The Seybold Report

ISSN 1533-9211

DOI: 10.5281/zenodo.8366457

BIOINFORMATICS ANALYSIS OF THE GENES INVOLVED IN THE EXTENSION OF PROSTATE CANCER TO ADJACENT LYMPH NODES USING ADAPTIVE ACTIVATION FUNCTIONS WITH DEEP KRONECKER NEURAL NETWORK

S. SUBASREE

Professor and Head, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, (Autonomous), Coimbatore 641 105, TN, India.

N. K. SAKTHIVEL *

Dean (Computing), Nehru Institute of Engineering and Technology, (Autonomous), Coimbatore 641 105, TN, India. *Corresponding Author Email: nksakthivel@gmail.com

S. PRIYA

Assistant Professor (SG), Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, (Autonomous), Coimbatore 641 105, TN, India.

M. MAHABOOBA

Assistant Professor, Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, (Autonomous), Coimbatore 641 105, TN, India.

Abstract

The objective of this research was to discover the genes involved in prostate cancer (PCa) patients' participation in extra lymph nodes and offer useful insights for identifying possible diagnostic biomarkers and pathogenic genes in PCa metastasis. Using Adaptive activation functions with deep Kronecker neural network (AAF-DKNN) and PCA with or without down sampling, the most significant candidate genes were determined. In total, twenty one genes were identified as related to the lymph node involvement. Between these, 9 genes were observed in metastatic prostate cancer, 6 were identified in another metastatic cancer and 4 were found in another local cancer. Additionally, augmentation of candidate genes was assessed in another PCa data sets. A verified set of genes that contribute to PCa metastasis was also identified. The SPAG1 and PLEKHF2 gene amplification was linked with a reduced chance of survival in prostate cancer patients.

Keywords: Gene Expression Analysis, Metastasis Prostate Cancer, Adaptive Activation Functions With Deep Kronecker Neural Network.

1. INTRODUCTION

Prostate cancer is accountable for 3rd highest count of cancer-associated deaths [1]. In fewer than 10% of prostate cancer cases, a positive family history is observed [2]. It is challenging to treat aggressive as well as metastatic forms of prostate cancer. The androgen receptor and TP53 are the most frequently impacted genes in patients with metastatic PCa who frequently exhibit somatic genomic changes [3]. Additionally, the tumour suppressor gene PTEN and the transcriptional regulator ETS have changed, albeit to a lesser degree. Advanced genetic analysis shows promise in identifying genetic alterations and discovering prognostic or

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

predictive biomarkers. This can lead to the development of various options to combat metastatic PCa and increase the quality of patient's life [4]. Moreover, advanced genetic analysis may identify patients in higher danger of developing metastatic prostate cancer and enable a better understanding of pathologic procedure [5].

The key contribution of this manuscript can be summarized as below:

- A comprehensive estimation of an mRNA expression data set has been conducted to identify verified genes with differential expression between cases as well as spread to nearby lymph nodes (N1) and those with a local category (N0).
- Deep learning technique AAF-DKNN was utilized. The data analysis methods employed in this study has demonstrated their usefulness in various cancer biology fields and has taken benefits of abundance of higher throughput experimental gene analysis data. These methods have also been utilized for survival analysis, categorization, clustering tasks, such as diagnosis and mechanistic analyses.

Remaining manuscript is structured as: Segment 2 analyses related works, Segment 3 describes proposed method, Segment 4 presents outcomes and discussions, and Segment 5 provides conclusion.

2. RELATED WORKS

This section covers the most recent research on bioinformatics analysis of genes implicated in predicting prostate cancer. A bioinformatics analysis of genes associated with prostate cancer spread to nearby lymph nodes was conducted through Shamsara et al. in 2020 [6]. The study utilized to identify the most significant candidate genes such as: (i) supervised (ii) unsupervised machine learning techniques. Among the machine learning methods employed were K-means clustering, neural networks, Nave Bayesian categorizations and PCa with or without down sampling.

Lai et al. presented an integrative bioinformatics analysis in 2022 [7], aiming to identify a gene signature linked to steroid hormone pathways that can predict the prognosis of PCa. The goal of this investigation was to detect core genes that regulate steroid hormone pathways and relate them to PCa progression. Genes associated with steroid hormones were chosen from operational data bases such (i) Gene Ontology (ii) KEGG (iii) Reactome. From TCGA, gene expression profiles and clinical data for PCa patients were acquired to investigate genes related to the steroid hormones. The study employed a machine learning approach to perform selection of main feature [13, 14].

Chen et al. published a paper in 2022 [8], which aimed to identify hub genes that could predict the growth of benign prostate hyperplasia to PCa and analyse their clinical value in PCa using bioinformatics evaluation. The differentially expressed genes among BPH and Prostate Cancer were discovered based on Gene Expression Omnibus (GEO) database. To identify pathways enriched with DEGs, analyses of Gene Ontology and Kyoto Encyclopaedia of Genes, Genomes data bases was activated. The STRING database was utilized to offer protein to protein

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

interaction network then identify centre genes on the network. To estimate the medical value of centre genes in Prostate Cancer 'R' software is used. The operation of these centre genes were experimented at various data bases, medical samples, Prostate Cancer cells.

In 2020, Wang and colleagues [9] conducted a bioinformatics analysis in multiple databases to identify core genes related to the PCa progression and result. The study utilized GEO2R in the Gene Expression Omnibus database to analyse variation in GSE38241, GSE69223, GSE46602, and GSE104749. Operation enrichment was assessed through DAVID 6.8, while, PPI network as well as important component was estimated through Cytoscape, STRING, and MCODE.GO. Twenty potential genes were found to be tightly associated to mitosis, cell division, cell cycle phases, and p53 signalling pathway by pathway analysis. In GSE21032 and TCGA PRAD, 6 independent prognostic variables were detected.

3. PROPOSED METHODOLOGY

This section focuses on the BI-PC-AAF-DKNN. The proposed method is depicted in Figure 1. The details of the BI-PC-AAF-DKNN approach are discussed below.

3.1 Dataset preparation

A Prostate Adenocarcinoma dataset [10] from the TCGA PanCancer Atlas was obtained through cBioPortal. The RNA expression values were normalized against a reference population's gene expression distribution and informed into log2 values, while the CNA data was informed into -2, -1, 0, +1, or +2. RNA data was originally analyzed, and CNA data was used for further validation. Samples were classified as N1 or N0, with N1 containing samples from PCa patients that have lymph node involvement and N0 containing models from prostate cancer patients without lymph node involvement. NA samples were excluded from the analysis.

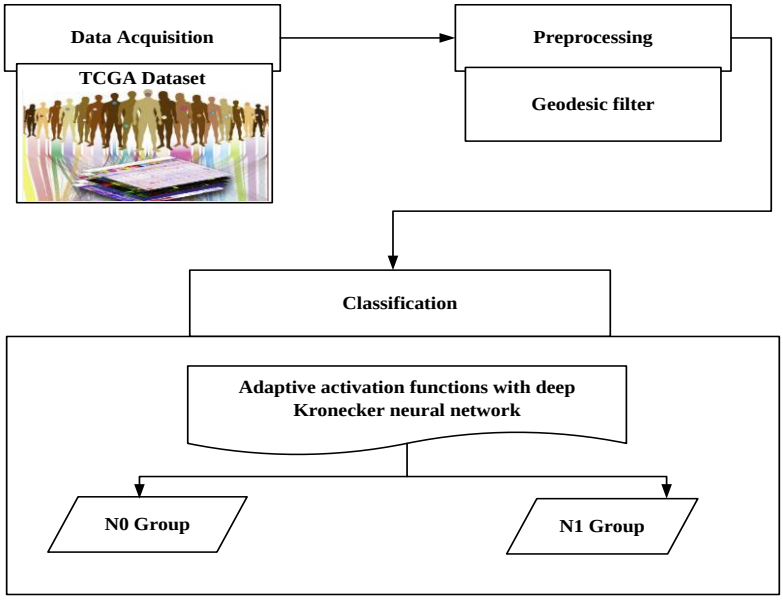

Figure 1: Overall block diagram of proposed BI-PC-AAF-DKNN methodology

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

3.2 Preprocessing Using Geodesic filter

Geodesic filter (GF) [11] is used for pre-processing to remove variables that have zero values for all instances in the input data. Genes that have '0' variance or high correlation with another gene (with a 0.9 correlation threshold) were also eliminated. The dataset was randomly split into test as well as training set of 80:20 ratios. ANOVA F-values were calculated to identify differentially expressed genes amongst N1 and N0 groups are used for selection of variable. Using ANOVA F-values, different datasets are generated as well as estimated.

Equation (1) represents the flow of conditional likelihood density operation on state vector, X regarding a parameter λ during the pre-processing step using Geodesic filter (GF).

$$\log P(X,\lambda) = \log g(x) + \lambda \log h(x) - \log K(\lambda) \quad (1)$$

The pre-processing using Geodesic filter (GF) involves a flow of conditional likelihood density operation on a state vector regarding parameter. This is expressed as equation (1), here $g(x)$ represents previous density, $h(x)$denotes probability, $K(\lambda)$ indicates normalization. As λ moves from 0 to 1, conditional likelihood $P(X,\lambda)$moves among prior and posterior. Unlike standard particle filtering, particles are flowed from prior position to posterior position based on measurements. Moreover, algorithm assumes that the particle flow from prior to posterior adheres with Ito stochastic differential equation, which is given by equation (2).

$$dX = f(X,\lambda)d\lambda + Q(X,\lambda)dW \quad (2)$$

After the pre-processing stage using Geodesic filter (GF), the data is transformed based on the flow function $f(X,\lambda)$ and Wiener W process along with the diffusion matrix $Q(X,\lambda)$. The resultant pre-processed data is then used for the classification stage.

3.3 Prostate Cancer Classification using AAF-DKNN

The AAF-DKNN, which stands for adaptive activation function with deep Kronecker neural network, is a deep learning method that can differentiate groups without supervision and can also serve as a supervised classification technique. It has the ability to calculate the weights of variables for group separation. In this research, AAF-DKNN was implemented to determine the variations in gene expression levels amongst the N1, N0 groups. The proposed methodology was employed to analyse the gene expression patterns that distinguish the N0 and N1 groups. The size of the adaptive activation function with deep Kronecker neural network (AAF-DKNN) [12] is times k larger compared to feed forward networks. Conversely, entire count of parameters only varies through $2K(D-1)$ owing to use of Kronecker product. Moreover, Kronecker networks are regarded into new kind of neural networks, which extend a category of existing feed-forward neural networks, specifically by utilizing adaptive activation functions. Kronecker networks are set up for supervised learning with square loss operation. The square loss is delineated through equation (3) for a $T_n = \{(x_i, y_i)\}_{i=1}^{n}$ set of n-training data points.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

$$L(\Im) = \frac{1}{2}\left(u_{\Im}^{type}(x_i) - y_i\right)^2 \quad (3)$$

Where, $u_{\Im}^{type}$ represents chosen network, which can or cannot be a standard network, such as a two-layer network that can be computed using equation (4),"

$$u_{\Im FF}^{FF}(x) = \sum_{i=1}^{M} c_i \phi_1\left(w_i^T x + b_i\right) \quad (4)$$

In the equation, $c_i\phi_1, w_i^T x$ represents the network parameter. The objective of the learning process is to obtain network parameters values, which minimize the loss operation, which is shown in equation (5).

$$\min_{\Im_{Type}} L\left(\Im_{type}\right) \quad where, type = N0\, or\, N1 \quad (5)$$

Finally, the adaptive activation function with deep Kronecker neural network (AAF-DKNN) was successfully applied to classify prostate cancer (PCa) samples into N1/N0 groups. Samples from patients with PCa involving lymph nodes made comprised the N1 group, while samples from patients in the N0 group had no involvement of the lymph nodes.

4. RESULT AND DISCUSSIONS

In this section, assess the efficiency of the proposed BI-PC-AAF-DKNN method for analysing the genes involved for prostate cancer spreading to nearby lymph nodes. The simulation was conducted using MATLAB, which is known for its accuracy in obtaining results. Performance metrics such as accuracy, computational time, and ROC were used to compare the results of BI-PC-AAF-DKNN with existing methods. The analysis shows that BI-PC-AAF-DKNN outperforms the other techniques based on these metrics. The comparison was made with two other methods, namely BI-PC-KMC [6] and BI-PC-SHAG [7], which employed supervised and unsupervised machine learning techniques and an integrative bioinformatics analysis, respectively.

4.1 Performance metrics

The proposed BI-PC-AAF-DKNN method was evaluated by comparing it with existing approaches in terms of accuracy, computational time, and ROC. To assess the performance metrics, the computation of a confusion matrix was necessary.

The following classes were used to calculate performance metrics for the proposed method:

- True positive (α): positive samples (samples in N1 group) predicted into positive.
- True negative (β): negative samples (samples in N0 group) predicted into negative.
- False positive (χ): negative samples predicted into positive.
- False negative (δ): positive samples predicted into negative.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

These metrics are employed for choosing ideal methods and determined for train, cross-verification, entire set or test.

4.1.1 Accuracy

The mathematical expression for Accuracy is given in equation (6).

$$Accuracy = \frac{\alpha + \beta}{\alpha + \chi + \delta + \beta} \tag{6}$$

The computation of AUC is performed using equation (7).

$$AUC = 0.5 \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{7}$$

4.3 Performance Analysis

The simulation results of the proposed BI-PC-AAF-DKNN method are illustrated in Figures 2-4. Furthermore, a comparison is made between the proposed method and existing methods, namely BI-PC-KMC and BI-PC-SHAG.

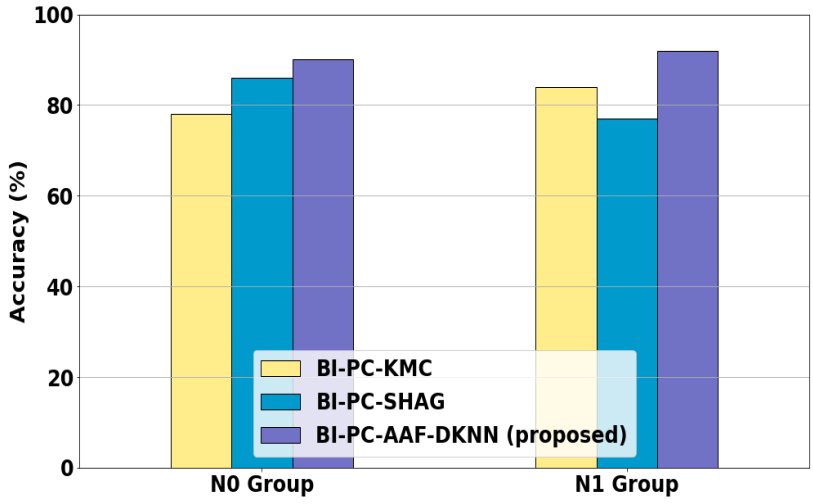

Figure 2: Accuracy Analysis

Figure 2 presents a comparison of accuracy analysis between the proposed BI-PC-AAF-DKNN method and existing methods, namely BI-PC-KMC and BI-PC-SHAG. The outcomes indicate that proposed technique outperforms the existing techniques with 33.34%, 29.56% higher accuracy for regular condition, 38.61%, 52.26%, 31.77% higher accuracy for N0 group classification, and 44.23% and 23.22% higher accuracy for N1 group classification, respectively.

ISSN 1533-9211

DOI: 10.5281/zenodo.8366457

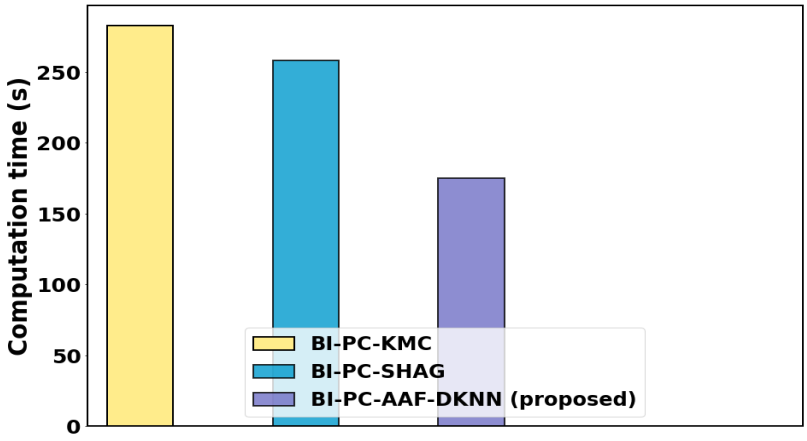

Figure 3: Computational Time Analysis

Figure 3 displays comparison of computational time analysis. The proposed BI-PC-AAF-DKNN method has been found to provide 33.34% and 29.56% lower computational time compared to existing methods, namely BI-PC-KMC and BI-PC-SHAG, respectively.

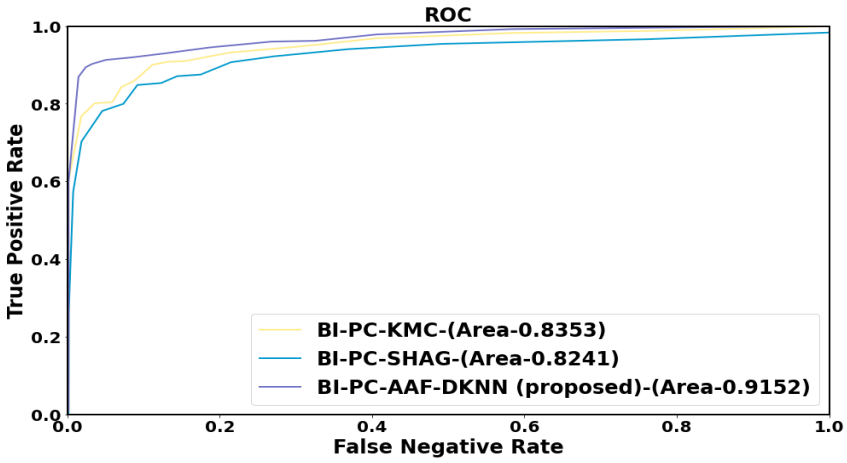

Figure 4: Analysis of RoC

In Figure 4, the RoC analysis is presented, showing that the proposed BI-PC-AAF-DKNN method achieves higher AUC values of 2.292% and 3.915% when compared to other methods such as BI-PC-KMC and BI-PC-SHAG.

5. CONCLUSION

The successful implementation of the BI-PC-AAF-DKNN method for analysing the genes involved for prostate cancer spreading to nearby lymph nodes in bioinformatics has been achieved. The approach was implemented using MATLAB and demonstrated superior performance with 7.69% and 10.805% higher accuracy when compared to existing methods such as BI-PC-KMC and BI-PC-SHAG, respectively.

The Seybold Report
ISSN 1533-9211
DOI: 10.5281/zenodo.8366457

Reference

1) Kendrick, J., Francis, R., Hassan, G.M., Rowshanfarzad, P., Jeraj, R., Kasisi, C., Rusanov, B. and Ebert, M., 2021. Radiomics for identification and prediction in metastatic prostate cancer: a review of studies. *Frontiers in Oncology*, *11*, p.771787.

2) Erdem, E. and Bozkurt, F., 2021. A comparison of various supervised machine learning techniques for prostate cancer prediction. *Avrupa Bilim ve Teknoloji Dergisi*, (21), pp.610-620.

3) Conti, D.V., Darst, B.F., Moss, L.C., Saunders, E.J., Sheng, X., Chou, A., Schumacher, F.R., Olama, A.A.A., Benlloch, S., Dadaev, T. and Brook, M.N., 2021. Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature genetics*, *53*(1), pp.65-75.

4) Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S.H., Salari, K., Kregel, S., Richter, C. and Arnoff, T.E., 2021. Biologically informed deep neural network for prostate cancer discovery. *Nature*, *598*(7880), pp.348-352.

5) Boschheidgen, M., Schimmöller, L., Arsov, C., Ziayee, F., Morawitz, J., Valentin, B., Radke, K.L., Giessing, M., Esposito, I., Albers, P. and Antoch, G., 2021. MRI grading for the prediction of prostate cancer aggressiveness. *European Radiology*, *32*(4), pp.2351-2359.

6) Shamsara, E. and Shamsara, J., 2020. Bioinformatics analysis of the genes involved in the extension of prostate cancer to adjacent lymph nodes by supervised and unsupervised machine learning methods: The role of SPAG1 and PLEKHF2. *Genomics*, *112*(6), pp.3871-3882.

7) Lai, Y.L., Liu, C.H., Wang, S.C., Huang, S.P., Cho, Y.C., Bao, B.Y., Su, C.C., Yeh, H.C., Lee, C.H., Teng, P.C. and Chuu, C.P., 2022. Identification of a steroid hormone-associated gene signature predicting the prognosis of prostate cancer through an integrative bioinformatics analysis. *Cancers*, *14*(6), p.1565.

8) Chen, X., Ma, J., Xu, C., Wang, L., Yao, Y., Wang, X., Zi, T., Bian, C., Wu, D. and Wu, G., 2022. Identification of hub genes predicting the development of prostate cancer from benign prostate hyperplasia and analyzing their clinical value in prostate cancer by bioinformatic analysis. *Discover Oncology*, *13*(1), p.54.

9) Wang, Y., Wang, J., Yan, K., Lin, J., Zheng, Z. and Bi, J., 2020. Identification of core genes associated with prostate cancer progression and outcome via bioinformatics analysis in multiple databases. *PeerJ*, *8*, p.e8786.

10) https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

11) Wu, H., Cheng, Y., Chen, X., Yang, Z., Li, X. and Wang, H., 2022. Geodesic Normal Coordinate-Based Manifold Filtering for Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, pp.1-15.

12) Jagtap, A.D., Shin, Y., Kawaguchi, K. and Karniadakis, G.E., 2022. Deep Kronecker neural networks: A general framework for neural networks with adaptive activation functions. *Neurocomputing*, *468*, pp.165-180.

13) NK Sakthivel, S Subasree, P Alias Priya M, AK Tyagi, "Breast lesion identification and categorization using mammography screening based on combined convolutional recursive neural network framework with parameters optimized", Concurrency and Computation: Practice and Experience 34 (28),2022.

14) S Subasree, NK Sakthivel, M Shobana, AK Tyagi. "Deep Learning based Improved Generative Adversarial Network for Addressing Class Imbalance Classification Problem in Breast Cancer Dataset", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Volume 31, Pp 387-412, 2023