Published online: 2024 September 1.

Research Article

Identifying Survival Subtypes of Esophageal Squamous Cell Carcinoma Patients: An Application of Deep Learning in Gene Expression Data Analysis

Zahra Kousehlou [1], Ebrahim HajiZadeh [1,*], Leili Tapak [2], Ahmad Shalbaf [3]

[1] Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2] Department of Biostatistics, School of Public Health and Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran
[3] Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[*] *Corresponding author:* Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. Email: hajizadeh@modares.ac.ir

Received 2024 February 18; **Revised** 2024 May 18; **Accepted** 2024 May 25.

Abstract

Background: Esophageal squamous cell carcinoma (ESCC) is one of the most lethal types of cancer. Late diagnosis significantly decreases patient survival rates.

Objectives: The study aimed to identify survival groups for patients with ESCC and find predictive biomarkers of time-to-death from ESCC using state-of-the-art deep learning (DL) and machine learning algorithms.

Methods: Expression profiles of 60 ESCC patients, along with their demographic and clinical variables, were downloaded from the GEO dataset. A DL autoencoder model was employed to extract lncRNA features. The univariate Cox proportional hazard (Cox-PH) model was used to select significant extracted features related to patient survival. Hierarchical clustering (HC) identified risk groups, followed by a decision trees algorithm which was used to identify lncRNA profiles. We used Python.3.7 and R.4.0.1 software.

Results: Inputs of the autoencoder were 8,900 long noncoding RNAs (lncRNAs), of which 1000 features were extracted. Out of the features, 42 lncRNAs were significantly related to time-to-death using the Cox-PH model and used as input for clustering of patients into high and low-risk groups (P-value of log-rank test = 0.022). These groups were then labeled for supervised HC. The C5.0 algorithm achieved an overall accuracy of 0.929 on the test set and identified four hub lncRNAs associated with time-to-death.

Conclusions: Novel discovered lncRNAs *lnc-FAM84A-1*, *LINC01866*, *lnc-KCNE4-2* and *lnc-NUDT12-4* implicated in the pathogenesis of death from ESCC. Our findings represent a significant advancement in understanding the role of lncRNAs on ESCC prognosis. Further research is necessary to confirm the potential and clinical application of these lncRNAs.

Keywords: Esophageal Squamous Cell Carcinoma, Deep Learning, Machine Learning, Survival, Gene Expression, Decision Trees

1. Background

Esophageal carcinoma (EC) is the sixth most common cause of cancer-related death and the seventh most common cancer worldwide (1). Esophageal squamous cell carcinoma (ESCC) is one of the main types of EC, comprising approximately 84% of all global cases of EC (2). Typically, ESCC is thought primarily to be a disease of the developing world (the Asian esophageal cancer belt, includes countries such as Iran, Turkey, Kazakhstan, and parts of China) (3).

Despite improvements in EC's prognosis due to advances in treatment methods such as chemotherapy, radiation therapy, and surgery, the 5-year survival rate remains very low (3). In Europe, the United States, and China, only 10 - 22% of patients show survival of more than 5 years after diagnosis (4). The low survival rate is due to the cancer being symptomless in its early stages, leading to late diagnosis and ineffective treatment (2). However, if that's diagnosed at an early stage, the survival rate can be as high as 85% (3).

Currently, no screening guidelines are available to identify the early stage of ESCC at the population level (5). Consequently, identifying risk factors through new diagnostic methods is crucial for ESCC screening, preventive measures and reducing the overall disease burden, as well as the development of treatment. The role of several factors including smoking, genetic family history, diet, alcohol, opium use and socioeconomic status in developing ESCC has been well-established (6).

Current clinical biomarkers are not suitable for prognosis and diagnosis of ESCC due to their low sensitivity and specificity (7). Recent studies have indicated that DNAs and RNAs including protein-coding RNA and non-coding RNA and proteins could be used as potential cancer biomarkers (8). Long noncoding RNAs (lncRNAs) are a large class of non-coding RNAs with a size exceeding 200 nucleotides and have an essential role in the progression and development of cancer by regulating genes related to cancer. Some lncRNAs have shown abnormal expression in tumor tissues in different stages of cancer (9).

Therefore, by discovering new lncRNAs affecting cancer, it is possible to reduce the complexity of cancer and help to better manage the treatment process. So, it is crucial to explore potential relationships and improve survival prediction using state-of-the-art models.

Deep learning (DL), an advanced computer technique, has experienced explosive growth in the fields of biomedical sciences and pattern recognition. This advanced computer technique has been successfully used in detecting and diagnosing various types of cancer due to its algorithms (10). Deep learning autoencoders (DL-autoencoder) are one type of DL (10) that have been successfully used to reduce high-dimensional gene expression (11, 12) and omics data (13) and predict patients' survival (10, 13-15).

Parallel to our study, Tapak et al utilized a DL-autoencoder approach to analyze gene expression profiles for extracting significant features. Their study focused on patients with oral cancer and used a different pipeline and did not consider lncRNAs (16).

In the current study the DL- autoencoder model was used to predict the prognostic factors of ESCC then created a prognostic stratification for estimating patients survival. Finally, a machine learning algorithm was employed to identify ideal biomarkers related to the prognosis of primary ESCC.

2. Objectives

The aim of this study was to discover lncRNAs markers associated with time-to-death from ESCC that have not been fully understood and remained obscure, using state-of-the-art machine learning and DL models.

3. Methods

A publicly available dataset of (ESCC) patients from the gene expression Omnibus (GEO) repository with accession series GSE53622 was utilized. The dataset was generated using the Affymetrix transcript version (microarray) with platform ID GPL18109.This dataset consisted of preprocessed expression data of 60 patients.

The quantile normalization, summarization and quality control of data were done using the Gene Spring software V11.5 (Agilent) (17). Time-to-death from ESCC in patients was considered as survival time, and the patients for whom death did not occur considered as censored.

In this study, we used a DL computational framework on lncRNA profiles related to time-to-death from ESCC in patients. For feature extraction, an auto-encoder framework was used as a DL implementation.

Typically, in an auto-encoder framework, the number of neurons in the first layer is equal to the number of input observations. Moving towards the center of the network, the number of neurons in each layer drops in some measure. The middle layer, called bottleneck layer, typically contains the fewest neurons, representing the extracted new neurons. The layers after the middle are a mirrored version of the layers before the middle one.

The Keras package (https://github.com/fchollet/keras) was utilized to build an autoencoder with three hidden layers (5000, 1000 and 5000 nodes). Each layer captures different levels of abstraction and complexity in the data. Rectified linear unit (ReLU), which is popular, was used as the activation function.

The activation function is used in neural networks to discover complex and non-linear patterns in data. This function is chiefly implemented in the hidden layers of a neural network. Its equation is represented as follows: $f(x) = \max(0, x)$, producing an output of x if x is positive and 0 otherwise.

Finally, the gradient descent approach with 10 epochs (iterations) and 50% dropout as reasonable starting points based on empirical experience with similar models were utilized to train the autoencoder as the learning algorithm. Each instance of training data is processed once by the learning algorithm for one epoch. The implementation codes are developed using Python 3.10 (18).

The 1000 lncRNAs that were extracted using autoencoder model for 60 patients were considered independent variables and the survival time of these people was considered as a dependent variable, then the univariate Cox proportional hazard (Cox-PH) model was used to find lncRNAs that have a significant relation with survival times, expressed as $h(t|xi) = h0(t)exp(bxi)$ (19). The R.4.0.1 software was used to select the significant lncRNAs ($P < 0.05$) (20).

These lncRNAs were then divided into two different groups using the hierarchical clustering algorithm (HCA) in the subsequent step. The group with a low median survival time was identified as the high-risk group and the other group with a high median survival time was identified as the low-risk group.

Hierarchical clustering algorithm is an unsupervised learning method that involves grouping data points into clusters based on their similarity. This clustering can be accomplished in two different ways. In this research, we used the agglomerative HC method, the most popular model, which involves combining small clusters to create larger clusters. The silhouette coefficient method was employed to establish the optimal number of clusters. The algorithm's parameters include minimum, maximum, average, and center distance (21).

The difference between the survival curves of these two groups was measured using the log-rank test (22, 23).

The C5.0 algorithm, which is used in supervised learning, is an extension of the ID3 (simple decision tree learning algorithm) and the C4.5 algorithm. It has improved in speed, memory, and efficiency compared to C4.5 and has proven its high detection accuracy in many fields of research. The accuracy criterion, which employs a confusion matrix, is used to compare machine learning models (24).

4. Results

In the presented study, 80% (n = 48) of patients were men. The number of patients who used tobacco and alcohol were 56.7% (n = 34) and 53.3% (n = 32), respectively. The clinical information of the patients is shown in Table 1. Thirty-three out of 60 patients died from ESCC. The mean and median follow-up time of the patients were 36.62 and 39.17 months (min=1.67 and max = 58.20 years), respectively. One- and three-year survival rates of the patients were 91% and 78%, respectively.

Table 1. Demographic and Clinical Information of two Identified Groups [a]

Index	Low Risk	High Risk	Total	P-Value
Gender				0.892
Total	21 (100)	39 (100)	60 (100)	
Female	4 (19.0)	8 (20.5)	12 (20.0)	
Male	17 (81.0)	31 (79.5)	48 (80.0)	
Use of tobacco				0.251
No	7 (33.3)	19 (48.7)	26 (43.3)	
Yes	14 (66.7)	20 (51.3)	34 (56.7)	
Use of alcohol				0.419
No	8 (38.1)	20 (51.3)	28 (46.7)	
Yes	13 (61.9)	19 (48.7)	32 (53.3)	
Tumor location				0.573
Lower	9 (42.9)	17 (43.6)	26 (43.3)	
Middle	11 (52.4)	17 (43.6)	28 (46.7)	
Upper	1 (4.8)	5 (12.8)	6 (10)	
Tumor grade				0.045
Poorly	3 (14.3)	14 (35.9)	17 (28.3)	
Moderately	12 (57.1)	22 (56.4)	34 (56.7)	
Well	6 (28.6)	3 (7.70)	9 (15.0)	
T stage				0.785
T1	2 (9.5)	2 (5.1)	4 (6.7)	
T2	3 (14.3)	4 (10.3)	7 (11.7)	
T3	16 (76.2)	32 (82.1)	48 (80.0)	
T4	0	1 (2.6)	1 (1.7)	
N stage				0.898
N0	10 (47.6)	19 (48.7)	29 (48.3)	
N1	8 (38.1)	12 (30.8)	20 (33.3)	
N2	2 (9.5)	7 (17.9)	9 (15.0)	
N3	1 ()4.8	1 (2.6)	2 (3.3)	
Arrhythmia				0.392
No	14 (66.7)	30 (76.9)	44 (73.3)	
Yes	7 (33.3)	9 (23.1)	16 (26.7)	
Pneumonia				0.722
No	20 (95.2)	37 (94.9)	57 (95.0)	
yes	1 (4.8)	2 (5.1)	3 (5.0)	
Anastomotic leak				0.650
No	21 (100)	38 (97.4)	59 (98.3)	
Yes	0	1 (2.6)	1 (1.7)	
Adjuvant therapy				0.978
No	8 (38.1)	15 (38.5)	23 (38.3)	
Yes	13 (61.9)	24 (61.5)	37 (61.7)	
Tnm stage				0.450
1	2 (9.5)	2 (5.1)	4 (6.7)	
2	11 (52.4)	19 (48.7)	30 (50.0)	
3	8 (38.1)	18 (46.2)	26 (43.3)	

[a] Values are expressed as No. (%).

Figure 1 illustrates the architecture of the autoencoder. The activity of the 1000 nodes from the bottleneck hidden layer was extracted as new lncRNAs.

Out of 1000 lncRNAs, 42 were statistically significant using the univariate Cox-PH model ($P < 0.05$) and were shown to be related to the survival of the patients.

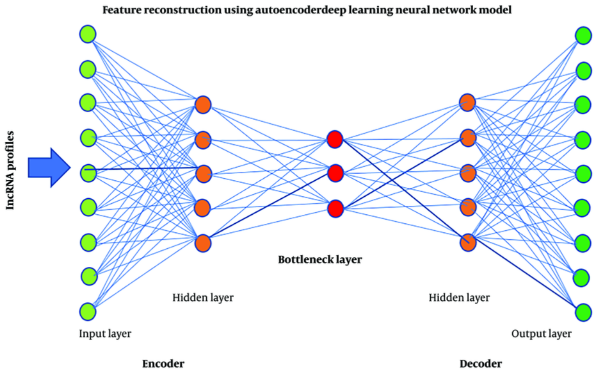


Figure 1. Architecture of the autoencoder

Using the silhouette indicator, the number of two clusters (k = 2) was optimal. Consequently, the lncRNAs were divided into two different groups. Figure 2 shows the heatmap of the 4 selected lncRNAs. Warmer tones like red and orange indicate over-expression of lncRNAs, while cooler tones like green represent under-expression of lncRNAs. High-risk patients (group 1) have warmer colors, indicating overexpression of lncRNAs. Table 1 displays the demographic and clinical information of low and high-risk groups identified through clustering utilizing 42 features.

Table 2 displays survival information of the two groups. The median survival time for the high-risk group was 29.87 months and for the low-risk group, it was more than 60 months. Furthermore, the survival analysis on the full data showed that the Kaplan-Meier survival curves in the median survival time in the high-risk group are significantly lower than the low-risk group (P = 0.022) (Figure 3).

Table 2. Survival Information of Two Identified Groups [a]

Subgroup	Patients	Events	Censor	Month	Median (Month)	SE
High risk	39 (0.65)	26 (66.67)	13 (33.33)	31.99 ± 3.36	29.87	1.55
Low risk	21 (0.35)	7 (33.33)	14 (66.67)	44.93 ± 4.33	-	-

[a] Values are expressed as No. (%) or mean ± SE.

Table 3 shows the results of employing different learning techniques. According to these results, the C5.0 and Chaid algorithms achieved a classification accuracy of 98.333 by selecting 4 genes and due to the simplicity of C5.0, it was employed in the continuation of the analysis. Using this method, 4 hub lncRNAs including *Lnc-FAM84A-1*, *LINC01866*, *Lnc-KCNE4-2*, *Lnc-NUDT12-4* were selected that were related to time-to-death from ESCC. This method with 4 selected lncRNAs provided total accuracy and AUC of 99.878% and 0.999, respectively.

Table 3. Results of Classification of High/Low Risk Survival Groups Using Machine Learning Models

Models	Fields [a]	Overall Accuracy	Total Accuracy (%)	AUC [b]
C5.0	4	98.333	99.878	0.999
Chaid	4	98.333	99.878	0.999
Quest	12	95.000	96.825	0.968
C & R tree	12	90.000	95.971	0.960

[a] Fields: The number of lncRNAs selected.

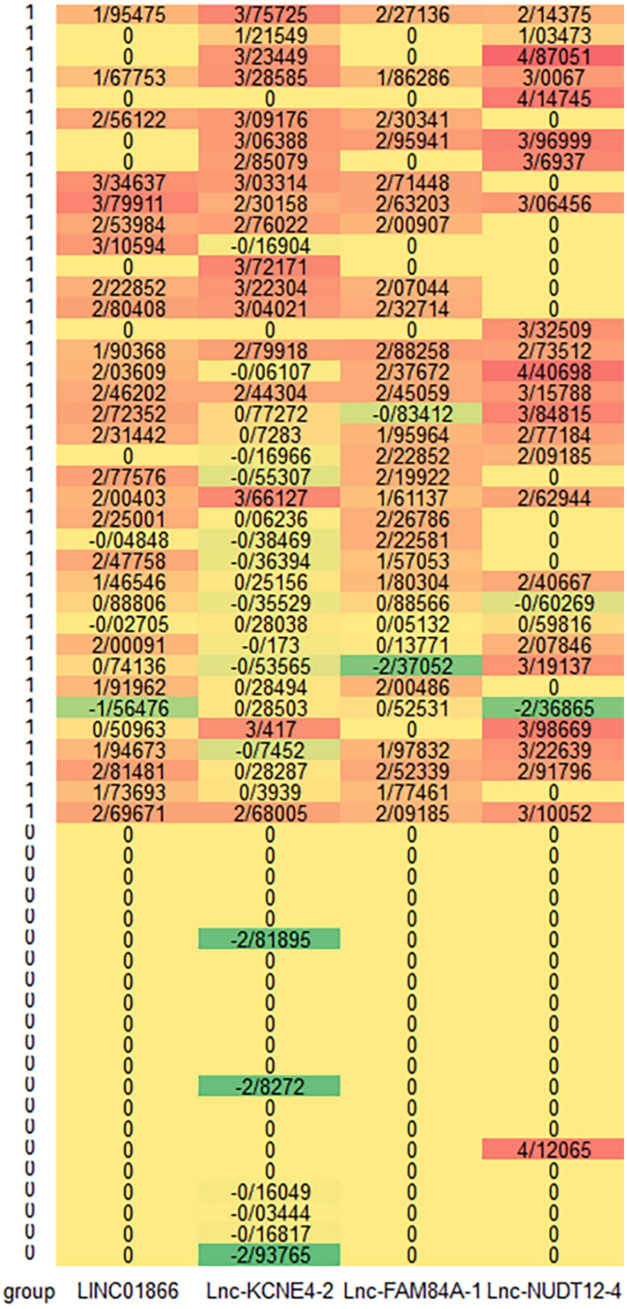
[b] Area under the ROC curve.

Table 4 shows the importance of 4 hub rank lncRNAs as over expression in high-risk patients.

Table 4. Hub Long Non-coding RNAs Identified by C5.0 Decision Tree Algorithm Method Through Variable Importance

Order	Ensembl Transcript ID	Nodes	Importance	Value in High-Risk Group
1	ENST00000450715.1	*Lnc-FAM84A-1*	0.30	Overexpressed
2	ENST00000421181.1	*LINC01866*	0.28	Overexpressed
3	ENST00000422118.1	*Lnc-KCNE4-2*	0.25	Overexpressed
4	ENST00000506337.1	*Lnc-NUDT12-4*	0.18	Overexpressed

5. Discussion



group	LINC01866	Lnc-KCNE4-2	Lnc-FAM84A-1	Lnc-NUDT12-4
1	1/95475	3/75725	2/27136	2/14375
1	0	1/21549	0	1/03473
1	0	3/23449	0	4/87051
1	1/67753	3/28585	1/86286	3/0067
1	0	0	0	4/14745
1	2/56122	3/09176	2/30341	0
1	0	3/06388	2/95941	3/96999
1	0	2/85079	0	3/6937
1	3/34637	3/03314	2/71448	0
1	3/79911	2/30158	2/63203	3/06456
1	2/53984	2/76022	2/00907	0
1	3/10594	-0/16904	0	0
1	0	3/72171	0	0
1	2/22852	3/22304	2/07044	0
1	2/80408	3/04021	2/32714	0
1	0	0	0	3/32509
1	1/90368	2/79918	2/88258	2/73512
1	2/03609	-0/06107	2/37672	4/40698
1	2/46202	2/44304	2/45059	3/15788
1	2/72352	0/77272	-0/83412	3/84815
1	2/31442	0/7283	1/95964	2/77184
1	0	-0/16966	2/22852	2/09185
1	2/77576	-0/55307	2/19922	0
1	2/00403	3/66127	1/61137	2/62944
1	2/25001	0/06236	2/26786	0
1	-0/04848	-0/38469	2/22581	0
1	2/47758	-0/36394	1/57053	0
1	1/46546	0/25156	1/80304	2/40667
1	0/88806	-0/35529	0/88566	-0/60269
1	-0/02705	0/28038	0/05132	0/59816
1	2/00091	-0/173	0/13771	2/07846
1	0/74136	-0/53565	-2/37052	3/19137
1	1/91962	0/28494	2/00486	0
1	-1/56476	0/28503	0/52531	-2/36865
1	0/50963	3/417	0	3/98669
1	1/94673	-0/7452	1/97832	3/22639
1	2/81481	0/28287	2/52339	2/91796
1	1/73693	0/3939	1/77461	0
1	2/69671	2/68005	2/09185	3/10052
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	-2/81895	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	-2/8272	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	4/12065
0	0	0	0	0
0	0	-0/16049	0	0
0	0	-0/03444	0	0
0	0	-0/16817	0	0
0	0	-2/93765	0	0

Figure 2. Heat-map of the 4 selected long noncoding RNAs (lncRNAs) using C5.0 decision tree algorithm related two identified survival groups (1= high-risk, 0 = low-risk).

In the present study, we identified four lncRNAs with prognostic significance in ESCC. According to the findings, *lnc-FAM84A-1* (ENST00000450715.1_transformed) was the first hub lncRNAs identified by the C5.0 model. These results indicated an overexpression of *lnc-FAM84A-1* in high-risk patients compared to the low-risk (25). *LINC01866*, *lnc-KCNE4-2* and *lnc-NUDT12-4* were the other lncRNAs found in this study.

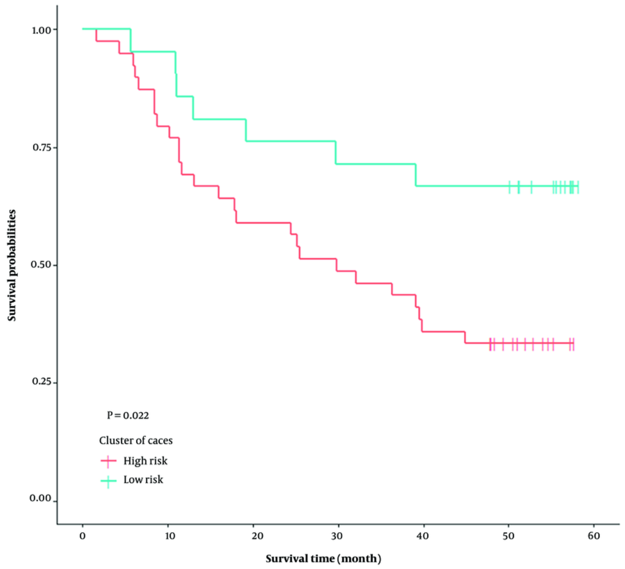


Figure 3. Kaplan Meier curve for two subgroups of survival time.

The mentioned lncRNAs have not been well studied and could be considered as novel biomarkers in ESCC. More research is needed to determine its specific function and biological significance.

The ability to predict a patient's time-to-death is critical for making informed treatment decisions, providing patients and their families with realistic expectations, and guiding end-of-life planning. Accurate prognostication is an essential aspect of managing patients with ESCC, one of the deadliest forms of cancer (26).

LncRNAs play a role in various aspects of cancer biology, including prognosis, diagnosis, tumorigenesis, and progression (27). They also have the ability to regulate various biological processes involved in cancer progression, such as cell proliferation, apoptosis, angiogenesis, and immune evasion. By modulating these processes, lncRNAs can affect the ability of cancer cells to survive and evade therapeutic interventions (28).

Since the expression of lncRNAs in cancer has been shown to correlate with overall survival (OS), metastasis, tumor stage, and tumor grade, these RNAs might serve as indicators for prognosis. For instance, HOTAIR as an important lncRNA, has been proven to be a prognosis biomarker of various cancer (29). In a study conducted by Svodoba et al., it was demonstrated that HOTAIR serves as a negative prognostic factor in colorectal cancer, exhibiting a sensitivity of 92.5%, a specificity of 67%, and an AUC of 0.8742 (30).

MALAT1 as another important lncRNA was also proved to have a role in the prognosis of different cancer types. It has been shown that high expression levels of this lncRNA are correlated with poor prognosis in breast cancer and hepatocellular carcinoma (31).

In a study by Cao et al., it was found that MALAT1 expression was significantly elevated in ESCC tissue compared to adjacent normal tissue samples ($P < 0.001$). Additionally, the level of MALAT1 was positively associated with the pT stage. Kaplan-Meier analysis revealed that high MALAT1 expression was correlated with poorer prognosis in ESCC patients (32).

CCAT2 is also a lncRNA with prognostic value and high expression levels of CCAT2 is associated with poor survival in ESCC. Recent studies have shown that these lncRNAs have potential value in predicting ESCC prognosis (33).

Integrating lncRNA expression data with gene mutations and DNA methylation profiles enhances understanding of ESCC tumorigenesis. This multi-omics approach identifies dysregulated pathways and biomarkers for personalized treatment strategies. Specific novel lncRNAs show promise as ESCC biomarkers, pending validation in larger cohorts. Their integration with molecular markers offers comprehensive insights into ESCC biology, advancing therapeutic targeting.

Future research should validate these lncRNAs in diverse ESCC cohorts to establish prognostic value and diagnostic assays. Integrating lncRNA expression into prognostic models could enhance outcome prediction, while further studies are needed to understand their roles in ESCC biology.

In this study, a large quantity of lncRNAs from ESCC patients was used to extract lncRNAs with an autoencoder framework. In similar studies the use of autoencoder, when compared with alternative methods, was more robust and much more efficient in identifying lncRNAs linked to survival (10). Then, we used a univariate Cox-PH model for the selection of significant lncRNAs.

We could not use the multivariate regression for this purpose because the number of unsupervised extracted lncRNAs (> 100) is more than the number of the sample size ($n = 60$) therefore, it is suggested to use penalized Cox regression model to select a subset of lncRNAs.

5.1. Conclusions

This study identified 4 hub lncRNAs including *lnc-FAM84A-1*, *LINC01866*, *lnc-KCNE4-2* and *lnc-NUDT12-4*, that have a role in the pathogenesis of developing ESCC. Further experimental investigations are required to well-understand the role of these lncRNAs.

Acknowledgements

The authors would like to thank Dr. S. Afshar for his expertise and assistance in the field of genetics.

Footnotes

Authors' Contribution: Z. K., conceived the research topic; Z. K., and E. H. and L. T. and A. Sh., explored that idea, performed the statistical analysis, and drafted the manuscript; S. A., participated in the interpretations and drafting of the manuscript. All authors read and approved the final manuscript.

Conflict of Interests Statement: The authors have no conflicts of interest to declare for this study.

Data Availability: The dataset presented in the study is available on request from the corresponding author during submission or after publication.

Ethical Approval: This study used a publicly available data set. All methods were carried out in accordance with relevant guidelines and regulations.

Funding/Support: present study was financially supported by Tarbiat Modares University Faculty of Medical Sciences, Tehran, Iran. (Grant number: Med/69161).

References

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):209-49. [PubMed ID: 33538338]. https://doi.org/10.3322/caac.21660.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;**68**(6):394-424. [PubMed ID: 30207593]. https://doi.org/10.3322/caac.21492.
3. Chela HK, Gangu K, Ertugrul H, Juboori AA, Daglilar E, Tahan V. The 8th Wonder of the Cancer World: Esophageal Cancer and Inflammation. *Dis*. 2022;**10**(3). [PubMed ID: 35892738]. [PubMed Central ID: PMC9326664]. https://doi.org/10.3390/diseases10030044.
4. Dubecz A, Gall I, Solymosi N, Schweigert M, Peters JH, Feith M, et al. Temporal trends in long-term survival and cure rates in esophageal cancer: a SEER database analysis. *J Thorac Oncol*. 2012;**7**(2):443-7. [PubMed ID: 22173700]. https://doi.org/10.1097/JTO.0b013e3182397751.
5. Uhlenhopp DJ, Then EO, Sunkara T, Gaduputi V. Epidemiology of esophageal cancer: update in global trends, etiology and risk factors. *Clin J Gastroenterol*. 2020;**13**(6):1010-21. [PubMed ID: 32965635]. https://doi.org/10.1007/s12328-020-01237-x.
6. Gholipour M, Islami F, Roshandel G, Khoshnia M, Badakhshan A, Moradi A, et al. Esophageal Cancer in Golestan Province, Iran: A Review of Genetic Susceptibility and Environmental Risk Factors. *Middle East J Dig Dis*. 2016;**8**(4):249-66. [PubMed ID: 27957288]. [PubMed Central ID: PMC5145292]. https://doi.org/10.15171/mejdd.2016.34.
7. Yang Y, Huang X, Zhou L, Deng T, Ning T, Liu R, et al. Clinical use of tumor biomarkers in prediction for prognosis and chemotherapeutic effect in esophageal squamous cell carcinoma. *BMC Cancer*. 2019;**19**(1):526. [PubMed ID: 31151431]. [PubMed Central ID: PMC6544972]. https://doi.org/10.1186/s12885-019-5755-5.
8. Zhang X, Sun XF, Shen B, Zhang H. Potential Applications of DNA, RNA and Protein Biomarkers in Diagnosis, Therapy and Prognosis for Colorectal Cancer: A Study from Databases to AI-Assisted Verification. *Cancers (Basel)*. 2019;**11**(2). [PubMed ID: 30717315]. [PubMed Central ID: PMC6407036]. https://doi.org/10.3390/cancers11020172.
9. Bolha L, Ravnik-Glavac M, Glavac D. Long Noncoding RNAs as Biomarkers in Cancer. *Dis Markers*. 2017;**2017**:7243968. [PubMed ID: 28634418]. [PubMed Central ID: PMC5467329]. https://doi.org/10.1155/2017/7243968.
10. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res*. 2018;**24**(6):1248-59. [PubMed ID: 28982688]. [PubMed Central ID: PMC6050171]. https://doi.org/10.1158/1078-0432.CCR-17-0853.
11. Shinde K, Itier V, Mennesson J, Vasiukov D, Shakoor M. Dimensionality reduction through convolutional autoencoders for fracture patterns prediction. *Appl Mathematical Mod*. 2023;**114**:94-113.
12. Kabir MF, Chen T, Ludwig SA. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*. 2023;**3**:100125.
13. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in

High-Risk Neuroblastoma. *Front Genet*. 2018;**9**:477. [PubMed ID: 30405689]. [PubMed Central ID: PMC6201709]. https://doi.org/10.3389/fgene.2018.00477.

14. Takahashi S, Asada K, Takasawa K, Shimoyama R, Sakai A, Bolatkan A, et al. Predicting Deep Learning Based Multi-Omics Parallel Integration Survival Subtypes in Lung Cancer Using Reverse Phase Protein Array Data. *Biomol*. 2020;**10**(10). [PubMed ID: 33086649]. [PubMed Central ID: PMC7603376]. https://doi.org/10.3390/biom10101460.
15. Yu J, Wu X, Lv M, Zhang Y, Zhang X, Li J, et al. A model for predicting prognosis in patients with esophageal squamous cell carcinoma based on joint representation learning. *Oncol Lett*. 2020;**20**(6):387. [PubMed ID: 33193847]. [PubMed Central ID: PMC7656101]. https://doi.org/10.3892/ol.2020.12250.
16. 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. [PubMed ID: 36849997]. [PubMed Central ID: PMC9972685]. https://doi.org/10.1186/s12920-023-01462-6.
17. Saintigny P, Zhang L, Fan YH, El-Naggar AK, Papadimitrakopoulou VA, Feng L, et al. Gene expression profiling predicts the development of oral cancer. *Cancer Prev Res (Phila)*. 2011;**4**(2):218-29. [PubMed ID: 21292635]. [PubMed Central ID: PMC3074595]. https://doi.org/10.1158/1940-6207.CAPR-10-0155.
18. Van Rossum G, Drake FL. *Python/C Api Manual-Python 3*. Scotts Valley, California: CreateSpace; 2009.
19. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1972;**34**(2):187-202.
20. R Core Team. *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/.
21. Murtagh F, Contreras P. Methods of hierarchical clustering. *arXiv*. 2011;**Preprint**.
22. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res*. 2010;**1**(4):274-8. [PubMed ID: 21455458]. [PubMed Central ID: PMC3059453]. https://doi.org/10.4103/0974-7788.76794.
23. Khalkhali HR, Gharaaghaji R, Valizadeh R, Kousehlou Z, Ayatollahi H. Ten Years' Survival in Patients with Cervical Cancer and Related Factors in West Azerbaijan Province: Using of Cox Proportion Hazard Model. *Asian Pac J Cancer Prev*. 2019;**20**(5):1345-51. [PubMed ID: 31127888]. [PubMed Central ID: PMC6857867]. https://doi.org/10.31557/APJCP.2019.20.5.1345.
24. Agaoglu M. Predicting instructor performance using data mining techniques in higher education. *Ieee Access*. 2016;**4**:2379-87.
25. Chang SC, Lin WL, Chang YF, Lee CT, Wu JS, Hsu PH, et al. Glycoproteomic identification of novel plasma biomarkers for oral cancer. *J Food Drug Anal*. 2019;**27**(2):483-93. [PubMed ID: 30987719]. [PubMed Central ID: PMC9296197]. https://doi.org/10.1016/j.jfda.2018.12.008.
26. Mittal SK, Abdo J, Adrien MP, Bayu BA, Kline JR, Sullivan MM, et al. Current state of prognostication, therapy and prospective innovations for Barrett's-related esophageal adenocarcinoma: a literature review. *J Gastrointest Oncol*. 2021;**12**(4):1197-214. [PubMed ID: 34532080]. [PubMed Central ID: PMC8421895]. https://doi.org/10.21037/jgo-21-117.
27. Zhu Y, Zhang F, Zhang S, Yi M. Predicting latent lncRNA and cancer metastatic event associations via variational graph auto-encoder. *Methods*. 2023;**211**:1-9. [PubMed ID: 36709790]. https://doi.org/10.1016/j.ymeth.2023.01.006.
28. Shirahama S, Miki A, Kaburaki T, Akimitsu N. Long Non-coding RNAs Involved in Pathogenic Infection. *Front Genet*. 2020;**11**:454. [PubMed ID: 32528521]. [PubMed Central ID: PMC7264421]. https://doi.org/10.3389/fgene.2020.00454.
29. Endo H, Shiroki T, Nakagawa T, Yokoyama M, Tamai K, Yamanami H, et al. Enhanced expression of long non-coding RNA HOTAIR is associated with the development of gastric cancer. *PLoS One*. 2013;**8**(10). e77070. [PubMed ID: 24130837]. [PubMed Central ID: PMC3795022]. https://doi.org/10.1371/journal.pone.0077070.
30. Svoboda M, Slyskova J, Schneiderova M, Makovicky P, Bielik L, Levy M, et al. HOTAIR long non-coding RNA is a negative prognostic factor not only in primary tumors, but also in the blood of colorectal cancer patients. *Carcinogenesis*. 2014;**35**(7):1510-5. [PubMed ID: 24583926]. https://doi.org/10.1093/carcin/bgu055.
31. Chou J, Wang B, Zheng T, Li X, Zheng L, Hu J, et al. MALAT1 induced migration and invasion of human breast cancer cells by competitively binding miR-1 with cdc42. *Biochem Biophys Res Commun*. 2016;**472**(1):262-9. [PubMed ID: 26926567]. https://doi.org/10.1016/j.bbrc.2016.02.102.
32. Cao X, Zhao R, Chen Q, Zhao Y, Zhang B, Zhang Y, et al. MALAT1 might be a predictive marker of poor prognosis in patients who underwent radical resection of middle thoracic esophageal squamous cell carcinoma. *Cancer Biomark*. 2015;**15**(6):717-23. [PubMed ID: 26406400]. https://doi.org/10.3233/CBM-150513.
33. Zhang X, Xu Y, He C, Guo X, Zhang J, He C, et al. Elevated expression of CCAT2 is associated with poor prognosis in esophageal squamous cell carcinoma. *J Surg Oncol*. 2015;**111**(7):834-9. [PubMed ID: 25919911]. https://doi.org/10.1002/jso.23888.