JOURNAL OF DECISION MAKING AND HEALTHCARE
Volume 3 (2026), No. 1, 39–45
https://doi.org/10.69829/jdmh-026-0301-ta04
Tulipa Opera Scholarum

# MINIMIZING GENES FOR CANCER DETECTION USING A GENETIC ALGORITHM

YEN-JU TSUI[1], FENG-SHENG TSAI[2,3,*], AND HAO-REN YAO[4]

[1] *Master Program for Biomedical Engineering, China Medical University, Taichung 40402, Taiwan*
[2] *Department of Biomedical Informatics, China Medical University, Taichung 40402, Taiwan*
[3] *Research Center for Interneural Computing, China Medical University Hospital, Taichung 40447, Taiwan*
[4] *Information Retrieval Lab, Georgetown University, Washington, DC 20057, USA*

Abstract. The cancer genome atlas database contains extensive genomic data on various cancers, while the catalogue of somatic mutations in cancer (COSMIC) provides a curated list of oncogenes. Although oncogenic data from TCGA can be extracted to analyze cancer types, utilizing the entire genomic dataset for screening is computationally resource-intensive. Therefore, this study aims to employ a genetic algorithm to identify the minimum number of genes required for accurate cancer detection. In this research, gene expression data from six types of cancer and their corresponding normal tissues were extracted. A subset of 716 genes and their expression values were randomly selected based on a gene activation probability $p$. Through the GA process, including data input, fitness evaluation, selection, crossover, mutation, and iteration, the classification accuracy for various values of $p$ was determined. The fitness function was calculated using a classification neural network, where the accuracy of the network, trained and tested on the activated gene expression values, served as the fitness score. The GA enables the activation of randomly selected genes to evolve through generations, increasingly optimizing the identification of genes necessary for classifying these six cancer types and normal tissues. Experimental parameter tuning involved the gene activation probability, crossover probability, and mutation probability. The result indicates that for any crossover probability and active mutation, a stable accuracy exceeding 93.5% is achieved when $p \geq 0.1$.

**Keywords.** Cancer Detection, Gene Wxpression, Genetic Algorithm, Neural Networks, Oncogenes.


## 1. Introduction

Recent studies have explored various machine learning approaches for cancer classification [4, 6, 9, 11, 12, 13]. Danaee et al. [4] proposed a framework based on stacked denoising autoencoders (SDAE), utilizing transcriptomic data from the cancer genome atlas (TCGA) [16] consisting of 1,097 breast invasive carcinoma (BRCA) samples and 113 healthy controls. To mitigate the effects of class imbalance, the synthetic minority over-sampling technique (SMOTE) [2] was employed during the pre-training phase to transform the data into a more balanced representation. The primary advantage of the SDAE model lies in its multi-layered architecture, which progressively transforms raw genomic data into low-dimensional, highly representative feature vectors. Regarding classification performance, the study provided rigorous quantitative results. Experimental outcomes demonstrated that when a neural network (NN) was applied to the identified deeply connected genes (DCGs), the classification accuracy reached 91.74%. Comparative analysis with other machine learning models indicated that a standard support vector machine (SVM) [3] achieved an identical accuracy of 91.74%, while the SVM utilizing a radial basis function (RBF) kernel attained a superior accuracy of 94.78%. Furthermore, by analyzing the

*Corresponding author.
E-mail address: lulu.taui1@gmail.com (Y.-J. Tsui), fstsai@mail.cmu.edu.tw (F.-S. Tsai), and hao-ren@ir.cs.georgetown.edu (H.-R. Yao)
Accepted: March 06, 2026.

connectivity matrices within the SDAE, the researchers isolated 500 critical DCGs. Bioinformatic analysis confirmed that these genes significantly overlap with biological pathways associated with tumor proliferation and suppression. This demonstrates that deep learning not only enhances the precision of cancer classification but also enables the identification of clinically significant biomarkers from complex genomic datasets.

Kim et al. [9] explored the feasibility of multi-cancer classifiers. They defined a "minimum gene set" capable of reliably distinguishing 21 cancer types and their corresponding normal tissues. ANOVA tests were utilized to compare cancer and normal samples, selecting the top 300 genes with the most significant differential expression as the final feature set. In terms of model comparison, the study evaluated the performance of NNs, SVMs, $k$-nearest neighbors (KNNs) [15], and random forests [1]. Experimental results demonstrated that the NN consistently outperformed other methods; specifically, based on the 300 essential genes, the NN achieved a Matthews correlation coefficient (MCC) of 0.89 and an accuracy of 0.9. In contrast, the KNNs method was the poorest performing classifier, yielding an MCC of only 0.71. Furthermore, gene set enrichment analysis (GSEA) confirmed that these 300 key genes are highly enriched in 10 core pathways, including the cell cycle and DNA replication, which are intrinsically linked to the dysregulated proliferation of tumor cells. This demonstrates that the NN not only enables effective cancer diagnosis but also allows for the extraction of biologically significant features from a condensed gene set.

Li et al. [11] proposed a hybrid architecture combining a genetic algorithm (GA) [8, 10] with KNNs. The GA is employed to simulate biological evolution through stages of initialization, evaluation, selection, crossover, and mutation. In the feature selection phase, the GA serves as an evolutionary search tool to iteratively screen and evolve high-dimensional gene expression data, ultimately identifying an optimal feature subset of only 20 genes. During the classification phase, the KNNs algorithm is utilized as a classifier to evaluate the discriminative efficacy of these gene combinations. Experimental results demonstrated that the model performed well in distinguishing between 31 cancer types, achieving an overall predictive accuracy of 95.6%. However, the study also noted that the GA/KNNs framework encounters limitations in efficiency and computational resource consumption when processing large-scale samples, primarily due to the requirement for numerous iterations (e.g., 300 generations) and frequent distance calculations. Furthermore, because the discriminative power of KNNs is inherently fixed, it lacks the capacity for performance enhancement through iterative training—a characteristic that distinguishes it from the NN and serves as a primary motivation for our study in this paper.

Addressing the limitations identified in prior research—specifically the low efficiency of KNN-based wrappers—this study implements a high-performance NN as the primary classifier within an evolutionary framework. Unlike KNNs, which possesses a fixed discriminative power, the NN's performance can be significantly enhanced through training, allowing for more robust handling of linear and non-linear relationships within the data. Our research utilizes a GA to simulate biological evolution through initialization, evaluation, selection, crossover, and mutation. We replace KNNs with NNs as the primary classifiers. A critical innovation in our methodology is the implementation of the gene activation probability $p$, which serves as a control mechanism to limit the total genomic input involved in the GA from the outset. We systematically analyze classification accuracy relative to the probability $p$ to identify the smallest gene subset capable of robust cancer detection. This reduction in genomic variables not only improves computational efficiency for clinical use but also highlights the critical "key genes" or "collaborative diagnostic genes" necessary for cancer detection.

## 2. Materials and Methods

The foundation of this study rests on the integration of two of the world's most comprehensive genomic repositories. The first is the cancer genome atlas (TCGA) [16] launched in 2006 following the Human Genome Project [7]. TCGA provides a systematic understanding of the molecular basis

of neoplasm formation, growth, and metastasis. It contains molecular characterizations of over 20,000 primary cancer samples across 33 distinct cancer types, offering a vast library of mRNA expression data that documents essential biological patterns and tumor behaviors beyond traditional clinical examinations. The second is the catalogue of somatic mutations in cancer (COSMIC) [14] recognized as the most exhaustive resource for exploring the impact of somatic mutations on human cancer. COSMIC features the cancer gene census (CGC). The CGC categorizes genes into Tier 1 (strong evidence of cancer relevance) and Tier 2 (emerging evidence in tumor development but with less robust mutation patterns). In this research, COSMIC-listed oncogenes are utilized to filter and extract the most relevant genomic features from the TCGA database.

Genomic expression values were extracted from the TCGA database for six common cancer types and their corresponding normal tissues: breast invasive carcinoma (BRCA), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), prostate adenocarcinoma (PRAD), thyroid carcinoma (THCA), and uterine corpus endometrioid carcinoma (UCEC). A total of 716 oncogenes were identified and extracted based on the COSMIC list to serve as the baseline features.

To strictly control the number of genes used, we implement a gene activation probability $p$, which establishes an upper limit on the total genomic data involved in the GA from the outset. We assume that each chromosome consists of 716 gene switches $G_n$, where $n$ denotes the index of the gene for $n = 1, 2, \ldots, 716$. The variable $G_n$ represents whether the $n$-th gene is activated, with a value of either 0 or 1. $G_n = 0$ indicates the $n$-th gene is not activated, whereas $G_n = 1$ indicates the $n$-th gene is activated. Various values of $p$ are applied, specifically: 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.025, and 0.0125. For example, if $p = 0.05$, then 5% of the 716 oncogenes will be activated, leading to a total of approximately 36 gene switches in the "on" state ($716 \times 0.05 = 36$). In each experiment of the GA, 10 chromosomes are used. The activation status of all 716 genes within the $m$-th chromosome, $m = 1, 2, \ldots, 10$, is denoted by $C_m$ as follows:

$$C_m = [G_1, G_2, G_3, \ldots, G_{716}], \text{ where } G_n \in \{0, 1\} \text{ for } n = 1, 2, \ldots, 716.$$

The calculation of the fitness value utilizes a neural network model. The neural network consists of three fully connected layers (dense layers). For the first two layers, ReLU (Rectified Linear Unit) is selected as the activation function, whereas the final layer utilizes the softmax function to handle multi-class classification problems. The input vector $X$ of the neural network is defined as:

$$X = [E_1 \cdot G_1, E_2 \cdot G_2, \ldots, E_{716} \cdot G_{716}],$$

where $E_n$ denotes the $n$-th gene expression. For optimization, stochastic gradient descent (SGD) is selected, and cross entropy is employed as the loss function. The classification categories consist of six cancer types BRCA, KIRC, LUAD, PRAD, THCA, UCEC, and normal tissues. Initially, 80 samples were picked from each cancer type to serve as the validation set, and another 80 samples were selected for the test set. Subsequently, from the respective normal control groups, 80 validation samples and 80 test samples were extracted proportionally to form the normal class. For the training set, the borderline-SMOTE [2, 5] was applied to the six cancer types as well as the normal class to expand the number of samples. A fitness value is assigned by the testing accuracy provided by the neural network. The fitness value serves as the criterion for evaluating the quality of the activation status of all 716 genes within a chromosome. A higher fitness value represents a superior degree of adaptation of a chromosome. Chromosomes with a higher degree of adaptation are then filtered and selected for the subsequent selection step.

The purpose of selection is to retain superior gene sequences so that good genes can be passed down to future generations. Selection probabilities are appropriately weighted according to the fitness values corresponding to different chromosomes. When a fitness value is higher, the individual has a higher probability of being selected, ensuring that high-quality gene sequences are more likely to be preserved.

This study employs elite selection, which preserves the most optimal genes in each generation. Ten sets of chromosomes $C_m$, $m = 1, 2, \ldots, 10$, undergo fitness scoring based on their fitness values. The top four elite chromosomes with the highest scores are selected. From these four elite chromosomes, two sets are randomly chosen to serve as parent chromosomes (Parent 1, Parent 2). These parents then undergo crossover and mutation to produce offspring.

This paper adopts uniform crossover to produce new chromosomes by recombining parental genetic information. Two chromosomes undergo crossover is controlled by a crossover probability $p_{cr}$. This process involves the generation of a crossover mask that matches the length of the parent chromosomes. Each position in the crossover mask is assigned a value of 1 with a probability of $p_{cr}$. When a value within the mask is 1, the switches corresponding to the genes in the parent chromosomes are swapped; otherwise, no exchange occurs. New offspring are generated using this method.

Similar to the crossover stage, changes are applied to specific gene switches within selected chromosomes based on a predefined mutation probability $p_{mu}$. We adopt one-point mutation, a method where a gene is selected from the chromosome with the probability $p_{mu}$ and its bit value is flipped between 0 and 1. This is mathematically represented as $G_n = 0 \to 1$ or $G_n = 1 \to 0$. To maintain stability, the following quantity control rule is implemented during mutation. Since $716 \cdot p \cdot p_{mu}$ represent the total gene switches transitioning from an activated to an inactivated state, and $716 \cdot (1 - p) \cdot p_{mu}$ represent the reverse. Given the parameter setting $p \leq 0.5$, it follows that $p \leq (1 - p)$; therefore, $716 \cdot p \cdot p_{mu} \leq 716 \cdot (1 - p) \cdot p_{mu}$. To prevent an excessive increase in activated genes and maintain the total count below the initial threshold $716 \cdot p$, we require that the number of activated gene switches after mutation is randomly reduced to a maximum of $716 \cdot p$.

As a result, the two selected elite parents (Parent 1 and Parent 2) undergo three crossover operations to produce six post-crossover chromosomes. These chromosomes then undergo mutation operations, resulting in the final six offspring chromosomes. The output for the current generation consists of the four elite chromosomes and the six offspring chromosomes (totaling ten sets), which serve as the input for the subsequent generation. This process is iterated for 30 generations. By testing accuracy across various values of $p$, $p_{cr}$, and $p_{mu}$, the research identifies the minimum gene requirement necessary to achieve high-precision classification.

## 3. Results

To test the efficiency of the GA with NNs, we fix the number of training epochs equaling 200 and vary the gene activation probability $p = 0.0125, 0.025, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$, the crossover probability $p_{cr} = 0.2, 0.3, 0.4, 0.5$, and the mutation probability $p_{mu} = 0, 0.1, 0.2, 0.3$.

A comparison of accuracy across various mutation probabilities reveals that the crossover probability has a negligible impact on the final performance. As illustrated in Figure 1, regardless of the values for all $p_{cr}$ and active mutation $p_{mu} = 0.1, 0.2, 0.3$, the accuracy (Acc) increases significantly as parameter $p$ increases. When $p \geq 0.1$, the accuracy for all configurations tends to stabilize, eventually converging at approximately 95%. The accuracy curves for different $p_{cr}$ values tend to overlap or follow nearly identical trajectories under the same mutation and gene activation conditions. This suggests that within the scope of this experiment, the crossover probability is not a dominant factor in determining the neural network's accuracy.

As illustrated in Figure 2, when mutation is disabled ($p_{mu} = 0$), the accuracy performance is relatively poor across almost all gene activation levels $p \leq 1$. All active mutation settings $p_{mu} = 0.1, 0.2, 0.3$ perform substantially better than the zero-mutation condition, highlighting that the presence of mutation, rather than its specific intensity, is vital for sustained improvement.

The gene activation probability serves as the most direct predictor of model accuracy. The experimental results demonstrate a positive correlation where higher gene activation probability yields higher


FIGURE 1. Regardless of the values for all $p_{cr}$ and active mutation $p_{mu} = 0.1, 0.2, 0.3$, the accuracy (Acc) increases significantly as parameter $p$ increases

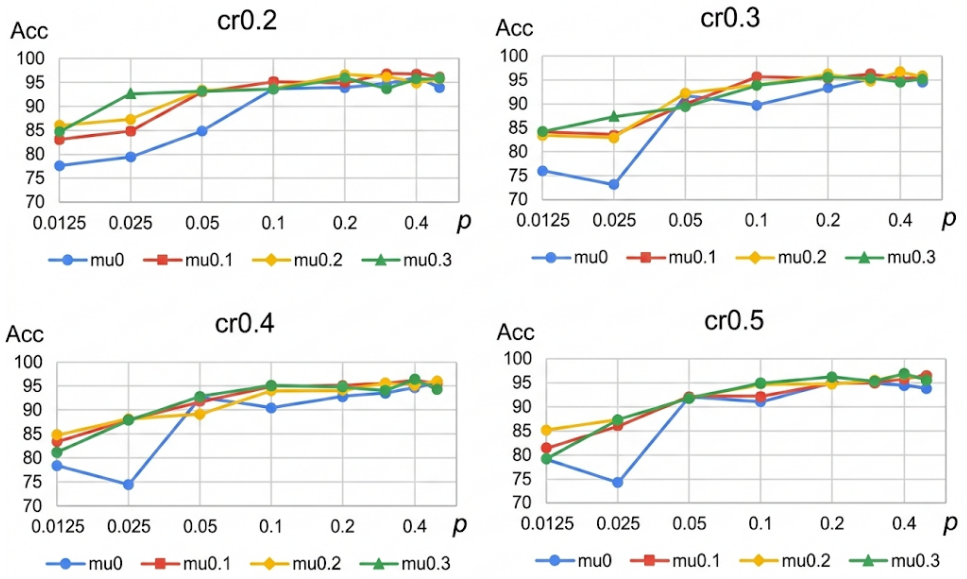


FIGURE 2. The presence of mutation is vital for accuracy performance especially when the gene activation probability $p \leq 1$

model accuracy. A key finding is that for all active mutation $p_{mu} = 0.1, 0.2, 0.3$, the accuracy curves consistently stay above 93.5% once the gene activation probability reaches $p \geq 0.1$.

In summary, the efficiency of the GA in this context is maximized by prioritizing a sufficient gene activation probability and ensuring an active mutation setting, while the crossover probability serves as a secondary, highly tolerant parameter.

## 4. Discussion

The findings of this study underscore the efficacy of integrating GA with NNs to streamline the identification of essential biomarkers for multi-cancer classification. A primary contribution in this research is the implementation of the gene activation probability $p$, which serves as a restrictive control mechanism for genomic input of GA. The experimental data reveals a positive correlation: as $p$ increases, classification accuracy improves significantly. Specifically, once $p \geq 0.1$, the NN model consistently achieves a stable accuracy exceeding 93.5%, eventually converging around 95%. This suggests that a relatively small subset of the genome contains sufficient information for robust cancer detection.

The comparative analysis of GA parameters highlights distinct roles for mutation and crossover. When mutation was disabled ($p_{mu} = 0$), the model's performance was notably poor across almost all

gene activation levels. The presence of active mutation ($p_{mu} = 0.1, 0.2, 0.3$) proved essential for sustained improvement. Conversely, the crossover probability $p_{cr}$ exerted a negligible impact on model performance. Accuracy trajectories for various $p_{cr}$ values were nearly identical, indicating that the NN's performance is highly tolerant of crossover variations within this GA/NNs framework. In conclusion, with GA/NNs, an efficient configuration for the identification of "collaborative diagnostic genes" prioritizes a sufficient gene activation probability and active mutation, while remaining flexible regarding crossover settings.

## Statements and Declarations

The authors declare that they have no conflict of interest. This work is adapted from the author Y.-J. Tsui's master's thesis.

## Acknowledgments

This research was based on data generated by the COSMIC Cancer Gene Census (https://cancer.sanger.ac.uk/census) and the TCGA Research Network (https://www.cancer.gov/tcga). This work was supported in part by the National Science and Technology Council, Taiwan, and in part by the China Medical University.

## References

[1] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[4] P. Danaee, R. Ghaeini, and D. A. Hendrix. A deep learning approach for cancer detection and relevant gene identification. *Pacific Symposium on Biocomputing*, 22:219–229, 2017.

[5] H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer, Berlin, Heidelberg.

[6] S.-Y. Hsu, M.-H. Shih, W.-H. Wu, H.-R. Yao, and F.-S. Tsai. Gene reduction for cancer detection using layer-wise relevance propagation. *Journal of Decision Making and Healthcare*, 1(1):30–44, 2024.

[7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[8] S. Katoch, S. S. Chauhan, and V. Kumar. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*, 80:8091–8126, 2021.

[9] B.-H. Kim, K. Yu, and P. C. W. Lee. Cancer classification of single-cell gene expression data by neural network. *Bioinformatics*, 36(5):1360–1366, 2020.

[10] A. Lambora, K. Gupta, and K. Chopra. Genetic algorithm–a literature review. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 380–384, Faridabad, India, 2019. IEEE, New York.

[11] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12):1131–1142, 2001.

[12] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics*, 18(1):Article ID 508, 2017.

[13] J. Oyelade, I. Isewon, F. Oladipupo, O. Aromolaran, E. Uwoghiren, F. Ameh, M. Achas, and E. Adebiyi. Clustering algorithms: Their application to gene expression data. *Bioinformatics and Biology Insights*, 10:237–253, 2016.

[14] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[15] K. Taunk, S. De, S. Verma, and A. Swetapadma. A brief review of nearest neighbor algorithm for learning and classification. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1255–1260, Madurai, India, 2019. IEEE, New York.

[16] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, and Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.