3 shows the computational time analysis. Here, the proposed CKCD-DWGAN method attains 7.45% and 10.18% lower computational time as compared with existing methods like CKCD-COST-HDL and CKCD-GSA respectively

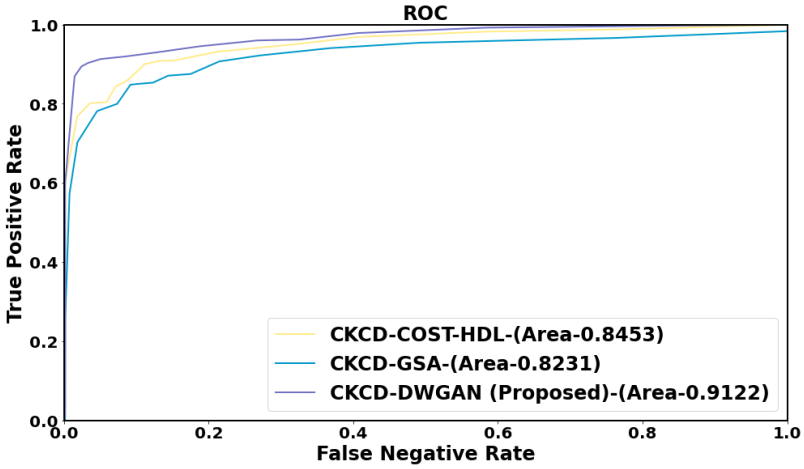

Figure 4: Analysis of RoC

"Figure 4 displays the RoC analysis results, which demonstrate that our proposed CKCD-DWGAN method achieves significantly higher AUC compared to existing methods such as CKCD-COST-HDL and CKCD-GSA. Specifically, the proposed method attains 1.551% and 3.915% higher AUC compared to CKCD-COST-HDL and CKCD-GSA, respectively."

5. CONCLUSION

This study describes the effectiveness of unsupervised non-linear EWT which act as a model for the extraction of features in Gene Expression data. The characteristics are connected with kidney cancer diagnosis, like type sample, primary detection, tumor level, and significant

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/KVUB7

status. An End-to-end DL frame work proposed is highly effective than Traditional Machine Learning approach. While comparing CKCD-DWGAN approach with traditional approaches, we can find that it got better output for prognosis in Gene Expression data. Feature extracted through EWT method differentiated better than features extracted by other methods on training and testing sets. Also, EWT model identified another class label. Our findings are really useful and can apply for feature extraction in Genome Biomarkers to diagnose Kidney Cancer due to several problems and contribute to prevent kidney cancer and perform early detection.

References

1. Rukhsar, L., Bangyal, W.H., Ali Khan, M.S., Ag Ibrahim, A.A., Nisar, K. and Rawat, D.B., 2022. Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification. Applied Sciences, 12(4), p.1850.
2. Shao, D., Huang, L., Wang, Y., He, K., Cui, X., Wang, Y., Ma, Q. and Cui, J., 2022. DeepSec: a deep learning framework for secreted protein discovery in human body fluids. Bioinformatics, 38(1), pp.228-235.
3. Gong, P., Cheng, L., Zhang, Z., Meng, A., Li, E., Chen, J. and Zhang, L., 2023. Multi-omics integration method based on attention deep learning network for biomedical data classification. Computer Methods and Programs in Biomedicine, 231, p.107377.
4. Kakati, T., Bhattacharyya, D.K., Kalita, J.K. and Norden-Krichmar, T.M., 2022. DEGnext: classification of differentially expressed genes from RNA-seq data using a convolutional neural network with transfer learning. BMC bioinformatics, 23(1), p.17.
5. Jena, L., Nayak, S. and Swain, R., 2020. Chronic disease risk (CDR) prediction in biomedical data using machine learning approach. In Advances in Intelligent Computing and Communication: Proceedings of ICAC 2019 (pp. 232-239). Springer Singapore.
6. Shon, H.S., Batbaatar, E., Kim, K.O., Cha, E.J. and Kim, K.A., 2020. Classification of kidney cancer data using cost-sensitive hybrid deep learning approach. Symmetry, 12(1), p.154.
7. Jena, L., Patra, B., Nayak, S., Mishra, S. and Tripathy, S., 2021. Risk prediction of kidney disease using machine learning strategies. In Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 2 (pp. 485-494). Springer Singapore.
8. Kim, B.H., Yu, K. and Lee, P.C., 2020. Cancer classification of single-cell gene expression data by neural network. Bioinformatics, 36(5), pp.1360-1366.
9. Ahsan, A.O., Ansar, M., Bin, A. and Minhaj, M.I., 2021. Cancer Classification with Deep Learning using Genomics Data Islamic University of Technology (IUT), Board Bazar, Gazipur-1704, Bangladesh).
10. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
11. Hong, F.T., Zhang, L., Shen, L. and Xu, D., 2022. Depth-aware generative adversarial network for talking head video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3397-3406).
12. Liu, W. and Chen, W., 2019. Recent advancements in empirical wavelet transform and its applications. IEEE Access, 7, pp.103770-103780.
13. S Subasree, NK Sakthivel, VR Balasaraswathi, AK Tyagi, 2022. Selection of Optimal Thresholds in Multi-Level Thresholding Using Multi-Objective Emperor Penguin Optimization for Precise Segmentation of Mammogram Images, Journal of Circuits, Systems and Computers 31 (07), 2250131