UDC: 616.24-006:036.65+004.8 DOI: 10.52532/2663-4864-2022-3-65-4-11

USING MACHINE LEARNING ALGORITHMS TO DEVELOP A MODEL FOR PREDICTING THE SURVIVAL OF LUNG CANCER PATIENTS IN THE REPUBLIC OF KAZAKHSTAN

V.A. MAKAROV[1,2], D.R. KAIDAROVA[3], S.E. YESSENTAYEVA[4], J. KALMATAYEVA[2], M.E. MANSUROVA[2], N. KADYRBEK[2], R.E. KADYRBAYEVA[3], S.T. OLZHAYEV[1], I.I. NOVIKOV[1]

[1]«Almaty Regional Multidisciplinary Clinic» MSE on REM, Almaty, the Republic of Kazakhstan;
[2]«Al-Farabi Kazakh National University» Non-Commercial JSC, Almaty, the Republic of Kazakhstan;
[3]«Kazakh Institute of Oncology and Radiology» JSC, Almaty, the Republic of Kazakhstan;
[4]«Kazakh-Russian Medical University» Non-Governmental Educational Institution, Almaty, the Republic of Kazakhstan

ABSTRACT

Relevance: *The 5-year overall survival rate(s) in NSCLC p-stage IA is 73%, and the recurrence rate in radically treated patients is almost 10%.*

The study aimed to *evaluate the prognostic significance of several clinical and morphological factors and apply machine learning algorithms to predict the results of the overall survival of patients with lung cancer.*

Methods: *The forms 030-6/y C34 – lung cancer (n=19,379) from the EROB database for 2014-2018 were analyzed, and the impact of risk factors on overall survival was assessed using the Kaplan-Meier method. Accordingly, the training data set for constructing forecasting models included 19,379 observations and 15 factors. The machine learning algorithms such as Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, and K Nearest Neighbors (KNN) Classifier were implemented in the Python programming language. The results were evaluated by constructing an error matrix and calculating classification metrics: the proportion of correctly classified objects (accuracy) during training and validation (validation), accuracy (precision), completeness (recall), Kappa-Cohen.*

Results: *In our study, 19,379 patients were analyzed, including 15,494 men (79.95%) and 3,885 women (20.04%). At the time of the study, 6,171 men (39.8%) and 1,962 women (49.5%) were alive. Median survival was 8.3 months (SE – 0.154 months, 95% CI – 7.96-8.56) in men and 15.43 months (SE – 1.0 months, 95% CI – 13.497-17.363) in women. At diagnosis, 1,037 patients (5.35%) had stage I disease, and 4,145 (21.38%) had stage II. Most patients (61.4%) had advanced stage NSCLC: 9,189 people (47.4%) were diagnosed with stage III, and 4,655 (24%) – with stage IV. The reliability of differences in median survival ($\chi 2=3991.6$, $p=0.00$) indicated the prognostic significance of the tumor process stage and its influence on the patient's survival. Also, the revealed significant difference in the median survival of patients with various morphological forms of lung cancer suggests the prognostic significance of the morphological factor (the difference between those indicators was statistically significant, $\chi 2=623.4$ $p=0.000$).*

Conclusion: *Machine learning models can predict the risk of fatal outcomes for patients after surgical treatment and registration in the EROB database. The creation of patient-oriented systems to support medical decision-making makes it possible to choose the optimal strategies for adjuvant therapy, dispensary observation, and frequency of diagnostic studies.*

Keywords: *lung cancer, prognostic significance, machine learning, relapses, overall survival.*

Introduction: In recent decades, cancer of thoracic organs has become one of the main causes of cancer cases and deaths. Despite early detection, some patients still die from relapse. According to R. Maeda, relapses in radically treatment patients approach 10% [1]. Determination of risks of relapse and/or fatal outcomes in patients with NSCLC remains an acute open issue. Modern tumor staging system (TNM 7 and 8) is the most common tool for predicting the course of NSCLC. However, this classification does not reflect all significant clinical and pathological predictors, so it cannot always determine a personalized approach in precision medicine [2-4]. Some studies demonstrate AI-based models to be more accurate than the standard TNM staging system since they analyze a large amount of data, reflecting both the biological and clinical features of the course of the disease [5]. Therefore, the models based on machine learning were recommended as prognostic tools alternative or supplementary to TNM classification [6]. A literature review revealed a successful use of machine learning algorithms, such as *Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, and K Nearest Neighbors (KNN) Classifier,* in the classification of LC patients by risk groups [7-12] and predicting the survival of patients with LC [13-14].

The study aimed to evaluate the prognostic significance of several clinical and morphological factors and apply machine learning algorithms to predict the results of the overall survival of patients with lung cancer.

Materials and methods: The forms 030-6/y C34 – lung cancer (n=19,379) from the EROB database for

2014-2018 were analyzed. The impact of risk factors (gender, age, TNM, histology, localization of metastatic foci) on overall survival was assessed using the Kaplan-Meier method. The database was created using Microsoft Excel. Accordingly, the training data set for constructing forecasting models included 19,379 observations and 15 factors. We identified three risk groups: Group 1 – survival from 0 to 12 months, Group 2 – survival from 12 to 24 months, and Group 1 – survival from 24 to 72 months, respectively.

The machine learning algorithms *(Random Forest Classifier*, *Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier)* were implemented in the Python programming language. The results were evaluated by constructing an error matrix and calculating classification metrics: the proportion of correctly classified objects (accuracy) during training and validation (validation), the measurement accuracy (precision), and completeness (recall) by Kappa-Cohen.

Results:

Evaluation of the gender factor impact on the survival of patients with LC in the Republic of Kazakhstan.

Our study involved 19 379 patients, including 15 494 men (79.95%) and 3 885 women (20.04%).

At the time of the study, 6 171 (39,8%) men were alive, with a median survival of 8.3 months (SE – 0.154 months, 95% CI 7.96-8.56). One-year survival in men was 44% (SE – 0.44), two-year – 31% (SE – 4.4), three-year – 26% (SE – 0.47), four-year– 24% (SE – 0.49), and five-year survival reached 23% (SE – 0.51).

Among women, 1 962 (49,5%) were alive, with a median survival of 15.43 months (SE – 1.0 month, 95% CI 13.497-17.363). One-year survival in women was 55% (SE – 0.84), two-year – 45% (SE – 0.9), three-year – 40% (SE – 0.95), four-year– 38% (SE – 1.0), and five-year survival reached 37% (SE – 1.03) (Figure 1).

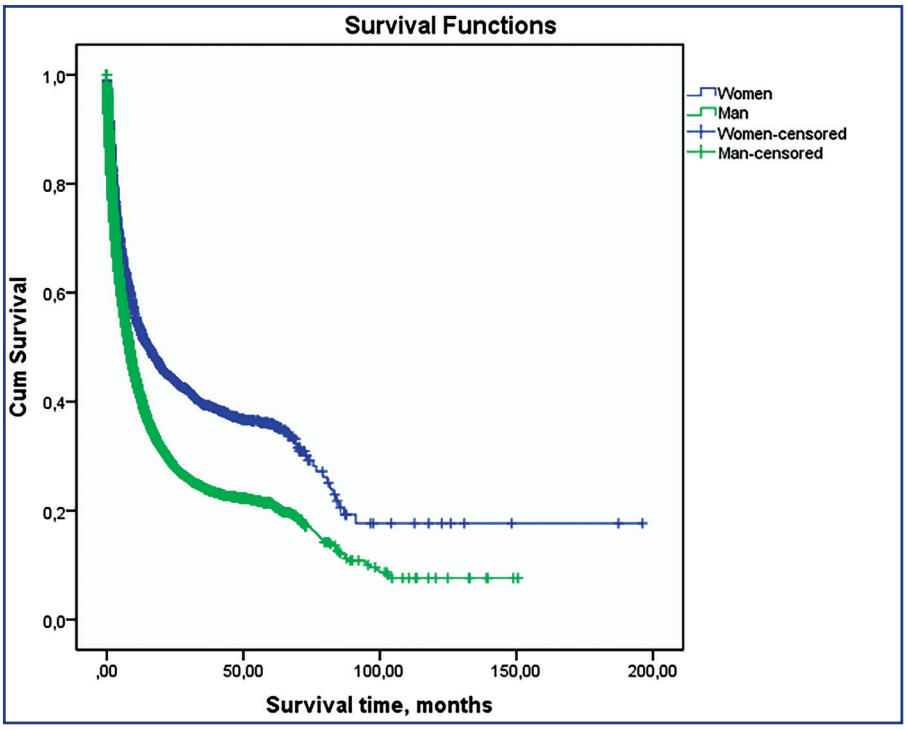

Figure 1 – Overall survival of patients depending on gender, by Kaplan-Meier method

Thus, it is clear that in our data set, the male gender was a risk factor for survival in LC. The difference in the median survival of men and women with LC was statistically significant: $\chi 2=219.03$, $p=0.00$.

Tumor stage impact on the remote outcome of patients with LC

Most patients with NSCLC (61.4%) had advanced cancer: stage III – 9 189 (47.4%) or stage IV – 4 655 (24%).

Among stage I patients, 845 (81.5%) were alive by the end of the study (2018). The median was not reached: the median survival was 125.6 months, SE – 9.6 months, 95% CI - 106.7-144.5. Among stage II patients, 2 366 (57.1%) were alive by the end of the study (2018). Their median survival was 26.1 months, SE – 1.4 months, 95% CI 23.3-28.8. Among 9 189 patients with stage III, 3 687 (40.1%) survived, with a median survival of 8.3 months, SE – 0.2 months, 95% CI – 8.0-8.7. By the end of 2018, only one-fourth of patients with stage IV had survived – 1 183 (25.4%). The median survival in that group was 3.3 months, SE – 0.1 month, 95% CI – 3.1-3.5 (Figure 2).

Assessing the significance of differences in median survival ($\chi 2=3991.6$, $p=0.00$) showed the prognostic significance and influence of the tumor process stage on patient survival.

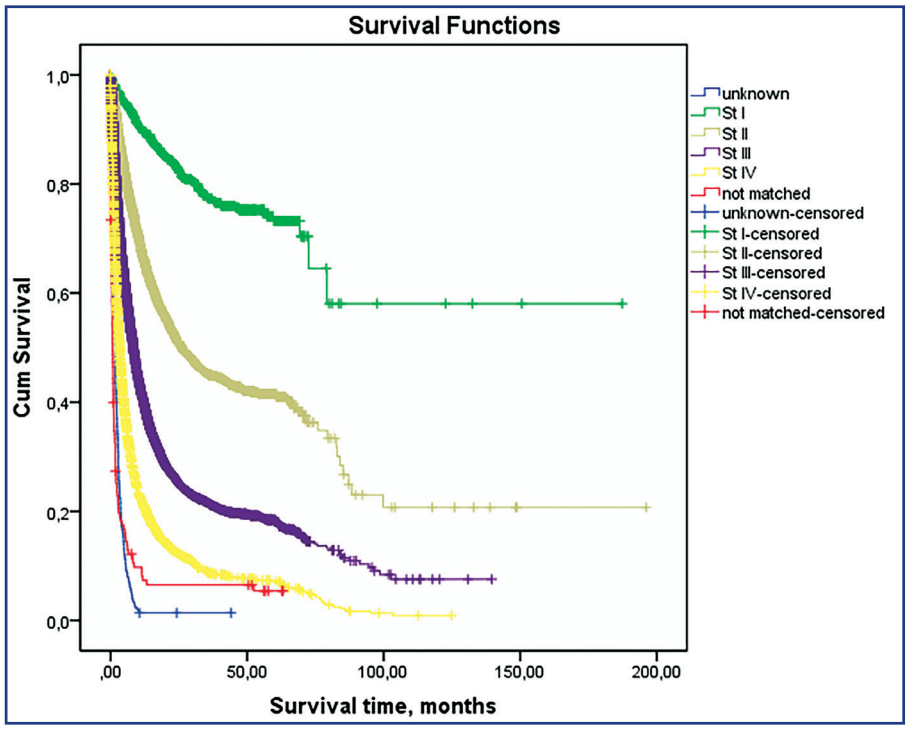

Figure 2 – Overall survival of patients depending on disease stage, by Kaplan-Meier method

Impact of the tumor morphological type on LC patient survival in the RK

Among 19 379 patients diagnosed with LC in 2014-2018, 18.5% (3 579) had adenocarcinoma. Of them, 1 738 (48.6%) were alive by the end of 2018; the median survival was 17.1 months, SE – 0.9 months, 95% CI – 15.2-19.1.

Patients with squamous cell cancer accounted for 27.0% (5 231), and 2 254 (43.1%) survived 2018. The median survival was 11.6 months, SE – 0.3 months, 95% CI – 10.9-12.3 (Figure 2).

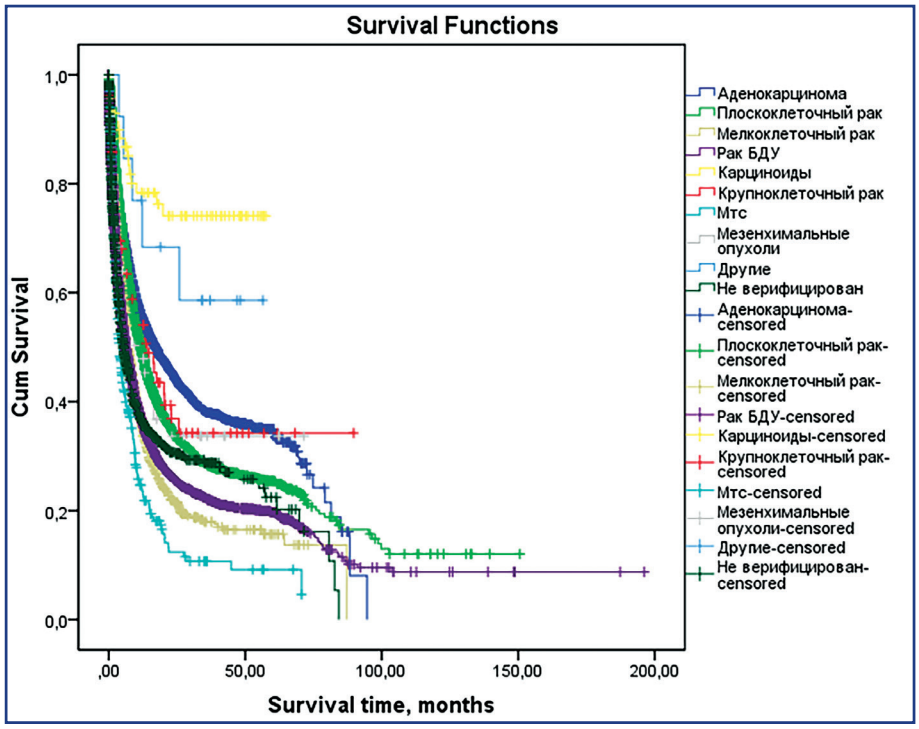

Figure 3 – Overall survival of patients depending on the tumor morphological type, by Kaplan-Meier method

Small-cell cancer (SCC) was diagnosed in 1 091 (5.6%) cases. Of them, 377 (34.6%) were alive by the end of 2018; the median survival was 7.2 months, SE – 0.3 months, 95% CI – 6.5-7.99. Lung cancer not otherwise specified (NOS) was detected in 7 643 (39.4%) cases; 2 922 (38.2%) were alive; the median survival was 6.2 months, SE – 0.2 months, 95% CI – 5.7-6.6.

Among patients with lung adenocarcinoma, one-, two-, three-, and four-year survival rates were 57% SE1, 45% SE1, 39% SE1, and 37% SE1, respectively. The five-year survival reached 36% SE1. In patients with squamous cell lung cancer, the one-, two-, three-, and four-year survival rates were slightly lower and amounted to 51% SE1, 35% SE1, 30% SE1, and 28% SE1, respectively. The five-year survival reached 27% SE1. In SCC, major survival rates were still below NSCLC: one-year survival was 39% SE2, two-year survival was 24% SE2, three-year survival was 21% SE2, and four-year survival was 19% SE2. The five-year survival did not exceed the 20% threshold and amounted to 18% SE2.

The survival rates in patients with lung cancer NOS correlated with those in SCC: 40% SE1, 28% SE1, 24% SE1 for one-, two-, and three-year survival, and 22% SE1 for four- and five-year survival. One-year survival in patients with carcinoids amounted to 78% SE5. The two-, three-, four-, and five-year survival rates were 74% SE6.

Thus, mandatory morphological identification of malignant neoplasms of the lung helps in choosing treatment tactics and selecting adequate anticancer drug therapy and the disease prognosis. The identified significant difference in median survival among patients with various morphological forms of lung cancer demonstrates the prognostic significance of the morphological factor (the difference between these indicators was statistically significant, $\chi 2=623.4$ $p=0.000$).

Predicting the marker of survival of patients with LC from the EROB database using a machine learning model.

After assessing potentially significant predictors from the EROB database, a training set was formed. Machine learning-based models can automatically classify patients, taking into account multifactorial data.

Figure 4 shows the prevailing number of patients in Risk Group 1.

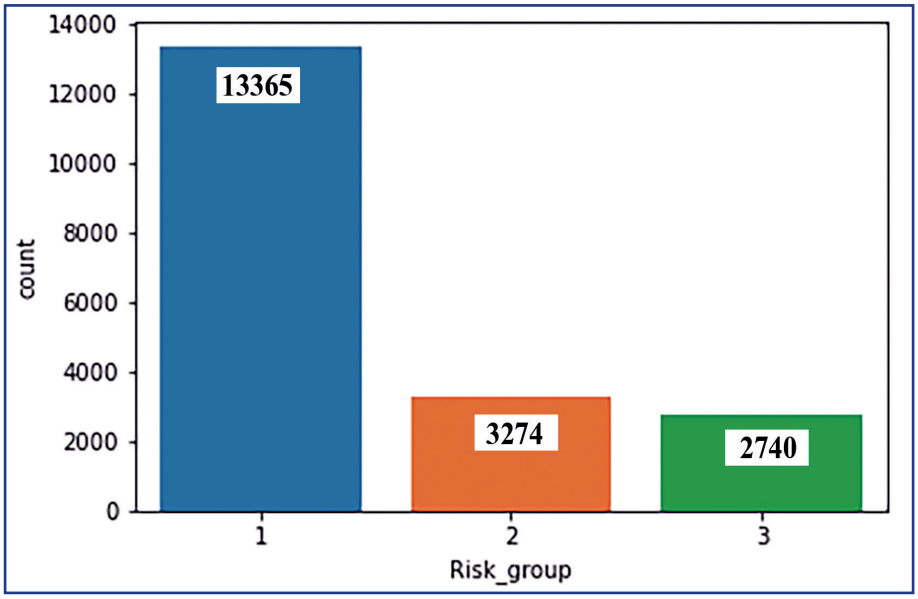

Figure 4 - Distribution of LC patients by risk groups

When building the machine learning model, we had to decide on significant and excessive (noise) parameters. Avoiding excessive parameters improves data interpretation and model accuracy. We optimized the list of selected parameters during model training to improve the simulation accuracy. The main predictors (~5%) chosen for those models were the tumor stage and size, the involvement of lymph nodes (N), and patient age (Table 1).

The developed machine learning models showed a high proportion of correctly grouped classification objects, i.e., high model *accuracy*. The highest accuracy of predictions on the training set was achieved using *Decision Tree* (0.86), *Gradient Boosting* (0.72), and *Random Forest* (0.70) algorithms. Validation of the obtained models revealed the following accuracy rates: for Gradient Boosting - 0.70, Random Forest - 0.70, and logistic regression - 0.69 (Figure 5, Table 2).

The Decision Tree algorithm showed the best characteristics (accuracy during training - 0.86, during validation – 0.63) on the given data set. After the optimal parameters for the model were selected, namely {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}, the accuracy during validation amounted to 69%. The quality of this model was tested using the

error matrix (Figure 6). During the test, the degree of measurement accuracy (precision) was 0.71, and the recall was 0.87. The Consistency Measure Indicator k-Cohen was 0.66, which indicates the good potential of this approach. Other measurements amounted to a true positive rate (TPR) of 0.98, a false positive rate (FPR) of 0.06, specificity was 0.94, and an area under the curve (AUC) was 0.98.

Table 1 - Calculation of the importance of parameters in algorithms,%

Parameter	Algorithm		
	Decision Tree Classifier	Random Forest Classifier	Gradient Boosting Classifier
Stage	16,0	34,7	59,3
Tumour_size	8,6	17,2	6,0
N	9,1	17,4	8,2
Metastasis	3,7	12,9	4,3
Brain_metastasis	0,7	0,3	0,3
Multiple_metastasis	2,1	0,9	0,2
Bone_metastasis	1,0	0,2	0,3
Liver_metastasis	1,7	0,2	0,4
Adenocarcinoma	2,7	4,3	4,0
Squamoscell_carcinoma	3,4	1,6	1,4
Smallcell_carcinoma	2,0	0,2	0,1
Carcinoid_tumours	0,4	0,4	0,8
Cancer_unknown	2,5	3,4	2,6
Gender	3,6	2,9	4,3
Age	42,4	3,4	7,8
Total	**100**	**100**	**100**

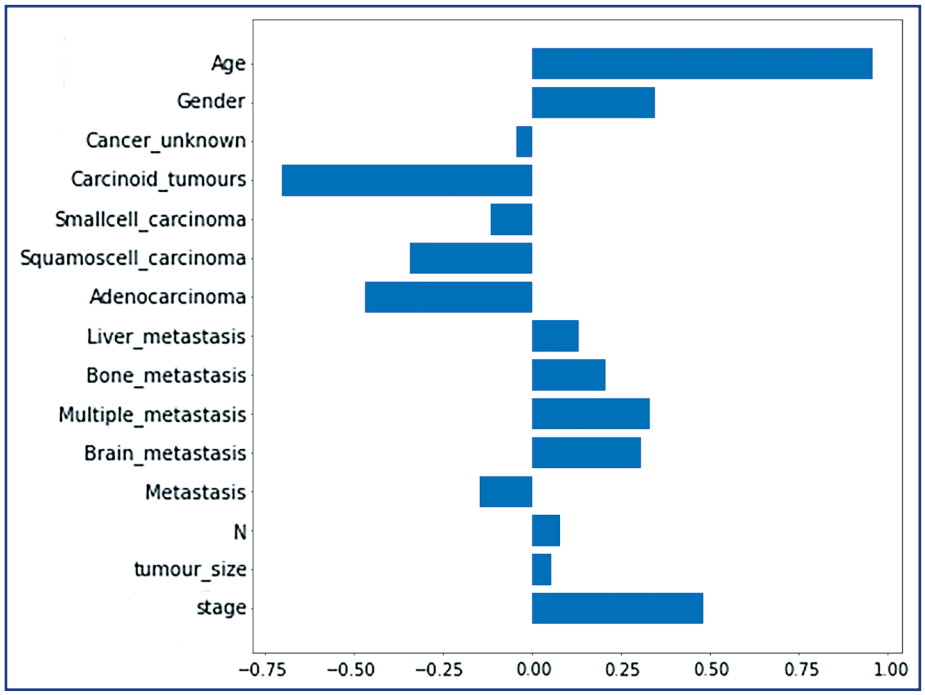

Figure 5 - Survival prediction model created using the logistic regression algorithm

Table 2 – Accuracy indicators of machine learning algorithms during training and validation

Machine learning algorithms	Accuracy during training	Accuracy during validation
DecisionTreeClassifier	0,86	0,63
RandomForestClassifier	0,71	0,70
GradientBoostingClassifier	0,72	0,70
LogisticRegressionModel	0,70	0,69
K NearestNeighborsClassifier	0,75	0,68

Predicting the marker of survival of patients registered at the EROB database during 2014-2018 (19 379 patients, 15 factors) using machine learning has shown that such machine learning algorithms as Random Forest, Gradient Boosting, Decision Tree, and logistic regression produce the best models, with the *accuracy of* 72% during training and 70% during validation on the test set. The *accuracy* of the Decision Tree Classifier during training was 87%

compared to 63% on the test set. After the optimal parameters for the model were selected {'C': 100, 'penalty': 'l2', 'solver': 'liblinear'}, the *validation accuracy* amounted to 69%.

The constructed models have reached the indicators acceptable for the application and therefore were recognized as applicable.

According to the error matrix, the *measurement precision* was 0.71, the *recall* – was 0.87, the k-Cohen was 0.66, TPR – 0.98, FPR – 0.06, and the specificity was 0.94, and AUC – 0.98.

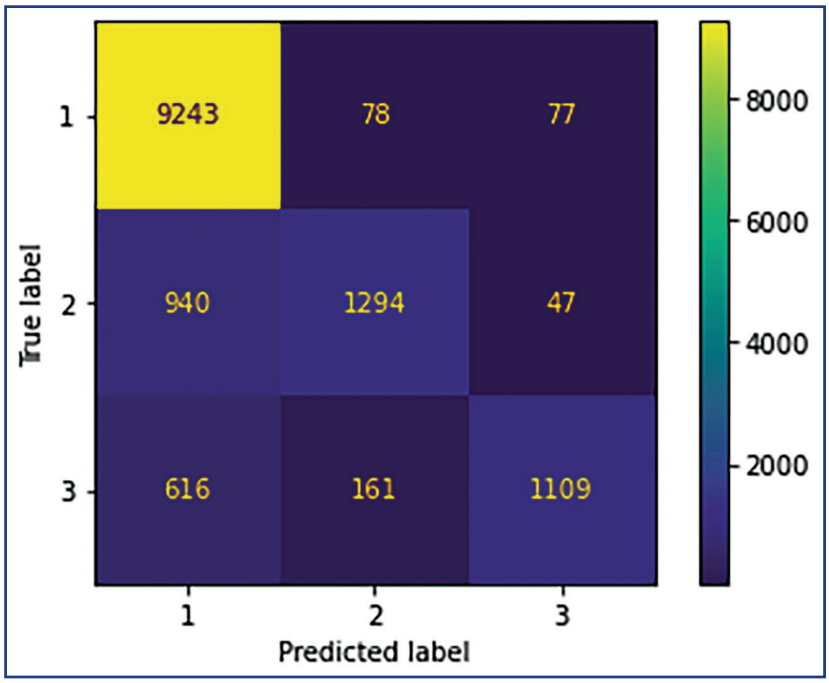

Figure 6 – Error matrix of the prediction model for three risk groups of patients from the EROB database created using the decision tree algorithm

Discussion: The TNM staging system remains an important and verified survival predictor, but it does not account for all disease parameters. This limits the TNM staging system's capacity to predict an unfavorable outcome. In this study, we used the survival impact assessment by Kaplan-Meier to evaluate the prognostic and predictive biomarkers that affect disease prognosis. E.g., the male gender was identified as a risk factor worth focusing on in future studies. Presumably, not the male gender as a genetic factor but the relationship of gender with lifestyle, behavioral factors, work, attitude to health, etc., should be explored.

The developed algorithms can also help identify the patient subgroups requiring more intensive follow-up and adjuvant treatment regimens. Stratifying risks based on the received data may contribute to changing established monitoring and treatment standards in favor of further drug therapy and the intensity of dispensary follow-up. Thus, high-risk patients shall reduce the dispensary follow-up to timely adjust treatment to changes in their oncological and functional status. Still, identifying patient groups with a high risk of relapse who can benefit from adjuvant therapy remains an issue. Selecting patients for chemotherapy based on a single risk factor may be ineffective because comprehensive prediction shall account for all disease attributes and the weight of each factor.

In this study, machine learning models showed the optimal combination between forecast and actual observation. This guarantees the reproducibility and reliability of the proposed model. More importantly, the proposed model fits the EROB cohort.

Machine learning algorithms deliver a more accurate prognosis than the TNM staging system and earlier developed predictive models. This tool allows doctors to better predict the patient's survival after surgery and determine the subgroups of patients who need a specific treatment strategy.

Takeaways:

1. The analysis of factors influencing LC survival revealed the male gender as a risk factor ($\chi2=219.03$, $p=0.00$), while the female gender was classified as a favorable prognostic factor.

2. An analysis of clinical and morphological factors showed a significant effect of such indicators as the stage of the disease ($\chi2=3991.6$, $p=0.00$) and the morphological type of tumor ($\chi2=623.4$ $p=0.000$) on survival in LC.

3. Significance of differences in median survival ($\chi2=3991.6$, $p=0.00$) indicates the prognostic significance and impact of the LC stage on survival.

4. The Random Forest, Gradient Boosting, and Decision Tree classifiers showed their applicability in predicting the risk group (marks) of overall survival in LC.

5. The machine learning model validation on the test set showed the admissibility of the Random Forest, Gradient Boosting, and Decision Tree classifiers to aid decision-making.

6. Data quality, as is an algorithm, is important for building a predictive model. Data shall be accurate and bulk, with a normal (Gaussian) distribution by risk groups (classes).

Conclusion: Machine learning models can help predict the risk of death in LC patients after surgical treatment or registration in the EROB database. Creating patient-oriented medical decision support systems will help choose the optimal strategies for adjuvant therapy, dispensary follow-up, and the frequency of diagnostic tests.

References:

1. Maeda R., Yoshida J., Ishii G., Aokage K., Hishida T., Nishimura M., Nishiwaki Y., Nagai K. Long-term outcome and late recurrence in patients with completely resected stage IA non-small cell lung cancer // J. Thorac. Oncol. – 2010. – Vol. 5. – P. 1246-1250. https://doi.org/10.1097/JTO.0b013e3181e2f247;

2. Amin M.B., Greene F., Edge S.B., Compton C.C., Gershenwald J.E., Brookland R.K., Meyer L., Gress D.M., Byrd D.R., Winchester D.P. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging // CA: Cancer J. Clin. – 2017. – Vol. 67(2). – P. 93-99. https://doi.org/10.3322/caac.21388;

3. Goldstraw P., Chansky K., Crowley J., Rami-Porta R., Asamura J., Eberhardt W.E.E., Nicholson A.G., Groome P., Mitchell A., Bolehack V., on behalf of the International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC lung cancer-staging project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer // J. Thorac. Oncol. – 2016. – Vol. 11(1). – P. 39-51. https://doi.org/10.1016/j.jtho.2015.09.009;

4. Rami-Porta R., Bolejack V., Crowley J., Ball D., Kim J., Lyons G., Rice T., Suzuki K., Thomas C.F. Jr., Travis W.D., Wu Y.-L., on behalf of the IASLC Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC Lung Cancer Staging Project: Proposals for the Revisions of the T Descriptors in the Forthcoming Eighth Edition of the TNM Classification for Lung Cancer // J. Thorac. Oncol. – 2015. – Vol. 10(7). – P. 990-1003. https://doi.org/10.1097/JTO.0000000000000559;

5. Balachandran V.P., Gonen M., Smith J.J., DeMatteo R.P. Nomograms in oncology: more than meets the eye // Lancet Oncol. – 2015. – Vol. 16(4). – P. e173-180. https://doi.org/10.1016/S1470-2045(14)71116-7;

6. Kourou K., Exarchos K.P., Papaloukas C., Sakaloglou P., Exarchos T., Fotiadis D.I. Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis // Comput. Struct. Biotechnol. J. – 2021. – Vol. 19. – P. 5546-5555. https://doi.org/10.1016/j.csbj.2021.10.006;

7. Lynch C.M., Abdollahi B., Fuqua J.D., de Carlo A.R., Bartholomai J.A., Balgemann R.N., van Berkel V.H., Frieboes H.B. Prediction of lung cancer patient survival via supervised machine learning classification techniques // Int. J. Med. Inform. – 2017. – Vol. 108. – P. 1-8. https://doi.org/10.1016/j.ijmedinf.2017.09.013;

8. Ramroach S., Joshi A., John M. Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers // Mol. Omics. – 2020. – Vol. 16(2). – P. 113-125. https://doi.org/10.1039/c9mo00198k;

9. Levitsky A., Pernemalm M., Bernhardson B.M., Forshed J., Kölbeck K., Olin M., Henriksson R., Lehtiö J., Tishelman C., Eriksson L.E. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model // Sci. Rep. – 2019. – Vol. 9(1). – Art. ID 16504. https://doi.org/10.1038/s41598-019-52915-x;

10. Zhang X., Wang J., Li J., Chen W., Liu C. CRlncRC: a machine learning-based method for cancer-related long noncoding RNA identification using integrated features // BMC Med. Genomics. – 2018. – Vol. 11(Suppl 6). – Art. ID 120. https://doi.org/10.1186/s12920-018-0436-9;

11. Gu Q., Feng Z., Liang Q., Li M., Deng J.., Ma M, Wang W.., Liu J., Liu P, Rong P. Machine learning-based radiomics strategy for prediction of cell proliferation in non-small cell lung cancer // Eur. J. Radiol. – 2019. – Vol. 118. – P. 32-37. https://doi.org/10.1016/j.ejrad.2019.06.025;

12. Bergquist S.L., Brooks G.A., Keating N.L., Landrum M.B., Rose S. Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data // Proc. Mach. Learn. Res. – 2017. – Vol. 68. – P. 25-38. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6287925/;

13. Sena G.R., Lima T.P.F., Mello M.J.G., Thuler L.C.S., Lima J.T.O. Developing Machine Learning Algorithms for the Prediction of Early Death in Elderly Cancer Patients: Usability Study // JMIR Cancer. – 2019. – Vol. 5(2). – Art. ID e12163. https://doi.org/10.2196/12163;

14. Siah K.W., Khozin S., Wong C.H., Lo A.W. Machine-Learning and Stochastic Tumor Growth Models for Predicting Outcomes in Patients With Advanced Non-Small-Cell Lung Cancer // JCO Clin Cancer Inform. – 2019. – Vol. 3. – P. 1-11. https://doi.org/10.1200/CCI.19.00046.

ТҰЖЫРЫМ

ҚАЗАҚСТАН РЕСПУБЛИКАСЫНДАҒЫ ӨКПЕНІҢ ҚАТЕРЛІ ІСІГІМЕН АУЫРАТЫН НАУҚАСТАРДЫҢ ӨМІР СҮРУ КӨРСЕТКІШІ НӘТИЖЕЛЕРІН БОЛЖАМДАУ МОДЕЛІН ҚҰРАСТЫРУДАҒЫ МАШИНАМЕН ОҚЫТУДЫҢ РӨЛІ

В.А. Макаров[1,2], Д.Р. Кайдарова[3], С.Е. Есентаева[4], Ж. Калматаева[2], М.Е. Мансурова[2], Н. Кадырбек[2], Р.Е. Кадырбаева[3], С.Т. Олжаев[1], И.И. Новиков[1]

[1]«Алматы Жергілікті Көпсалалы Клиникасы» КМК ШЖҚ, Алматы, Қазақстан Республикасы;
[2]«Әл-Фараби атындағы Қазақ Ұлттық университеті» Коммерциялық емес АҚ, Алматы, Қазақстан Республикасы;
[3]«Қазақ онкология және радиология ғылыми-зерттеу институты», Алматы, Қазақстан Республикасы;
[4]«Қазақ-Ресей Медицна Университеті» Мемлекеттік емес білім беру мекемесі, Алматы, Қазақстан Республикасы

Өзектілігі: *Іа сатысындағы өкпенің қатерлі ісігінің 5 жылдық жалпы өмір сүру деңгейі 73% құрайды, ал радикалды емделген пациенттерде рецидив жиілігі шамамен 10% құрайды.*

Зерттеу мақсаты *– бірқатар клиникалық және морфологиялық факторлардың болжамды маңыздылығы мен өкпе қатерлі ісігі бар науқастардың жалпы өмір сүру нәтижелерін болжау үшін машиналық оқыту алгоритмдерін қолдану мүмкіншілігін бағалау.*

Әдістер: *030-6/у с34 – өкпе обыры (n=19379) нысандарына 2014-2018 жж. ОНЭР деректер базасынан талдау жүргізілді, Каплан-Мейер әдісі бойынша жалпы өмір сүруге қауіп факторларының әсерін бағалау жүргізілді. Тиісінше, болжау модельдерін құруға арналған оқыту жиынтығы 19379 бақылау мен 15 факторды қамтиды. Жұмыста қолданылатын Машиналық оқыту алгоритмдері (Random*

Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier) Python бағдарламалау тілінде іске асырылған. Нәтижелер қате матрицаны құру, жіктеу өлшемдерін есептеу арқылы бағаланды: оқыту және тексеру (validation), дәлдік (precision), толықтық (recall), Каппа-Коэн кезінде дұрыс жіктелген объектілердің үлесі (accuracy).

***Нәтижелері:** Біздің зерттеуімізде 19 379 науқас талданды, оның ішінде 15 494 ер адам (79,95%) және 3 885 әйел (20,04%). Зерттеу барысында қазіргі күні ерлер арасында 6 171 науқас (39,8%) тірі екендігі анықталды, бұл ретте өмір сүру медианасы 8,3 айды құрады (SE – 0,154 ай, 95% ДИ – 7,96-8,56). Әйелдер арасында 1 962 науқас (49,5%) тірі, бұл ретте өмір сүру медианасы 15,43 айды құрады (SE – 1,0 ай, 95% ДИ – 13,497-17,363). 1 037 пациентте (5,35%) аурудың I сатысында және 4 145 (21,38%) II сатысында анықталды. ӨҰЖЕҚІ науқастардың көпшілігінде (61,4%) кең таралған сатыда диагноз қойылған: 9 189 адамда (47,4%) – III сатыда, 4 655-те (24%) – IV сатыда. Өмір сүру медианасындағы айырмашылықтардың дұрыстығын бағалау ($\chi2=3991,6$, $p=0,00$) ісік процесінің болжамды маңыздылығын және науқастардың өмір сүруіне әсерін көрсетеді. Сондай-ақ, өкпе қатерлі ісігінің әртүрлі морфологиялық формалары бар науқастар арасында өмір сүру медианасындағы айтарлықтай айырмашылық морфологиялық факторлардың болжамды маңыздылығы туралы айтуға мүмкіндік береді (статистикалық тұрғыдан алғанда, бұл көрсеткіштер арасындағы айырмашылық сенімді болды, $\chi2=623,4$ $p=0,000$).*

***Қорытынды:** Машиналық оқыту модельдері хирургиялық емдеуден кейін де, ОНЭР дерекқорына тіркелгеннен кейін де науқастардың өлім қаупін болжауға мүмкіндік береді. Науқасқа бағдарланған медициналық шешімдер қабылдауды қолдау жүйесін құру адъювантты терапияның, диспансерлік бақылаудың және диагностикалық зерттеулер жиілігінің оңтайлы стратегияларын таңдауға мүмкіндік береді.*

***Түйінді сөздер:** өкпенің қатерлі ісігі, болжамды маңыздылығы, машиналық оқыту, қайталанулар, жалпы өмір сүру көрсеткіші.*

АННОТАЦИЯ

РОЛЬ МАШИННОГО ОБУЧЕНИЯ В РАЗРАБОТКЕ МОДЕЛИ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ВЫЖИВАЕМОСТИ БОЛЬНЫХ РАКОМ ЛЕГКИХ В РК

***В.А Макаров*[1,2], *Д.Р. Кайдарова*[3], *С.Е. Есентаева*[4], *Ж. Калматаева*[2], *М.Е. Мансурова*[2], *Н. Кадырбек*[2], *Р.Е. Кадырбаева*[3], *С.Т. Олжаев*[1], *И.И. Новиков*[1]**

[1]КГП на ПХВ «Алматинская Региональная Многопрофильная Клиника», Алматы, Республика Казахстан;
[2]НАО «Казахский Национальный Университет им. аль-Фараби», Алматы, Республика Казахстан;
[3]АО «Казахский Научно-Исследовательский Институт Онкологии и Радиологии», Алматы, Республика Казахстан;
[4]НУО «Казахско-Российский Медицинский Университет», Алматы, Республика Казахстан

***Актуальность:** В ряде исследований было показано, что модели, созданные с помощью искусственного интеллекта, являются более точными, чем обычная система стадирования TNM, поскольку они строятся на анализе большого объема данных, отражающих как биологические, так и клинические особенности течения болезни. На этом основании модели, созданные с помощью машинного обучения, были рекомендованы в качестве альтернативных или дополняющих TNM классификацию прогностических инструментов.*

***Цель исследования** – оценить прогностическую значимость ряда клинико-морфологических факторов и применить алгоритмы машинного обучения для прогнозирования результатов общей выживаемости больных с раком легких.*

***Методы:** Проведен анализ истории болезни пациентов с раком легкого (n=19379) из базы данных ЭРОБ за 2014-2018 гг., произведена оценка влияния факторов риска на общую выживаемость по методу Каплана-Мейера. Примененные в работе алгоритмы машинного обучения (Random Forest Classifier, Gradient Boosting Classifier, Logistic Regression Model, Decision Tree Classifier, K Nearest Neighbors (KNN) Classifier) реализованы на языке программирования Python.*

***Результаты:** В нашем исследовании были проанализированы истории болезни 19 379 пациентов. На момент исследования среди мужчин были живы 6 171 больных (39,8%), при этом медиана выживаемости составила 8,3 месяцев (SE – 0,154 месяцев, 95% ДИ – 7,96-8,56). Среди женщин были живы 1 962 больных (49,5%), при этом медиана выживаемости составила 15,43 месяцев (SE – 1,0 месяц, 95% ДИ – 13,497-17,363). У большинства (61,4%) пациентов НМРЛ был диагностирован в распространенной стадии: у 9 189 человек (47,4%) – на III стадии, у 4 655 (24%) – на IV стадии. Оценка достоверности различий в медиане выживаемости ($\chi2=3991,6$, $p=0,00$) указывает на прогностическую значимость и влияние стадии опухолевого процесса на выживаемость больных.*

***Заключение:** Модели машинного обучения позволяют прогнозировать риск развития летального исхода больных как после хирургического лечения, так и после постановки на учет в базу данных ЭРОБ. Создание пациент-ориентированных систем поддержки принятия врачебных решений позволяет выбрать оптимальные стратегии адъювантной терапии, диспансерного наблюдения и частоты диагностических исследований.*

***Ключевые слова:** рак легкого, прогностическая значимость, машинное обучение, рецидивы, общая выживаемость.*

***Transparency of the study:** Authors take full responsibility for the content of this manuscript.*
***Conflict of interests:** Authors declare no conflict of interest.*
***Financing:** This study was financed under the NTP BR11065390 (TF of the Ministry of Health of the Republic of Kazakhstan).*
***Вклад авторов:** Authors' input: contribution to the study concept – Makarov V.A., Kaidarova D.R., Esentaeva S.E., Olzhaev S.T., Novikov I.I.; study design – Makarov V.A., Esentaeva S.E., Kalmataeva Zh.A.; execution of the study – Makarov V.A., Mansurova M.E., Kadyrbek N., Kadyrbaeva R.E.; interpretation of the study – Makarov V.A., Mansurova M.E., Kadyrbek N.; preparation of the manuscript – Makarov V.A., Mansurova M.E., Kadyrbek N., Kadyrbaeva R.E.*

Authors' data:
Makarov Valery Anatolyevich – Head of surgical department of Almaty regional multidisciplinary clinic, Almaty, the Republic of Kazakhstan, tel: +77017750830; e-mail: makaroff_valeriy@mail.ru, ID ORCID: https://orcid.org/0000-0003-2120-5323;
Kaydarova Dilyara Radikovna – MD, Professor, Academician of the national academy of sciences of the Republic of Kazakhstan, Board chairman of KazNIIOiR, JSC, President of the Association of oncologists and radiologists of the Republic of Kazakhstan, Almaty, the Republic of Kazakhstan, tel: +77017116593; e-mail: kazior@onco.kz, ID ORCID: https://orcid.org/0000-0003-2120-5104;
Yesentaeva Suriya Ertugyrovna – MD, Professor, Head of the oncology department, Kazakh-Russian medical university, Almaty, the Republic of Kazakhstan, tel: +77077942910, e-mail: surya_esentay@mail.ru, ID ORCID: https://orcid.org/0000-0001-7087-1440;
Zhanna Kalmataeva – MD, Professor, Dean of the medicine and public health faculty, Al-Farabi Kazakh national university, Almaty, the Republic of Kazakhstan, tel: +77772187666, e-mail: Zhanna.Kalmatayeva@kaznu.kz, ID ORCID: https://orcid.org/0000-0002-5562-1969;
Madina Mansurova – Teacher of Al-Farabi Kazakh national university, Almaty, the Republic of Kazakhstan, tel: + 77014151960, e-mail: madina.mansurova@kaznu.kz, ID ORCID: https://orcid.org/0000-0002-9680-2758;
Kadyrbek Nurgali – Teacher of Al-Farabi Kazakh national university, Almaty, the Republic of Kazakhstan, tel: +77079890989, e-mail: mailto:mnurgaliqadyrbek@gmail.com, ID ORCID: https://orcid.org/0000-0002-5461-8899;
Kadyrbaeva Rabiga Esengalikyzy (corresponding author) – Chemotherapist of Kazakh research Institute of oncology and radiology, JSC, Almaty, the Republic of Kazakhstan, tel: +77074023344; e-mail: rabiga-92@mail.ru, ID ORCID: https://orcid.org/0000-0001-8254-8675;
Olzhaev Sayakhat Turarbekovich – Director of Almaty regional multidisciplinary clinic, Almaty, the Republic of Kazakhstan, tel: +77017749999, e-mail: S.Olzhayev20@gmail.com, ID ORCID: https://orcid.org/0000-0002-3312-323X;
Novikov Igor Igorevich – Deputy director for the medical part of Almaty regional multidisciplinary clinic, Almaty, the Republic of Kazakhstan, tel: +77773640684, e-mail: migor-novikov-1982@mail.ru, ID ORCID: https://orcid.org/0000-0001-7015-6770.