MATHEMATICAL MODELING AND COMPUTING, Vol. 12, No. 2, pp. 525–531 (2025)

A robust network topology approach for survival analysis

Wan Yusoff W. N. S.[1], Muhammad N.[1], Abd Mutalib S. S. S.[2], Zakaria Z. A.[3]

[1]*Pusat Sains Matematik, University Malaysia Pahang Al-Sultan Abdullah, Lebuhraya Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia*
[2]*Faculty of Computer Science and Mathematics, University Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia*
[3]*Faculty of Informatics and Computing, University Sultan Zainal Abidin, Kuala Terengganu, Malaysia*

(Received 18 December 2024; Revised 29 May 2025; Accepted 30 May 2025)

Network topology can be used to simplify the complexity of the datasets. We are exploring its function in performing survival analysis to identify the most important factors that contributed to the survival time from diagnosis to death. This technique has the potential to illustrate easily some types of complex interactions in dataset. Then, based on those interactions, the most important factors in survival analysis are identified. However, network topology based on classical estimator is extremely sensitive to outlying observations, i.e., conclusions drawn from contaminated network topology could be misleading. Hence, in this paper, the classical estimator, i.e., classical correlation matrix of network topology is substituted with robust estimator, i.e., robust correlation matrix which is developed based on Index Set Equality (ISE). Then, the interpretation of that robust network topology is delivered by using centrality measure, i.e., degree centrality. A case study of the survival time for cervical cancer patients is presented and discussed. Robust network topology revealed that the most important factors that influence the survival of cervical cancer patients is stage at diagnosis (STG). The higher stage of cervical cancer led to shorter survival time of cervical cancer patients. Consequently, early diagnosis is very important. Early diagnosis of cancer allows early intervention to try to slow or prevent cancer development and lethality, hence, the survival improves.

Keywords: *network topology; robust estimator; survival analysis; cervical cancer.*
2010 MSC: 62N02, 93B35 **DOI:** 10.23939/mmc2025.02.525

1. Introduction

Survival analysis is a statistical method used to analyze and interpret data on the time until an event of interest occurs. The event could be anything from the time until a patient relapses, the failure time of a mechanical part, or the duration until a person finds employment and etc. Survival analysis was initially developed in biomedical sciences to understand the beginning of certain diseases but is now used in engineering, insurance, and other disciplines. This paper focused on cancer studies where survival time was calculated from time of diagnosis to death for cancer deaths or to date of last contact or death from other causes for censored patients [1]. A unique feature of survival data is that typically not all patients experience the event (example: death) by the end of the observation period, so the actual survival times for some patients are unknown. There are many factors that influence the survival period, see for example [2–4]. Schober and Vetter [5] mentioned that the most common statistical techniques used to analyze the survival data are the Kaplan–Meier estimator, log-rank test, and the Cox proportional hazards (PH) model. These techniques also have been discussed in [6–9]. These three methods are examples of univariate analysis; they describe the survival with respect to the factor under investigation, but unavoidably ignore the impact of any others [10]. Hence, in order to present more

This research was funded by a grant from the International Matching Grant, RDU222708/ UIC221521 of University Malaysia Pahang Al-Sultan Abdullah.

details on the survival time with respect to several factors simultaneously, multivariate analysis will be performed in this study. Multivariate analysis refers to set of statistical techniques that simultaneously look at three or more variables with the aim of identifying or clarifying the relationships between them [11]. Due to that and also the complexity of underlying data sets, multivariate analysis requires much computational effort. Therefore, to simplify the complexity of the multivariate analysis, network topology approach will be used in this study. However, network topology based on classical estimator is extremely sensitive to outlying observations, i.e., outliers, conclusions drawn from contaminated network topology could be misleading [12, 13]. Hence, in order to overcome this problem, the classical estimators of network topology will be substituted with robust estimators [14, 15]. Various robust estimators such as M-estimator, Minimum Volume Ellipsoid (MVE) estimator, Minimum Covariance Determinant (MCD) estimator, Fast-MCD (FMCD) estimator, Minimum Variance Vector (MVV), Covariance Matrix Equality (CME) and Index Set Equality (ISE) have been presented in the previous studies [16, 17]. Mutalib et al. [17] mentioned that ISE is as effective as FMCD and MVV and have a lower computation time. Hence, in this study, ISE will be used to substitute the classical correlation matrix. An example of the survival time for cervical cancer patients will be discussed to illustrate the structure of network topology and a recommendation will be presented. The rest of the paper is organized as follows. In the Section 2, we present the methodology of network topology, followed by the results and discussion of corresponding example in Section 3. At the end, this paper will be closed with a conclusion in Section 4.

2. Methodology

This section will discuss the characteristics of cervical cancer dataset used in this study and also robust network topology will be proposed.

2.1. Case study: survival time of cervical cancer

Table 1. Description of variables used.

Variable	Description
TIME	Survival time from diagnosis of cervical cancer to death (in months)
AGECAT	Age at diagnosis
	$\leqslant 39 = 0$
	$40 - 49 = 1$
	$50 - 59 = 2$
	$\geqslant 60 = 3$
STG	Stage at diagnosis
	I = 0
	II = 1
	III – IV = 2

An example used in this study is about the survival time for cervical cancer patients. There are 120 patients with two predictor variables involved in this study. The data was retrieved from local hospital in Malaysia. An event of interest for this data is survival time *(TIME)*. Those variables are age at diagnosis *(AGECAT)* and stage at diagnosis *(STG)*. In this study, qualitative variables which are ethnicity *(ETH)*, lymph node involvement *(LN)*, histologic type *(HIS)*, primary treatment *(PT)* and distant metastasis *(DM)* cannot be used because of singularity problem that related to the method used.

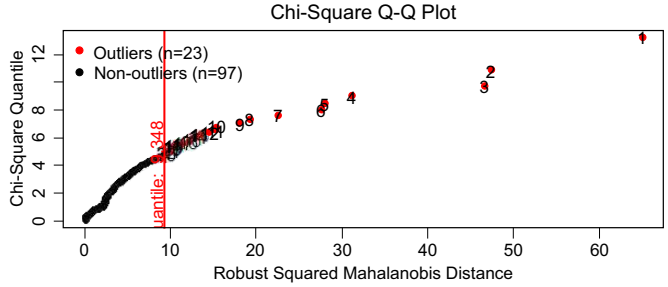

Fig. 1. Illustration of outliers' situation of cervical cancer patients.

For the analysis, a response variable, i.e., survival time *(TIME)* is added to determine which predictor variables give significant influence to the response variable. The description of variables used as in Table 1. Meanwhile, Figure 1 illustrates the outliers' situation in this dataset by using Mahalanobis distance in R Programming. Mahalanobis distance is the most commonly used distance metric to detect outliers for the multivariate setting [18]. Figure 1 shows that the cervical cancer dataset used in this study contaminated with the outliers where 23 out of 120 patients are the outliers. Croux and Haesbroeck [12] and Fitrianto and Midi [13] mentioned that any conclusion drawn from contaminated data could be misleading.

2.2. Robust network topology

Network topology starts with classical correlation matrix followed by transforming it into a distance matrix [19]. From this matrix, a minimum spanning tree (MST) is constructed as suggested Kruskal [20], by using Kruskal's algorithm provided in Matlab version R2018b. Kruskal's algorithm performed better than Prim's algorithm in terms of computational speed [21]. From MST, we construct the network topology of all variables. This is a simplification of the complex system of corresponding correlation matrix which will be used to summarize the most important information. The visualization of MST can be made possible by using the open source called 'Pajek' [22]. In order to make the network topology more attractive and easier to interpret, we use the Kamada–Kawai procedure provided in Pajek. The interpretation of that network will be delivered by using the centrality measure, i.e., degree centrality measure [22].

In this paper, robust network topology is proposed by substituting the classical correlation matrix with robust correlation matrix which developed based on ISE. The robust correlation matrix was obtained by using this following algorithm [23,24]. This algorithm was implemented in R Programming.

Step 1: Choose an arbitrary subset H_{old} containing h different observations, where h is the smallest integer $\geqslant (n+p+1)/2$, where p is the number of variables and n is the sample size. H_{old} is an initial subset containing h observations were chosen from n observations.

Step 2: Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}.

Step 3: Compute $d^2_{H_{old}}(i) = \big(X_i - \bar{X}_{H_{old}}\big)' S^{-1}_{H_{old}} \big(X_i - \bar{X}_{H_{old}}\big)$ for $i = 1, 2, \ldots, n$. $d^2_{H_{old}}(i)$ is the distance for each observation in H_{old}.

Step 4: Sort $d^2_{H_{old}}(i)$ for $i = 1, 2, \ldots, n$ in increasing order $d^2_{H_{old}}(\pi(1)) \leqslant d^2_{H_{old}}(\pi(2)) \leqslant \ldots \leqslant d^2_{H_{old}}(\pi(n))$ where π is a permutation on $\{1, 2, \ldots, n\}$.

Step 5: Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(h)}\}$ and then calculate $\bar{X}_{H_{new}}$,$S_{H_{new}}$ and $d^2_{H_{new}}(i)$ for $i = 1, 2, \ldots, n$. H_{new} is a new subset obtained from Step 4 where only h observations are chosen after $d^2_{H_{old}}(i)$ is sorting. $\bar{X}_{H_{new}}$ and $S_{H_{new}}$ is the sample mean and covariance matrix of H_{new}. While $d^2_{H_{new}}(i)$ is the distance for each observation in H_{new}.

Step 6: If $I_{new} \neq I_{old}$, let $H_{old} := H_{new}$, calculate $\bar{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stopped. I_{old} and I_{new} is the index sets correspond to the observations in H_{old} and H_{new}, respectively.

Then, robust correlation matrix is obtained based on covariance matrix, $\mathbf{S}_{ISE}$ which obtained in the last step. According to Singh [25], classical estimators fail to be efficient in practical scenarios when data is riddled with outliers, hence, robust estimators approaches which are insensitive to outliers are used in such cases.

Since the cervical cancer dataset used in this study is contaminated with the outliers, then, the use of robust correlation matrix in network topology is a necessity to ensure the validity of conclusion drawn from the analysis.

3. Results & discussion

The correlation matrix of survival data consists of 3 variables as nodes connected by $((3-1)\times(3/2)) = 3$ links each of which corresponds to the correlation between two different nodes. However, by using the MST as in Figure 2, we only have to consider $[3-1] = 2$ links. The number of links shows that the complexity of multivariate analysis has been reduced. MST is a subgraph that connects all the variables (nodes) whose total weight, i.e., total distance i. Figure 2 shown that, STG relate directly to the survival time *(TIME)* of cervical cancer patients.

In Figure 3, the network topology where the colour of the node (predictor variables) represents the rank of importance based on degree centrality is presented. The colours used in this analysis, ordered decreasingly in terms of the rank of importance: green and black. The higher the score of the centrality measures of a particular node, the more dominance that node is. From Figure 3 *STG* has the highest

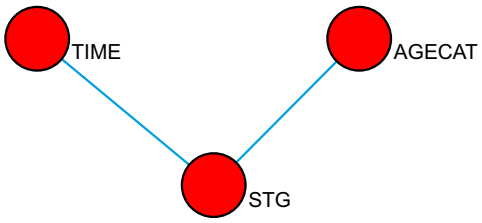

Fig. 2. Minimum Spanning Tree (MST) – illustrate which variable relate significantly to the response variable, *TIME*.

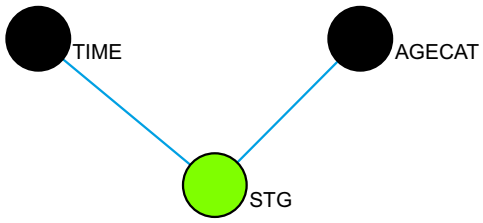

Fig. 3. Degree centrality – indicates the level of importance of a variable in terms of its connectivity with other variables and provides information on how many edges incident upon a given variable.

(green node) number of links, i.e., 2 links in the network. It plays the most important role in the network. This means that *TIME* is strongly influenced by the *STG*.

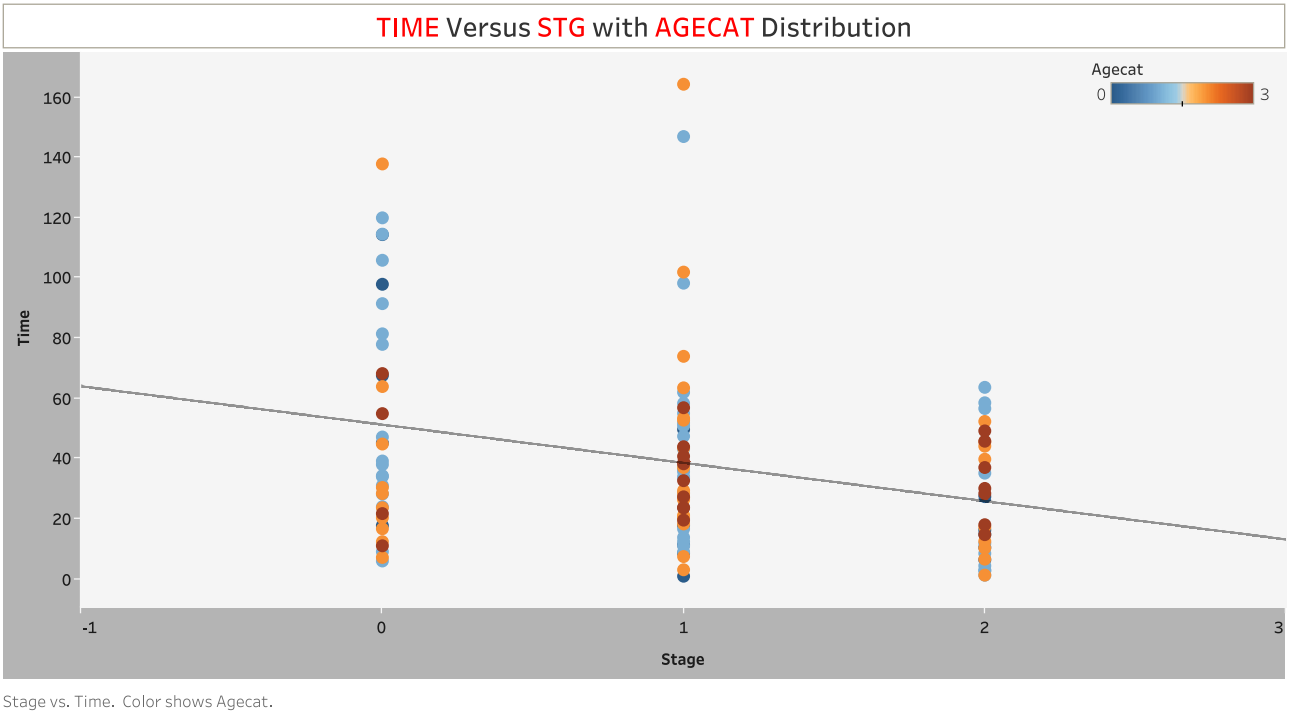

Fig. 4. Scatter plot: *TIME* versus *STG* with *AGECAT* distribution.

P-value: 0.0009915
Equation: Time = -12.7046*Stage + 51.2856

Coefficients

Term	Value	StdErr	t-value	p-value
Stage	-12.7046	3.7616	-3.37746	0.0009915
intercept	51.2856	4.58134	11.1945	< 0.0001

Fig. 5. Description of scatter plot's linear trend.

Since *AGECAT* is not directly influenced the survival time *(TIME)* of cervical cancer patients, a scatter plot between *TIME* and *STG* with distribution of *AGECAT* was constructed by using Tableau as in Figure 4 to illustrate their relationships. Figure 4 shows the negative trend between the *TIME* and *STG*, indicate that the survival time *(TIME)* of cervical cancer patients decreases when stage at diagnosis *(STG)* increases. This finding supported by Tabatabaei et al. [26] who mentioned that higher stage of cervical cancer led to shorter survival. Figure 4 also shows that age of patients at diagnosis *(AGECAT)* influenced indirectly the survival time *(TIME)* where the older diagnose patients led to shorter survival which implied that early diagnosis is very important. Early diagnosis of cancer allows early intervention to try to slow or prevent cancer development and lethality, hence, the survival improves [27]. Figure 5 illustrated the description of linear trend on the scatter plot between *TIME* and *STG*. It can be seen that the regression model of scatter plot is significant to use to predict the survival time since P-value (0.0009915) is less than

standard significance level, α which is 0.05. This finding justified that *STG* strongly influenced the survival time of cervical cancer patients.

4. Conclusion

Based on the analysis on *MST* in Figure 2, we learn that the survival time of cervical cancer patients *(TIME)* is directly influenced by *STG*. Further analysis based on degree centrality measure justified the finding by *MST* that *TIME* is strongly influenced by the *STG*. Furthermore, scatter plot between *TIME* and *STG* with *AGECAT* distribution demonstrated that higher stage of cervical cancer and older diagnose patients led to shorter survival time of cervical cancer patients. These findings verified by the findings by Tabatabaei et al. [26] and Carneiro et al. [28] who discovered that the stage of cervical cancer influenced the survival time of cervical cancer patients. These findings shown that the robust network topology is reliable to use even though that the dataset used is contaminated with outliers but yet robust network topology able to identify the significant factor that influenced the survival time of cervical cancer patients.

Consequently, these analyses concluded that *STG* is the most important factor that influence the survival of cervical cancer patients which implied that early diagnosis is very important. Early diagnosis of cancer allows early intervention to try to slow or prevent cancer development and lethality, hence, the survival improves [27]. For the future research, it is suggested to use a robust network topology in these following situations.

(i) Use the dataset which not encountered the singularity problem since ISE algorithm involved the inverse of sample covariance matrix.
(ii) Propose a new robust network topology by substitute the classical estimator with robust estimator which is free from singularity problem.

Acknowledgement

This research was funded by a grant from the International Matching Grant, RDU222708/ UIC221521 of University Malaysia Pahang Al-Sultan Abdullah. The authors would like to thanks the late Dr. Nuradhiathy binti Abd Razak for the cervical cancer dataset.

[1] Clark T. G., Bradburn M. J., Love S. B., Altman D. G. Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer. **89**, 232–238 (2003).

[2] Bacha R. H., Jabir Y. N., Asebot A. G., Liga A. D. Risk Factors Affecting Survival Time of Breast Cancer Patients: The case of Southwest Ethiopia. Journal of Research in Health Sciences. **21** (4), e00532 (2021).

[3] Ebrahimi V., Khademian M. H., Masoumi S. J., Morvaridi M. R., Jahromi S. E. Factors influencing survival time of hemodialysis patients; time to event analysis using parametric models: a cohort study. BMC Nephrology. **20**, 215 (2019).

[4] Byeon K. H., Kim D. W., Kim J., Choi B. Y., Choi B., Cho K. D. Factors affecting the survival of early COVID-19 patients in South Korea: An observational study based on the Korean National Health Insurance big data. International Journal of Infectious Diseases. **105**, 588–594 (2021).

[5] Schober P., Vetter T. R. Survival Analysis and Interpretation of Time-to-Event Data: The Tortoise and the Hare. Anesthesia & Analgesia. **127** (3), 792–798 (2018).

[6] Govindarajulu U., D'Agostino R. B. Review of current advances in survival analysis and frailty models. WIREs Cimputational Statistics. **12** (6), e1504 (2020).

[7] Hazra A., Gogtay N. Biostatistics Series Module 9: Survival Analysis. Indian Journal of Dermatology. **62** (3), 251–257 (2017).

[8] Andrade C. Survival Analysis, Kaplan–Meier Curves, and Cox Regression: Basic Concepts. Indian Journal of Psychological Medicine. **45** (4), 434–435 (2023).

[9] Lee S. W. Kaplan–Meier and Cox proportional hazards regression in survival analysis: statistical standard and guideline of Life Cycle Committee. Life Cycle. **3**, e8 (2023).

[10] Wang G., Li X., Xiong R., Wu H., Xu M., Xie M. Long-term survival analysis of patients with non-small cell lung cancer complicated with type 2 diabetes mellitus. Thoracic Cancer. **11** (5), 1309–1318 (2020).

[11] Hazra A., Gogtay N. Biostatistics Series Module 10: Brief Overview of Multivariate Methods. Indian Journal of Dermatology. **62** (4), 358–366 (2017).

[12] Croux C., Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika. **87** (3), 603–618 (2000).

[13] Fitrianto A., Midi H. A Comparison Between Classical and Robust Method in a Factorial Design in the Presence of Outlier. Journal of Mathematics and Statistics. **9** (3), 193–197 (2013).

[14] Herwindiati D., Hendryli J., Mulyono S. Robust Kurtosis Projection Approach for Mangrove Classification. Recent Advances in Information and Communication Technology 2018. 93–103 (2019).

[15] Syed Abd Mutalib S. S., Satari S. Z., Wan Yusoff W. N. S. A New Robust Estimator to Detect Outliers for Multivariate Data. Journal of Physics: Conference Series. **1366**, 012104 (2019).

[16] Mohamad Mokhtar M. A. A., Yusoff N. S., Liang C. Z. Robust Hotelling's T2 statistic based on M-estimator. Journal of Physics: Conference Series. **1988**, 012116 (2021).

[17] Syed Abd Mutalib S. S., Satari S. Z., Wan Yusoff W. N. S. A Review on Outliers-Detection Methods for Multivariate Data. Journal of Statistical Modeling & Analytics (JOSMA). **3** (1), (2021).

[18] Li X., Deng S., Li L., Jiang Y. Outlier Detection Based on Robust Mahalanobis Distance and Its Application. Open Journal of Statistics. **9** (1), 15–26 (2019).

[19] Mantegna R., Stanley H. An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press (1999).

[20] Kruskal J. B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. Proceedings of the American Mathematical Society. **7**, 48–50 (1956).

[21] Ayegba P., Ayoola J., Asani E. O., Okeyinka A. A Comparative Study of Minimal Spanning Tree Algorithms. 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS). 1–4 (2020).

[22] Yusoff N. S., Mohamad N., Liang C. Z., Sharif S., Ken T. L. A Network Topology Approach to Diagnose the Shift of Covariance Structure. MATEC Web of Conferences. **189**, 03027 (2018).

[23] Lim H. A., Habshah M. Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. Computational Statistics. **31**, 859–877 (2016).

[24] Syed Abd Mutalib S. S., Satari S. Z., Wan Yusoff W. N. S. Comparison of robust estimators for detecting outliers in multivariate datasets. Journal of Physics: Conference Series. **1988**, 012095 (2021).

[25] Singh G. N., Bhattacharyya D., Bandyopadhyay A. Robust estimation strategy for handling outliers. Communications in Statistics – Theory and Methods. **53** (15), 5311–5330 (2024).

[26] Tabatabaei F. S., Saeedian A., Azimi A., Kolahdouzan K., Esmati E., Maddah A. Evaluation of Survival Rate and Associated Factors in Patients with Cervical Cancer: A Retrospective Cohort Study. Journal of Research in Health Sciences. **22** (2), e00552 (2022).

[27] Crosby D., Bhatia S., Brindle K. M., Coussens L. M., Dive C., Emberton M., Esener S., Fitzgerald R. C., Gambhir S. S., Kuhn P., Rebbeck T. R., Balasubramanian S. Early detection of cancer. Science. **375** (6586), eaay9040 (2022).

[28] Carneiro S. R., De Araújo Fagundes M., De Jesus Oliveira do Rosário Pricila, Neves L. M. T., Da Silva Souza G., Da Conceição Nascimento Pinheiro M. Five-year survival and associated factors in women treated for cervical cancer at a reference hospital in the Brazilian Amazon. PLoS One. **12** (11), e0187579 (2017).

Надійний підхід до топології мережі для аналізу виживання

Ван Юсофф В. Н. С.[1], Мухаммад Н.[1], Абд Муталіб С. С. С.[2], Закарія З. А.[3]

[1] *Центр математичних наук, Університет Малайзії Паханг Аль-Султан Абдулла, Лебухрайя Персіаран Тун Халіл Якоб, 26300 Куантан, Паханг, Малайзія*
[2] *Факультет комп'ютерних наук та математики, Університет Малайзії Теренгану, 21030 Куала-Нерус, Теренгану, Малайзія*
[3] *Факультет інформатики та обчислювальної техніки, Університет Султана Зайнала Абідіна, Куала-Теренгану, Малайзія*

Топологію мережі можна використовувати для спрощення складності наборів даних. Ми досліджуємо її функцію у проведенні аналізу виживання, щоб визначити найважливіші фактори, що сприяли часу виживання від постановки діагнозу до смерті. Цей метод може легко проілюструвати деякі типи складних взаємодій у наборі даних. Потім, на основі цих взаємодій, визначаються найважливіші фактори в аналізі виживання. Однак, топологія мережі, яка заснована на класичній оцінці, надзвичайно чутлива до вихідних спостережень, тобто висновки, які зроблені на основі забрудненої топології мережі, можуть бути оманливими. Тому в цій статті класична оцінка, тобто класична кореляційна матриця топології мережі, замінюється надійною оцінкою, тобто надійною кореляційною матрицею, яка розроблена на основі рівності індексних наборів (ISE). Потім інтерпретація цієї надійної топології мережі здійснюється за допомогою міри центральності, тобто ступеня центральності. Подано та обговорено дослідження часу виживання пацієнтів з раком шийки матки. Надійна топологія мережі показала, що найважливішим фактором, що впливає на виживання пацієнток з раком шийки матки, є стадія постановки діагнозу (STG). Вища стадія раку шийки матки призводить до коротшого часу виживання пацієнток з раком шийки матки. Тому рання діагностика дуже важлива. Рання діагностика раку дозволяє раннє втручання, щоб спробувати уповільнити або запобігти розвитку раку та летальності, отже, покращує виживання.

Ключові слова: *топологія мережі; надійна оцінка; аналіз виживання; рак шийки матки.*