Cell J
Cell Journal(Yakhteh)
Homepage: https://www.celljournal.org/
Original Article
Vol 26, No 12, December 2024, Pages: 688-699

Unlocking The Potential of Hybrid Models for Prognostic Biomarker Discovery in Oral Cancer Survival Analysis: A Retrospective Cohort Study

Leila Nezamabadi Farahani, M.Sc.[1], Anoshirvan Kazemnejad, Ph.D.[1]*, Mahlagha Afrasiabi, Ph.D.[2], Leili Tapak, Ph.D.[3, 4]

1. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Department of Computer, Hamedan University of Technology, Hamedan, Iran
3. Modeling of Noncommunicable Diseases Research Center, Institute of Health Sciences and Technologies, Hamadan University of Medical Sciences, Hamadan, Iran
4. Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran

Abstract

Objective: This study aimed to develop a hybrid model for variable selection in high-dimensional survival analysis using a support vector regression (SVR), to identify prognostic biomarkers associated with survival in oral cancer (OC) patients through the analysis of gene expression data.

Materials and Methods: In this retrospective cohort study, gene expression profiles (54,613 probes) related to 97 patients from the GSE41613 dataset from the GEO repository were used. First of all, martingale residuals were obtained using a Cox regression without covariates, and were used as pseudo-survival outcome. Then, the particle swarm optimization (PSO) and genetic algorithm (GA) were used in combination with SVR for selecting features related to pseudo-survival outcome. Concordance index (C-index), mean absolute error (MAE), mean squared error (MSE) and R-squares, were used to evaluate the performance of the models using selected features. Functional enrichment analysis was performed using DAVID database, and external validation utilized three independent datasets (GSE9844, GSE75538, GSE37991, GSE42743).

Results: The findings indicated that the PSO-based method outperformed the GA-based method, achieving a smaller MAE (0.061) and MSE (0.005), R-square (0.99) and C-index (0.973), selecting 291 probes from 1069 screened. A protein-protein interaction (PPI) network was constructed, including 200 nodes and 120 edges. Eleven key genes with the highest degree, including *RBM25, SMC3, PRPF40A, POLE, SRRT, BCLAF1, PDS5B, HNRNPR, JAK1, MED23,* and *SULT1A1* were identified as significant biomarkers associated with OC survival.

Conclusion: The PSO-based hybrid model effectively improved SVR performance in survival prediction for OC patients and identified key prognostic biomarkers. Despite its promising results and validation on independent datasets, limitations in generalizability and signs of overfitting suggest the model is not yet ready for clinical use. Further studies with larger, diverse datasets are recommended.

Keywords: Feature Selection, Gene Expression, Genetic Algorithm, Mouth Neoplasms, Particle Swarm Optimization

Citation: Nezamabadi Farahani L, Kazemnejad A, Afrasiabi M, Tapak L. Unlocking the potential of hybrid models for prognostic biomarker discovery in oral cancer survival analysis: a retrospective cohort study. Cell J. 2024; 26(12): 688-699. doi: 10.22074/cellj.2025.2034704.1618
This open-access article has been published under the terms of the Creative Commons Attribution Non-Commercial 3.0 (CC BY-NC 3.0).

Introduction

Oral cancer (OC) is one of the top 10 cancers in men and the 12th most common in women worldwide, with oral squamous cell carcinoma (OSCC) accounting for over 90% of head and neck cancers (1). Despite advancements in screening and treatment, OC incidence and mortality rates remain high, largely due to late-stage diagnoses and poor prognosis, with a five-year survival rate below 50% (1, 2).

Recent studies highlight the need for identifying biomarkers for OSCC diagnosis and prognosis, especially since most treatments have limited efficacy and only 15% of drugs are approved (3). Biomarkers predicting OSCC progression and patient survival are not well understood, highlighting the need for new prognostic markers. Advances in this area are crucial for better patient monitoring and treatment. In this regard, machine learning models are increasingly explored to identify such biomarkers and improve outcome prediction in OSCC (4).

In recent decades, many studies have developed models for high-dimensional data such as gene expression (4-6). Among these, support vector machines (SVMs), including their regression variant known as support vector regression (SVR), are widely used in the medical field for their effectiveness in classification and regression tasks (6). SVM/SVR stands out in high-dimensional data analysis because of its adaptability and several key advantages. Its flexible kernel selection allows it to handle diverse datasets, and by storing only support vectors, it remains memory-efficient. The convex nature of its optimization

Received: 06/July/2024, Revised: 26/April/2025, Accepted: 20/May/2025
*Corresponding Address: P.O.Box: 14115111, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
Email: kazem_an@modares.ac.ir

Royan Institute
Cell Journal (Yakhteh)

ensures stable, globally optimal solutions, while the maximum margin decision boundary supports strong generalization across different applications (7). However, SVM/SVR performance largely depends on proper hyperparameter selection, such as the penalty parameter (C) and kernel parameters. This poses challenges in high-dimensional data, including risks of overfitting and high computational cost (8). Traditional hyperparameter tuning methods like grid or random search are often inefficient, time-consuming, and may yield suboptimal results. To overcome this, we used a hybrid approach employing genetic algorithms (GA) for hyperparameter optimization. Additionally, heuristic methods such as GA and particle swarm optimization (PSO) can further improve SVM/SVR by efficiently selecting relevant features from large datasets (5, 7).

Studies have shown that PSO surpasses the GA and grid search in achieving higher accuracy and faster computation (8). PSO is simple and effective, especially in handling optimization with multiple local optima (9), while GA is also effective for feature selection in machine learning models (10). By integrating these heuristic methods with SVR, studies have shown that the model accuracy and robustness can improve, ultimately leading to better predictions (11-13). Methodologically, only one study has used GA for SVR feature selection in low-dimensional problems, demonstrating high sensitivity in identifying informative variables (14).

In practice, several studies have combined GA and PSO with SVM and other ML methods to identify key biomarkers with high diagnostic accuracy for various cancers and diseases in high-dimensional settings (5, 6). However, these studies primarily focused on binary outcomes (dead/alive) without considering time-to-event data or censoring issues. Ignoring censored observations can lead to an underestimation of survival probabilities beyond fixed time points (5, 6).

Despite widespread use of PSO/GA-optimized SVR for classification, no hybrid method has been developed for selecting prognostic biomarkers in survival analysis with high-dimensional data, including OSCC. Moreover, there no study was found to develop direct SVM model for high dimension survival data. On the other hand, studies have shown through simulations that the survival response can be converted into a regression problem through residuals of a Cox regression model (e.g. martingale residuals) (15-17). Their results have shown that this approach resulted in promising performances. However, for the context of high-dimension survival data, no attempt has been done yet. This gap highlights the need for innovative methods that effectively identify prognostic biomarkers to enhance survival prediction and patient care.

To address this gap, we present a novel hybrid method that integrates PSO/GA and SVR for survival prediction in high-dimensional settings. Applied to OSCC, this approach uses PSO/GA for efficient feature selection and SVR for survival outcome modeling, enabling identification of key prognostic biomarkers. This method advances biomarker discovery and clinical decision-making by improving survival prediction and patient care. The main aim of this study is to assess whether integrating PSO/GA-based feature selection with SVR modeling can improve the prediction of prognosis and identify key prognostic biomarkers in patients with OSCC.

Materials and Methods

In this retrospective cohort study, we describe the step-by-step process of our proposed hybrid survival analysis approach. First, martingale residuals are introduced and used as survival outcomes. Next, we detail the SVR model, which forms the core of our prediction framework. We then explain the optimization algorithms—PSO and GA—employed for feature selection and parameter tuning. The two-step hybrid algorithm and cross-validation (CV) strategy are outlined, followed by the experimental design, data preprocessing steps, and performance evaluation metrics.

Martingale residuals

Let us consider $N_i(t)$ be the counting processes indicating the occurrence of the event of interest, and $\Lambda_i(t)=\int_0^t \lambda_o(u)\,du$ be the cumulative intensity function. The martingale residual for subject i over the time interval from 0 to t is defined as:

$$M_i(t) = N_i(t) - \Lambda_i(t) \quad [1]$$

Therneau et al. (18) showed that martingale residuals, obtains from a null Cox model (without covariates), can be used as the outcome variable in regression problems. They demonstrated that this approach worked well for survival data. In our study, we extend their approach by combining martingale residuals with heuristic optimization algorithms (PSO/GA) for high-dimensional variable selection and survival prediction.

Support vector regression

In the regression problem, we have several covariates X and a response variable y_i, which represents the martingale residuals, the pseudo-survival outcome. These residuals are calculated by fitting the Cox proportional hazards (PH) model, considering a vector of response (T_i, δ_i), where T indicates the time to developing OC and δ is the censoring status in the training dataset. Thus, the SVR model takes the following form:

$$f(x) = W.\phi(X) + b \quad [2]$$

In a SVR, the norm of the coefficient vector (i.e. $1/2\ \|w\|^2$) is considered as the objective function and is minimized. The $f(x)$ should be as flat as possible and has the least deviation from the actual target for all training data, it means all the residuals must be less than or equal to the certain value called, maximum error (ε):

$$|y_i - w^T.\phi(x_i) - b| \leq \varepsilon \quad [3]$$

The function f(x) might not exist to satisfy [3] and [2] for all points; in order to deal with this issue, we can denote

any deviation from ε as and for each point, named slack variables. It makes constrain more flexible. The primal objective function includes these slack variables, and an additional hyperparameter, C > 0, controls the penalty for samples lying outside the (ε) and helps prevent overfitting (6). By this value we could make a trade-off between flatness and tolerance for samples outside the (ε) by minimizing:

$$1/2\ \|w\|^2+C\sum(\xi_i+\xi_i^*) \quad [4]$$

subject to:

$$\begin{cases} (y_i-w^T.\phi(x_i) - b \le \varepsilon + \xi_n \\ w^T.\phi(x_i)+b-y_i \le \varepsilon + \xi_n^* \\ \xi, \xi^* \ge 0 \end{cases} \quad [5]$$

Optimization problems can be defined as primal and Lagrange dual formulation. When the number of variables is large, but the number of samples is small, we need to use the kernel trick, and in this case, the dual form is preferred, which is written as follows (6):

$$L = 1/2\ \|w\|^2 + C\sum_n^{i=1}(\xi_i+\xi_i^*) - \sum_n^{i=1}(\eta_i\xi_i+\eta_i^*\xi_i^*) \quad [6]$$
$$-\sum_n^{i=1}\alpha_i\ (\varepsilon + \xi_n - y_i + w^T.\ \phi(x_i) + b)$$
$$-\sum_n^{i=1}\alpha_i^*\ (\varepsilon+\xi_n^* + y_i - w^T.\ \phi(x_i) - b)$$

Particle swarm optimization algorithm

PSO, introduced by Kennedy and Eberhart, is a population-based search method inspired by the social behavior of bird flocks (9). In PSO, individuals (known as particles) traverse a hyperdimensional search space. Each particle's position is updated based on both its own experiences and the success of other particles within the swarm. This process enables the swarm to dynamically adjust, leveraging collective knowledge to explore promising areas of the search space. In this algorithm, a particle's movement is influenced by its own best-known position (cognitive component) and the best-known position of the entire swarm (social component). Each particle is treated as a potential solution with a velocity and position, flying through the problem space.

Let and represent the velocity and position of the i-th particle at iteration t. denotes the particle's personal best position. represents the global best position found by the entire swarm.

The velocity and position are updated as follows:

$$v_i(t+1) = wv_i(t) + c_1 r_1 (Pb_i(t)-x_i(t)) + c_2 r_2 (gb_i(t) - x_i(t)) \quad [7]$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad [8]$$

Here, w is the inertia weight balancing exploration and exploitation, while c_1 and c_2 are acceleration coefficients controlling the cognitive and social components. Random values r_1 and r_2, generated between 0 and 1, add stochastic variation to the process, representing the particle's memory or learning capability (9).

To prevent unnecessary wandering, particle movement is often restricted within defined boundaries. The algorithm terminates either when a maximum number of generations is reached or when no significant improvements occur in the best particle's position. This enables PSO to provide robust and efficient solutions to complex optimization problems (9).

Genetic algorithm

GA is one of the most commonly employed optimization algorithms, inspired by the principle of natural selection. GA mimics the biology of changes occurring throughout the DNA sequence via mutation and crossover and assesses the set of solution through a fitness function (here, it is defined for the SVR for the survival response). Similar to PSO, GA is a population-based search algorithm that operates based on the survival of fittest concept (19). By applying various genetic operators: mutation, crossover, deletion and insertion to the subjects of the original population, new populations are created.

The key elements of GA include the chromosome representation and fitness function computation, along with the genetic operations. In GA, the fitness function is computed for every element (chromosome) in a randomly initialized population of size n (the number of chromosomes). Based on the fitness values obtained, in the previous step, two chromosomes (solution) are selected from the initial population, and operations like crossover and mutation are applied to these selected chromosomes to generate offspring. This process is repeated over the population until the new population is complete (19).

The search process in GA dynamically changes by selecting various crossover and mutation probabilities until reaching to the optimal solution. GA has a good global search capabilities, as it can generate optimal solutions (19). In this study, we aimed to handle parameter optimization along with the variable selection, the chromosomes were an arrangement of parameter genes (e.g. C, ε,). Solutions were considered as binary strings (0/1). The chromosome's parameters were employed for SVR modeling, and the mean squared error (MSE) and determination coefficient (R^2) of the SVR were used as the fitness function. A higher fitness value indicated a better individual.

The two-step survival hybrid algorithm

The proposed algorithm consists of two main steps:

Step 1: Fit a null survival model (without covariates) to estimate the baseline intensity functions (using the Breslow estimator) (20). The martingale residuals are

calculated as risk indices for each patient.

Step 2: Apply the optimal SVR model, optimized by GA, to select the relevant features and fine-tune the model's parameters.

The first step provides quantitative risk indices for survival prediction. The second step uses PSO/GA to optimize the SVR model, improving feature selection and using GA for parameter tuning. The final subset of gene expression features is selected for survival prediction. The pseudo codes were provided in Supplement Figures 1, 2 (See Supplementary Online Information at celljournal.org).

Cross-validation strategy

We employed a 10-fold CV approach to evaluate the models. The dataset was split into 10 subsets, with each subset serving as the validation set once, while the remaining data were used for training. Performance metrics were averaged across folds to assess model generalization and ensure robust evaluation.

Experiment

Data source and preprocessing

We utilized a public dataset of HPV-negative OSCC obtained from the GEO repository (accession number GSE41613). The dataset was generated using Affymetrix U133 Plus microarrays, specifically under Platform ID GPL570. This dataset can be accessed at the NCBI Gene Expression Omnibus via the following link: GEO Accession viewer (nih.gov)

This dataset was selected for its relevance to identifying survival related biomarkers and the availability of survival data. Specifically, this dataset contains Affymetrix array data from 97 HPV-negative OSCC patients, and has been used in several studies (21). It comprises 54,613 expression data point related to 97 patients with OC, of whom 51 died during follow-up and the remaining were censored. Survival time was defined as the interval between diagnosis and death due to OSCC, with censored cases representing patients alive at the end of the study. Clinical information includes age, sex, and disease stage.

Patient characteristics were as follows: 32% female and 68% male, with age distribution ranging from 19 to 88 years (6.2% aged 19-39, 16.5% aged 40-49, 28.9% aged 50-59, and 48.5% aged 60-88). Most patients (57.7%) had tumors staged III/IV, while the rest (42.3%) were staged I/II. Data were normalized using the quantile normalization (QN) procedure implemented with the bestNormalize R package. QN was chosen to correct technical variability and ensure comparable scales across samples. Genes with an adjusted P<0.01 in univariate Cox PH models were retained (1,096 probes).

Initial screening

To identify relevant genes, we applied a univariate Cox PH regression model to the 54,613 gene expression profiles. Considering gene expression as explanatory variables and T and C as event and censoring times, the univariate Cox PH model was defined as:

$$h(t \mid X) = h_0(t) \times e^{\beta X} \quad [9]$$

where $h(t \mid X)$ indicates the hazard function and the $h_0(t)$ is the baseline hazard. Gene expressions with an adjusted P<0.01 were retained for further analysis. Of the initial 54,613 probe sets, a number of 1,096 probes were selected.

Implementation

The proposed hybrid model was implemented in MATLAB (version R2021a), utilizing toolboxes for PSO, GA, NN, 1. Data were randomly divided into training (70%) and test (30%) subsets.

PSO-SVR: Feature selection was performed by ranking features based on importance scores derived from the PSO-SVR method, which used 100 particles and 50 iterations (Fig.S3, See Supplementary Online Information at celljournal.org). Parameters were set as w=0.7298, C1=C2=1.4962 (Equation 7).

GA-SVR: GA (population size=5, crossover rate=0.8, mutation rate=0.01) were chosen based on grid search results to balance computational efficiency and accuracy. Figures depicting prediction error curves, concordance indices, and area under the receiver operating characteristic (ROC) curves were generated using R (version 4.0.0).

Evaluation criteria

We evaluated the model's performance using a set of appropriate metrics. Since the survival outcomes were transformed into continuous values via Martingale residuals, some survival-specific metrics such as the Brier Score were no longer applicable. However, the C-index was retained, as it remains meaningful for evaluating the model's ability to rank predicted outcomes, even in a regression context. Therefore, we focused on suitable regression-based metrics such as MSE, MAE, and R^2, alongside the C-index for assessing concordance.

MSE which measures the amount of error in the used model as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(M_i - \hat{M}_i)^2 \quad [10]$$

In the above formula the M_i and $\hat{M}_i$ represents the martingale residual and its predicted value from the model for the subject i, respectively.

Mean absolute error which is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|M_i - \hat{M}_i| \quad [11]$$

The Pearson correlation coefficient which is defined as follows:

$$r = \frac{\sum (M_i - \bar{M})(\hat{M}_i - \bar{\hat{M}})}{\sqrt{\sum (M_i - \bar{M})^2 \sum (\hat{M}_i - \bar{\hat{M}})^2}} \quad [12]$$

The concordance index (C-index) is a metric used to assess the ability of a model to correctly rank pairs of subjects based on predicted survival times. The C-index is calculated as the proportion of concordant pairs, where a pair is considered concordant if the individual with the higher predicted survival time also has the longer actual survival time. It can be calculated using the following formula:

$$C = \sum_{i<j} I(\hat{T}_i > \hat{T}_j)\, I(T_i > T_j) / (\sum_{i<j} I(T_i \neq T_j) \quad [14]$$

Where and are the predicted survival time for individual i and j, and and are the actual survival times. A C-index value closer to 1.0 indicates better predictive accuracy, while a value of 0.5 indicates random predictions. MATLAB codes is available on GitHub at:

https://github.com/LeylaFarahani/Hybrid_PSO_SVM_Survival

Bioinformatic analysis

Enrichment analysis: In this section, we enriched the genes identified with the method that performed best.

Protein-protein interaction (PPI) Network: The PPIs for selected probes, identified using the survival hybrid PSO+SVR algorithm, were obtained from the STRING database (version 11.0) Cytoscape (version 3.7.2) was used as the network visualization software to construct the PPI network. The protein interactions were filtered to include only those with a confidence score of 0.700 or higher. Network characteristics including node degree, betweenness centrality, and clustering coefficient were calculated using built-in functions in Cytoscape and were used as selection criteria.

Functional enrichment analysis: The DAVID database (https://david.ncifcrf.gov/) was utilized to conduct a comprehensive functional enrichment analysis of the selected genes. The DAVID database is a widely recognized and reliable tool for functional enrichment analysis that incorporates the GO resources, which include the biological process, molecular function, and cellular component, as well as the KEGG resources. This increases the comprehensiveness and precision of the outcomes.

Drug-gene interaction network: In this study, the Drug Gene Interaction Database (DGIdb) was utilized to identify potential drug candidates that target the feature genes. The DGIdb is linked to 22 other relevant databases and retrieves information on target genes from 24 related databases. Only experimentally validated interactions were considered.

Ethics approval

This study was approved by the Ethical Committee of the Tarbiat Modares University (IR.MODARES.REC.1401.102).

Results

Comparison of the models

Table 1 presents the results of the two-step survival hybrid algorithm and the performance of the SVR with and without using PSO/GA. The first row of the table pertains to the utilizing of the PSO algorithm in the second step. As shown, this method selected 291 probe sets out of 1069 screened probes and it achieves a smaller MAE, and MSE values, with a compatible R^2 compared with the algorithm using GA. Moreover, the results indicated that, compared to the method without using PSO and GA, the two-step hybridized method provided a better performance with a smaller subset of probe sets.

Table 1: The performance criteria of the two-step algorithm with SVR and GA/PSO-SVR

Method	The number of used/selected features	MAE	MSE	R^2	C-index
PSO+SVR (linear kernel)	291	0.061	0.005	0.99	0.973
PSO+NN	293	0.294	0.141	0.63	0.644
GA+NN	529	0.401	0.262	0.99	0.607
GA+SVR (linear kernel)	517	0.299	0.135	0.99	0.821
SVR (linear kernel)	1069	0.306	0.158	0.58	0.571
SVR (RBF kernel)	1069	0.480	0. 493	0.45	0.543
NN	1069	0.666	0.699	0.19	0.417

PSO; Particle swarm optimization, SVR; Support vector regression, GA; Genetic algorithm, NN; Neural networks, C-index; Concordance index, MAE; Mean absolute error, MSE; Mean squared error, and RBF; Radial basis function.

Protein-protein interaction network analysis

The constructed PPI network, consisting of 200 nodes and 120 edges, was shown in Figure 1. The 11 genes with highest degree, including *RBM25, SMC3, PRPF40A, POLE, SRRT, BCLAF1, PDS5B, HNRNPR, JAK1, MED23,* and *SULT1A1* were selected.

GO and KEGG pathway analysis

The analysis of KEGG pathways (Fig.2) revealed that the ribosome pathway was significantly enriched among the 181 selected genes. The GO enrichment analysis results indicated that the molecular function GO terms of structural constituent of the ribosome and poly (A) RNA binding were enriched. The enriched cellular compartments that included cytosolic small ribosomal subunit, ribosome, membrane, nucleoplasm, respiratory chain complex IV, intracellular ribonucleoprotein complex, nucleolus, and small ribosomal subunit. The enriched biological process terms were translational initiation, SRP-dependent cotranslational protein targeting to membrane, viral transcription, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, rRNA processing, and translation.

Prediction error curve and concordance index for selected genes to predict survival

Figure 3 displays the prediction error curve for two models: one consisting solely of clinical variables, and the other model, which includes both clinical variables and the selected genes identified by the proposed method. As shown, the addition of the selected genes could decrease the prediction error of survival of the patients compared to the model based only on clinical variables.

Concordance index (C-index) values for two models are shown in Figure 4: one model contains only clinical variables, while the other model includes both clinical variables and the selected genes identified by the proposed method. Again, the model with clinical and selected genes provided a better C-index over time compared to the model including only clinical variables.

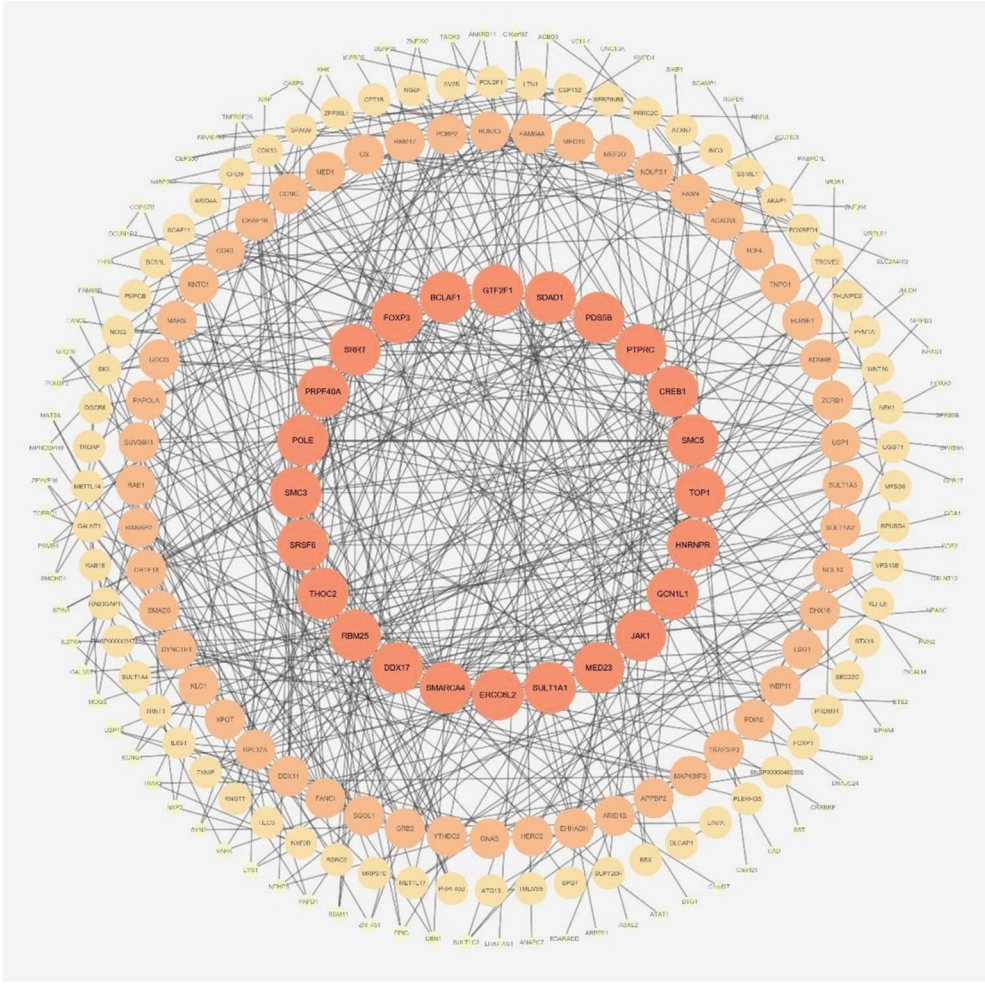

Fig.1: Protein-protein interaction (PPI) network of the top-ranked genes based on their degree centrality. The PPI network was constructed using STRING and visualized with Cytoscape software. Proteins are represented by nodes, and interactions between two proteins are described by edges. Nodes are ranked from the inside out based on their interaction degree.

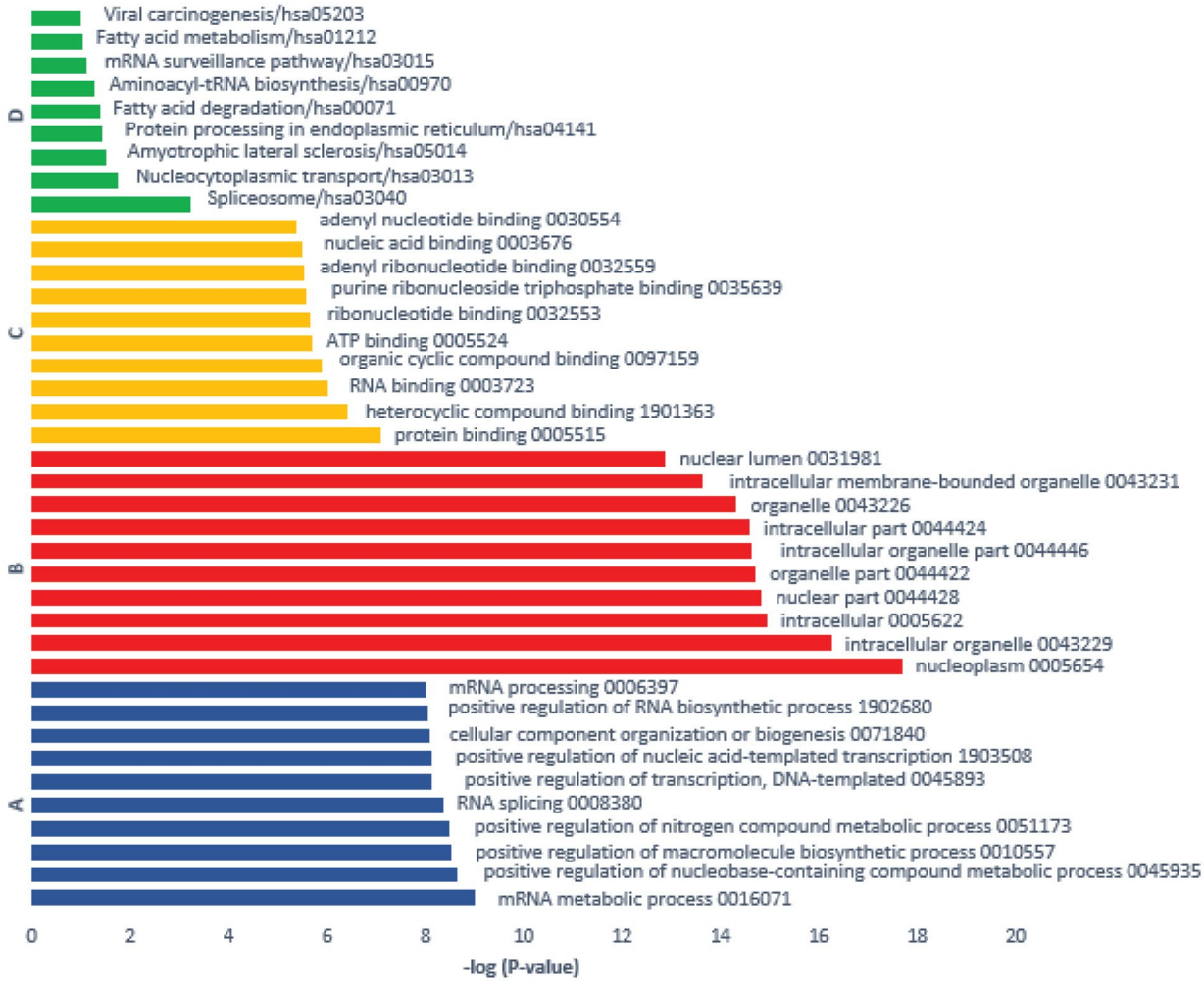

Fig.2: Gene ontology (GO) and KEGG pathway enrichment analysis of differentially expressed genes (DEGs) identified by the proposed two-step method. DEGs were derived by comparing high-risk and low-risk groups of OSCC patients. Results from the DAVID database is categorized as follows: (A; Blue) Biological process, (B; Red) Cellular component, (C; Yellow) Molecular, and (D; Green) KEGG pathway. OSCC; Oral squamous cell carcinoma.

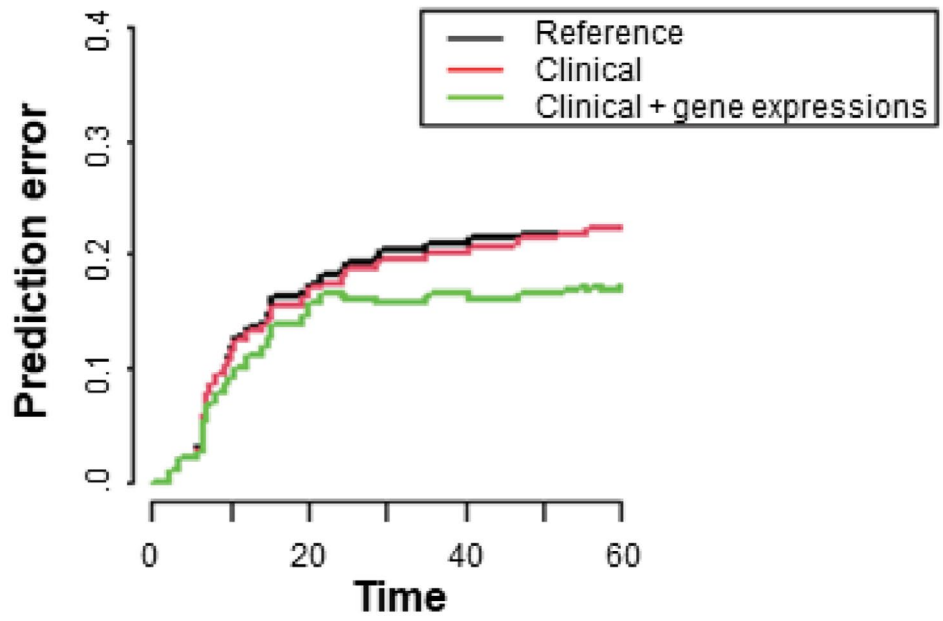

Fig.3: Prediction error curves compared the model with only clinical variables with the model integrating clinical variables and selected genes identified by the proposed method in data GSE41613.

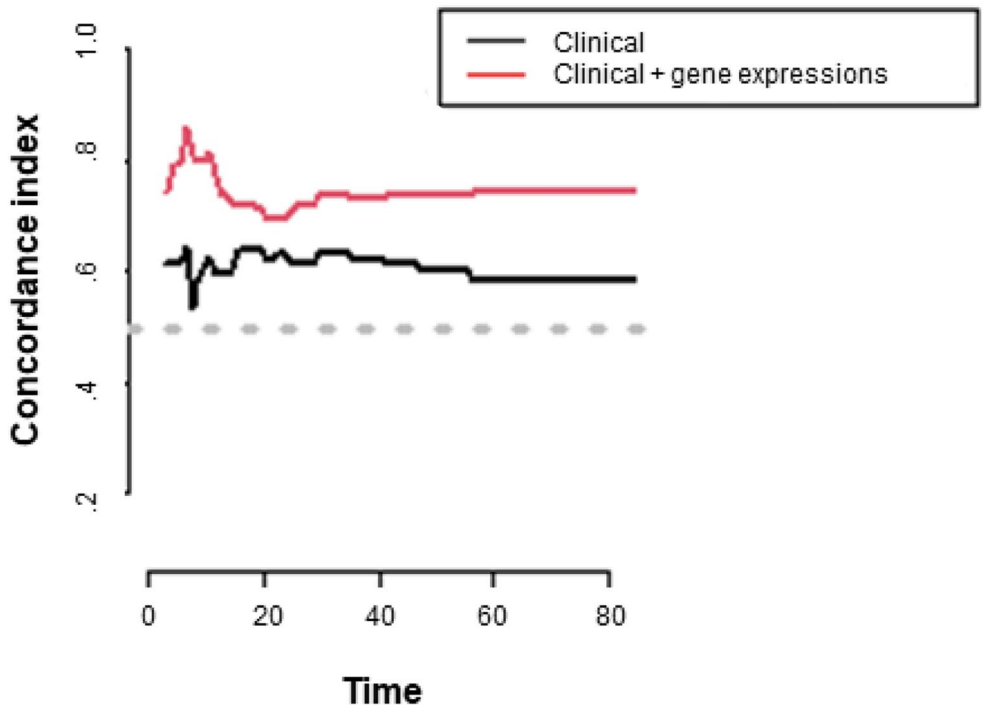

Fig.4: Concordance index compared between the model with only clinical variables and the model integrating clinical variables with selected genes identified by the proposed method in data GSE41613.

Drug network

Using DGIdb database, 187 candidate drugs were identified that target the feature genes associated with OC. These drugs may have a regulatory effect on the progression of the disease. The drug-gene interaction network was reconstructed using the identified drugs and feature genes, and visualized with Cytoscape v.3.7.2 software. The resulting network is presented in Supplement Figure 4 (See Supplementary Online Information at celljournal.org). Based on the constructed network, seven genes with known targeting drug were identified. Fludarabine, Adriamycin, Myricetin and Irinotecan and Cisplatin were the drugs which targeted more than one gene in the network. Cisplatin was considered as the most important drug in this network which targets three of our feature genes.

In Silico external validation

To biologically validate the selected genes, we utilized datasets GSE9844, GSE75538, and GSE37991. The results of the external validation were presented in Supplement Figure 5A-C (See Supplementary Online Information at celljournal.org), which displays the area under the ROC curve (AUC) for distinguishing OC patients from healthy controls using the selected genes. These results were used solely for the biological validation of the relevance of the selected genes in differentiating cancer from healthy controls, rather than for validating the predictive power of the survival model. The outcome suggests that the selected genes were effective in distinguishing OC patients from healthy controls.

To further validate the potential prognostic relevance of these genes, we analyzed an additional dataset, GSE42743, which included survival outcomes measured as time to death due to OSCC. Cox regression analysis was applied to calculate the prognostic index (βX) for each patient. Patients were then stratified into two distinct survival groups—high risk and low risk—based on the median prognostic index.

The survival analysis results are illustrated in Supplement Figure 6 (See Supplementary Online Information at celljournal.org), where survival curves between the two groups differ significantly, as determined by the log-rank test (Chi-square test statistic: 11.629; $P<0.001$). The mean survival time for the low-risk group (blue line) was 2073 days (SD: 249), compared to 657 days (SD: 93) for the high-risk group. These results suggest that the identified genes may have a critical role in influencing survival outcomes in OSCC patients.

Additionally, Brier score and C-index plots (Figs.S7, S8, See Supplementary Online Information at celljournal.org) were generated. The model incorporating the selected genes demonstrated a lower Brier score and a higher C-index over time, indicating that the selected genes provide valuable prognostic information.

To address the generalizability of the survival prediction model, we first trained and optimized the SVR using the GSE41613 dataset, including feature selection and hyperparameter tuning. Subsequently, we applied the trained SVR model directly to the same selected genes in GSE42743 dataset, ensuring that the feature set and

model parameters remained fixed across both datasets. We evaluated the model's performance on the GSE42743 dataset using the following metrics: MAE: 0.166, MSE: 0.123, R^2: 0.77 and C-index: 0.951.

These results demonstrate the model's ability to generalize to an independent dataset. Although a notable increase in error was observed compared to the training dataset (MAE: 0.061 in train vs. MAE: 0.166 in test, and similar trends in other metrics), this may be partially attributed to overfitting or differences in data distribution. Nevertheless, the model retained strong predictive performance and generalizability. Importantly, despite this drop, it still outperforms other existing methods, underscoring its effectiveness in survival prediction.

Discussion

This study proposed a new approach for gene selection in high dimension survival data based on two steps by hybridizing PSO/GA and SVR algorithms applying on the martingale residuals as risk index (survival outcome) with an application to identify prognostic biomarkers of OC. Our findings showed that the PSO-based method outperformed the GA-based ones.

We utilized a dataset including survival outcomes and high-dimensional feature space (gene expression signatures) to illustrate the implementation of the proposed method. To the best of our knowledge, this is the first study to hybridize heuristic algorithms and ML models address the feature selection problem while considering censored observations. Therefore, this study represents a pioneering attempt to hybridize heuristic algorithms (PSO and GA) with SVR for feature selection problem in survival data analysis, particularly in high dimension setting (i.e. $p>n$). Previous studies have shown that the evolutionary algorithms like GA and PSO can potentially lead to more accurate and efficient solutions to complex optimization problems in classification and regression problems (14, 22, 23). Our approach effectively selected a small subset of genes related to patient survival (considering censored observations) that were significantly enriched in signaling pathways, and highlighted how heuristic methods such as PSO and GA can improve the performance SVM model greatly. This finding is consistent with the results of other studies conducted in classification problems (4).

There have been few studies focused on identifying OC survival-related biomarkers. For instance, Tapak et al. (4) employed an indirect approach by utilizing autoencoders to extract features from comprehensive gene expression profiles, subsequently clustering patients based on these features to identify high-risk OC patients. However, their study framed the survival as a classification problem based on the time to OC development. In contrast, our hybrid model directly tackles the complexities of survival analysis by integrating PSO and SVR for feature selection, while also accounting for censored observations. This methodological advancement not only enhances the robustness of our findings but also establishes a new benchmark for future research in prognostic biomarker discovery in oral cancer. By effectively addressing the challenges associated with survival data, our approach offers a more nuanced understanding of patient outcomes compared to previous models.

We identified a small novel subset of prognostic biomarkers related to the OC, including *RBM25, SMC3, PRPF40A, POLE, SRRT, BCLAF1, PDS5B, HNRNPR, JAK1, MED23,* and *SULT1A1*. Some of these genes have been previously investigated and shown to be related to OC. For instance, França et al. reported high immunoexpression of *PDS5B* in OC (24). Ge et al. (25) have shown a significantly higher expression level of *POLE2* gene in OSCC compared to the normal tissues. A literature review related to the screened genes has been provided in Supplementary File S1 (See Supplementary Online Information at celljournal.org).

The drug-gene interaction network identified drugs such as Fludarabine, Adriamycin, Myricetin, Irinotecan, and Cisplatin, which target multiple genes in the network. Of these, cisplatin is most notably linked to oral cancer. Cisplatin is a type of chemotherapy medication that is employed to treat various cancers, such as testicular cancer, ovarian cancer, cervical cancer, bladder cancer, head and neck cancer, esophageal cancer, lung cancer, mesothelioma, brain tumors, and neuroblastomas (26). Cisplatin is frequently prescribed as induction, adjuvant, neoadjuvant, or palliative chemotherapy for OSCC (27). In this context, we presented an elucidative discourse on the contributions of these genes toward advancing our comprehension of OC progression and enhancing patient survival. At the heart of these potential prognostic biomarkers lies a diverse array of biological functions critical in cancer biology. RBM25, an RNA-binding protein, regulates alternative splicing, a process integral to cancer-related mRNA isoform generation. SMC3, a component of the cohesin complex, ensures genome stability by aiding in chromosome segregation and DNA repair (28). PRPF40A, another splicing regulator, influences pre-mRNA splicing, impacting gene expression patterns in OC cells (29).

While implementation of these ML-based biomarkers in clinical practice faces challenges, including obtaining large, high-quality datasets for training robust ML models and ensuring the interpretability of complex ML algorithms, their integration could revolutionize OC diagnostics.

Regulatory approval, a critical step in translating ML biomarkers, necessitates transparency, reliability, and adherence to ethical standards. Similarly, the integration of survival-related biomarkers requires extensive longitudinal studies and collaboration between biostatisticians and clinicians for accurate interpretation (30).

Early detection, often the linchpin in successful cancer treatment, stands to benefit profoundly. With these biomarkers, OC could be identified in its nascent stages,

enabling timely interventions that substantially improve survival rates and reduce the invasiveness of treatments. Moreover, the molecular intricacies uncovered in our research pave the way for personalized treatment strategies.

Our findings can be validated independently and reproduced by other researchers, which is crucial for establishing their reliability. We performed two types of validation:

Biological validation of selected genes, where we examined their prognostic relevance using binary datasets (GSE9844, GSE75538, and GSE37991) and survival dataset (GSE42743).

The binary datasets were used to confirm the discriminative power of the selected genes between tumor and normal tissues, while the survival dataset allowed us to categorize patients into high- and low-risk groups based on survival outcomes, thereby evaluating the prognostic utility of the selected genes. The suggested experimental validation method involves assessing gene expression in oral tumor samples by conducting studies to evaluate the expression levels of the final 11 genes identified in this study, comparing these levels with normal tissue through expression profiling to understand the variations within the tumor microenvironment, performing a survival correlation studies to analyze the expression patterns of the selected genes between oral tumor samples and adjacent normal tissue to identify dysregulated genes specific to oral tumors, and developing a prognostic panel that could be integrated into clinical practice to serve as a valuable tool for clinicians in making informed decisions about patient management.

Model validation, where we assessed the predictive performance of our trained SVR model on the GSE42743 dataset without retraining. By keeping the selected features and hyperparameters fixed, we ensured that the model's generalizability was fairly evaluated. The model's performance on GSE42743 was evaluated using MAE, MSE, R^2, and C-index, ensuring its robustness across independent data.

Although this study relied on a single dataset (GSE41613), this approach was chosen to avoid batch effects commonly encountered when integrating multiple datasets, which can obscure biological signals. To ensure the robustness of our findings, we validated our results using internal validation techniques and additional validation sets, enhancing the reliability and applicability of our study.

The proposed model for survival-related feature selection has several potential limitations and challenges as well. First of all, it is a complex model with many hyperparameters. If the model is not trained properly, it can overfit the training data and perform poorly on new data. This problem was handled through CV and use of GA for hyperparameter optimization to some extent. Nevertheless, an increase in error metrics was observed when the model was applied to the independent validation set (GSE42743). This discrepancy may be attributed to several factors such as the high dimensionality of microarray data combined with limited sample sizes, which inherently increase the risk of overfitting, as well as potential biological heterogeneity between datasets or technical differences (e.g., platform or lab conditions). While these factors are not uncommon in high-dimensional biomedical studies, they may contribute to reduced generalization performance. Although our model still outperforms other baseline methods in terms of predictive accuracy, we acknowledge that the observed increase in MAE and MSE on the test data and discrepancy between training and test results, strongly suggests overfitting and insufficient generalizability. Therefore, at this stage, we do not recommend this model for clinical application or decision-making, and its use should be limited to research purposes until further technical improvements and extensive external validations are performed. Future work may address this limitation by implementing more robust feature selection or dimensionality reduction techniques, applying batch effect correction methods to harmonize cross-dataset variability, and exploring the use of survival SVMs in combination with feature selection methods for high-dimensional, censored survival data, and comparing the results with our approach and other existing methods.

Second, computational complexity was a challenge. The proposed model is a hybrid model that combines PSO/GA with SVR, both of which are computationally expensive algorithms. Therefore, the model can be slow to train and evaluate, especially on large datasets. Third, the model is susceptible to potential feature selection bias. As the model uses a heuristic search algorithm to select features, the feature selection process is not guaranteed to be optimal. Additionally, the PSO algorithm may be biased towards certain types of features. Fourth, limitation is the availability of appropriate external validation data. Despite an extensive search, no other datasets were found that met our criteria for OSCC-related survival data with appropriate sample sizes and consistent platforms. As a result, only GSE42743 as a survival data was used for external validation. Future studies may benefit from newly deposited datasets in GEO or other repositories with similar criteria.

Despite these challenges, the proposed hybrid model demonstrates several strengths. By combining PSO/GA for feature selection with SVR for prediction, it effectively optimizes high-dimensional data while capturing complex non-linear relationships and handling censored data through martingale residuals. This approach not only enhances predictive performance but also outperforms conventional methods, making it a valuable tool for survival analysis and biomarker discovery, despite its computational demands. Future work could explore alternative feature selection strategies, ensemble methods, or even direct survival-SVM formulations for high-

dimensional data, in order to improve generalizability and minimize overfitting.

Conclusion

In this research, a combination of PSO and GA with SVR was proposed as a hybrid technique for selecting variables in survival data analysis. The efficacy of the proposed method was evaluated using a genuine dataset and compared to the conventional SVR approach. The findings showed that the PSO+SVR method had better performance than the standard SVR approach, suggesting that using PSO for feature selection in survival data is feasible. Additionally, the study recommends examining other heuristic approaches like ant colony or gray wolf optimization techniques in future research, based on the results of this hybrid method.

While the reliance on a single dataset limits generalizability, this approach minimizes the challenges associated with batch effects that often arise from integrating data from multiple sources. To address these concerns, we validated our findings using independent datasets, which showed consistent performance. Despite challenges related to model complexity and feature selection bias, our approach offers a promising framework for biomarker discovery in high-dimensional data.

However, due to substantial limitations in model generalizability and stability, and signs of overfitting in external validation, we emphasize that the current version of this model is not suitable for clinical application or decision-making. Further research with larger, multi-center and independent datasets, along with technical improvements, is essential to enhance clinical applicability.

Acknowledgments

The authors would like to thank Siamak Salimy and Amirhossein Ahmadieh-Yazdi for their assistance with the bioinformatics analyses. This research was financially supported by Tarbiat Modares University. The authors declare that they have no conflict of interest.

Authors' Contributions

A.K., L.T.; Conceptualization, Supervision, Writing-Review, and Editing. A.K.; Project administration, Resources, and Funding acquisition. L.T., L.N.F., M.A.; Methodology, Software, Formal analysis, and Data Curation. L.T., L.N.F.; Validation, Investigation, Writing-original Draft, and Visualization. All authors read and approved the final manuscript.

Reference

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71(3): 209-249.
2. Jain AK. Oral cancer screening: insights into epidemiology, risk factors, and screening programs for improved early detection. Cancer Screening and Prevention. 2024; 3(2): 97-105.
3. Falzone L, Lupo G, La Rosa GRM, Crimi S, Anfuso CD, Salemi R, et al. Identification of novel microrna and their diagnostic and prognostic significance in oral cancer. Cancers (Basel). 2019; 11(5): 610.
4. Tapak L, Ghasemi MK, Afshar S, Mahjub H, Soltanian A, Khotanlou H. Identification of gene profiles related to the development of oral cancer using a deep learning technique. BMC Med Genomics. 2023; 16(1): 35.
5. Joshi AA, Aziz RM. Soft computing techniques for cancer classification of gene expression microarray data: a three-phase hybrid approach. Computational intelligence for data analysis. Netherlands: Bentham Science Publishers; 2024; 92-113.
6. Tapak L, Afshar S, Afrasiabi M, Ghasemi MK, Alirezaei P. Application of genetic algorithm-based support vector machine in identification of gene expression signatures for psoriasis classification: a hybrid model. Biomed Res Int. 2021; 2021: 5520710.
7. Agrawal AK, Chakraborty G. On the use of acquisition function-based Bayesian optimization method to efficiently tune SVM hyperparameters for structural damage detection. Struct Contr Health Monit. 2021; 28(4): e2693.
8. Wang G, Carr TR, Ju Y, Li C. Identifying organic-rich marcellus shale lithofacies by support vector machine classifier in the Appalachian basin. Comput Geosci. 2014; 64: 52-60.
9. Gad AG. Particle swarm optimization algorithm and its applications: a systematic review. Arch Computat Methods Eng. 2022; 29(5): 2531-2561.
10. Amini F, Hu G. A two-layer feature selection method using genetic algorithm and elastic net. Expert Syst Appl. 2021; 166: 114072.
11. Kamala R, Thangaiah RJ. An improved hybrid feature selection method for huge dimensional datasets. IAES Int J Artif Intell. 2019; 8(1): 77.
12. Jayaraman R, Velumani B. A hybrid PSO-GA approach for biomarker discovery. Iioab Journal. 2015; 6(4): 107-119.
13. Marwai HA, AL-Gaphari GH. A hybrid feature selection method based on binary PSO algorithm for microarray data classification. Sana'a University Journal of Applied Sciences and Technology. 2024; 2(4): 375-380.
14. McKearnan SB, Vock DM, Marai GE, Canahuate G, Fuller CD, Wolfson J. Feature selection for support vector regression using a genetic algorithm. Biostatistics. 2023; 24(2): 295-308.
15. Arayeshgar M, Tapak L, Parami S, Alafchi B. Variable selection for recurrent events using heuristic approaches: identifying informative variables for rehospitalization in schizophrenia patients. J Biostat Epidemiol. 2023; 9(1): 103-130.
16. Alafchi B, Tapak L, Doosti H, Chesneau C, Roshanaei G. A two-step algorithm to estimate variable importance for multi-state data: an application to COVID-19. Comput Model Eng Sci. 2023; 135(3): 2047-2064.
17. Duan R, Fu H. Estimate variable importance for recurrent event outcomes with an application to identify hypoglycemia risk factors. Stat Med. 2015; 34(19): 2743-2754.
18. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. Biometrika. 1990; 77(1): 147-160.
19. Zbigniew M. Genetic algorithms+ data structures= evolution programs. Berlin: Springer Science & Business Media; 2013; 372-373.
20. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol. 1972; 34(2): 187-202.
21. Xie F, Ning Y, Yuan H, Goldstein BA, Ong MEH, Liu N, et al. Autoscore-survival: developing interpretable machine learning-based time-to-event scores with right-censored survival data. J Biomed Inform. 2022; 125: 103959.
22. Vijayashree J, Sultana HP. A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. Program Comput Soft. 2018; 44: 388-397.
23. Rustam Z, Kintandani P. Application of support vector regression in indonesian stock price prediction with feature selection using particle swarm optimisation. Model Simul Eng. 2019; 2019(1): 1-5.
24. França JA, Diniz MG, Bernardes VF, Costa-Silva RC, Souza RP, Gomez RS, et al. Cohesin subunits, STAG1 and STAG2, and cohesin regulatory factor, PDS5b, in oral squamous cells carcinomas. J Oral Pathol Med. 2017; 46(3): 188-193.
25. Ge S, Wang K, Meng Y, He Z, Yang X, Shang W, et al. Silenc-

ing POLE2 promotes apoptosis and inhibits proliferation of oral squamous cell carcinomas by inhibiting PI3K/AKT signaling pathway. Med Oncol. 2023; 40(10): 304.
26. Kim MJ, Lee JH, Kim YK, Myoung H, Yun PY. The role of tamoxifen in combination with cisplatin on oral squamous cell carcinoma cell lines. Cancer Lett. 2007; 245(1-2): 284-292.
27. Madhulaxmi M, Iyer K, Periasamy R, Gajendran P, Lakshmi T. Role of cisplatin in oral squamous cell carcinoma–a review. JAdv Pharm Edu Res 2017; 7(2): 39-42.
28. Yueh WT, Singh VP, Gerton JL. Maternal Smc3 protects the integrity of the zygotic genome through DNA replication and mitosis. Development. 2021; 148(24): dev199800.
29. Blijlevens M, Li J, van Beusechem VW. Biology of the mRNA splicing machinery and its dysregulation in cancer providing therapeutic opportunities. Int J Mol Sci. 2021; 22(10): 5110.
30. Ahmadieh-Yazdi A, Mahdavinezhad A, Tapak L, Nouri F, Taherkhani A, Afshar S. Using machine learning approach for screening metastatic biomarkers in colorectal cancer and predictive modeling with experimental validation. Sci Rep. 2023; 13(1): 19426.