

Information about the authors

Moldir Seilkhanova* – Master's degree, doctoral student of the Department of Mathematical and Computer Modeling, L.N. Gumilyov Eurasian National University, Kazakhstan; e-mail: mbt_kz@mail.ru. ORCID: <https://orcid.org/0009-0003-5282-700X>.

Keylan Alimhan – PhD, Professor of the Department of Mathematical and Computer Modeling, L.N. Gumilyov Eurasian National University, Kazakhstan; e-mail: keylan@live.jp. ORCID: <https://orcid.org/0000-0003-0766-2229>.

Редакцияға енуі 30.03.2026

Өңдеуден кейін түсуі 14.05.2026

Жариялауға қабылданды 15.05.2026

[https://doi.org/10.53360/2788-7995-2026-2\(22\)-9](https://doi.org/10.53360/2788-7995-2026-2(22)-9)



FTAXP: 28.23.27

**Д.Ж. Шырынханова¹, И.Б. Карымсакова², Г. Жомартқызы¹, Д.О. Кожаметова²,
Т.А. Устинова²**

¹Д. Серікбаев атындағы Шығыс Қазақстан техникалық университеті,
070004, Қазақстан Республикасы, Өскемен қ., Серікбаев көшесі, 19

²Шәкәрім университеті,
071412, Қазақстан Республикасы, Семей қ., Глинка көшесі, 20 А

*e-mail: karymsakova.indira@mail.ru

МАШИНАЛЫҚ ОҚЫТУ (ML) МОДЕЛЬДЕРІН ҚОЛДАНА ОТЫРЫП БАУЫР ҚАТЕРЛІ ІСІГІН БОЛЖАУ

Аңдатпа: Ашық қолжетімді құрылымдалған клиникалық деректер негізінде бауыр қатерлі ісігі қаупін болжауға арналған машиналық оқыту модельдері салыстырмалы түрде бағаланды. Зерттеуде Kaggle платформасынан алынған деректер жиынтығы қолданылды ($N=5000$): мақсатты айнымалы Liver_cancer, 13 белгі (9 сандық, 4 категориялық), кластар теңгерімсіз (0: 78,22%, 1: 21,78%). Деректерді алдын ала өңдеу, модельдерді үйрету және бағалау кезеңдері бірыңғай эксперименттік протокол шеңберінде орындалды. Сандық айнымалылар стандарттау арқылы нормаланды; категориялық айнымалылар бинарлы форматта кодталды.

Салыстырмалы талдауға логистикалық регрессия (Logistic Regression), тірек векторлар әдісі (Support Vector Machine, RBF ядросымен), k ең жақын көршілер (k -Nearest Neighbors), кездейсоқ орман (Random Forest) және градиенттік бустинг (Gradient Boosting) модельдері енгізілді. Барлық модельдер дәлдік (accuracy), нақтылық (precision), сезімталдық (recall/sensitivity), ерекшелік (specificity), $F1$ -өлшемі ($F1$ -score) және ROC қисығы астындағы аудан (ROC-AUC) көрсеткіштері бойынша бағаланды.

Нәтижелер бойынша градиенттік бустинг (Gradient Boosting) ең жоғары жиынтық тиімділікті көрсетті: ROC-AUC=0.999498, Accuracy=0.9754, Precision=0.998967, Sensitivity=0.887971, Specificity=0.999744, $F1$ =0.940204. Салыстыру деректерінде SVM (RBF) және Random Forest тұрақты дискриминация бергенімен, сезімталдық пен $F1$ бойынша Gradient Boosting-мен төмен болды. Жұмыстың нәтижелері құрылымдалған клиникалық белгілер негізінде модель таңдауда метрикалардың кешенді талдауы қажет екенін көрсетеді; шектеулер ретінде деректердің теңгерімсіздігі және сыртқы валидацияның болмауы атап өтіледі.

Түйін сөздер: бауыр қатерлі ісігі, құрылымдалған клиникалық деректер, машиналық оқыту, градиенттік бустинг, салыстырмалы талдау, ROC-AUC, сезімталдық, ерекшелік.

Кіріспе

Бауыр қатерлі ісігі (бауыр ісігінің ең жиі түрі – гепатоцеллюлярлық карцинома, HCC) әлемдік деңгейде өзектілігі жоғары аурулардың қатарына жатады. Халықаралық онкологиялық деректерге сүйенсек, 2022 жылы дүние жүзінде бауыр қатерлі ісігінің шамамен 866 мыңға жуық жаңа жағдайы тіркеліп, оған байланысты 758 мыңнан астам өлім тіркелген [1]. Бауыр ісіктерінің

ішінде НСС үлесі басым: көптеген шолуларда оның барлық жағдайлардың шамамен 75-85% аралығын құрайтыны көрсетіледі [2]. Аурудың клиникалық ерекшелігі – оның жиі түрде созылмалы бауыр аурулары (вирустық гепатиттер, фиброз, цирроз және т.б.) аясында дамуы және ерте кезеңде клиникалық белгілердің айқын болмауы; бұл уақытылы анықтауды қиындатады. Осы себепті қауіп тобындағы пациенттерді жүйелі бақылау және тәуекел деңгейіне қарай жіктеу (risk stratification) ерте диагностиканың негізгі бағыттарының бірі ретінде қарастырылады [3].

Клиникалық тәжірибеде тәуекелді бағалау пациенттің жасы, жынысы, өмір салты, зертханалық және клиникалық көрсеткіштері сияқты құрылымдалған деректерге сүйенеді. Мұндай ақпарат медициналық мекемелерде күнделікті жиналатындықтан, оны талдай отырып бауыр қатерлі ісігінің ықтималдығын ертерек болжау мүмкін болады. Дәстүрлі клиникалық шкалалар – PAGE-B, aMAP – осы мақсатта жасалған, бірақ олар шектеулі санды параметрге сүйеніп, деректердегі күрделі өзара байланыстарды толық қамти алмауы мүмкін. Осы олқылықты жою үшін машиналық оқыту (ML) алгоритмдерін қолдану соңғы жылдары белсенді зерттелуде [4].

Халықаралық зерттеулер ML модельдерінің дәстүрлі тәсілдерден айтарлықтай озық нәтиже бере алатынын дәлелдеді. Осы бағыттағы негізгі жарияланымдарды жүйелі түрде салыстырсақ, модельдердің өнімділігі қолданылған деректер жиынтығына, популяция ерекшелігіне және валидация тәсіліне байланысты ауытқитыны байқалады. Дегенмен жалпы үрдіс анық: ML негізіндегі тәсілдер жеке биомаркерлер мен қарапайым клиникалық шкалаларға қарағанда жоғары дискриминациялық қабілет көрсетеді. Төменде осы бағыттағы негізгі зерттеулердің сипаттамалары жинақталған.

Кестедегі (кесте 1) зерттеулерді салыстырғанда, диагностикалық модельдер жоғары AUROC мәнін көрсетсе де (0.93-0.99), олардың көпшілігі тек бір ауруханның немесе бір дерекқордың деректерінде ғана тексерілген. Ал басқа елдердің пациенттерінде де тексерілген модельдер (Wong et al., Kim et al.) нәтижесі сәл төмен болғанымен, нақты тәжірибеде қолдануға жарамды екенін дәлелдеді. Сонымен қатар модельге кіретін белгілерді дұрыс таңдау болжам дәлдігін айтарлықтай арттыратыны байқалды. Осыны ескере отырып, біздің зерттеуімізде бес алгоритм бірдей шарттарда салыстырылып, нәтижелер бірнеше өлшем арқылы бағаланды.

Кесте 1 – Бауыр қатерлі ісігін анықтауда жоғары нәтиже көрсеткен МО модельдері

Автор	Модель	AUROC	Сезім-талдық	Популяция
Wong G. et al., 2022 [5]	Ridge Regression, Random Forest	0.844 / 0.837	90.0%	124 006 пациент, Гонконг, В/С гепатит
Kim H.Y. et al., 2021 [6]	Gradient Boosting Machine (GBM)	0.79-0.81	–	13 508 пациент, Корея және Батыс орталықтары, В гепатит
Sato M. et al., 2019 [7]	Gradient Boosting	0.940	93.3%	1 582 пациент, Токио, В/С гепатит
Hashem S. et al., 2020 [8]	Alternating Decision Tree (ADTree)	0.955-0.99	86.3-91.8%	4 423 пациент, Египет, С гепатит
Mostafa G. et al., 2024 [9]	SVM, Naive Bayes, Neural Network	0.932-0.993	89.2%	377 пациент, TCGA дерекқоры
Zhang Y. et al., 2025 [10]	Artificial Neural Network (ANN)	0.82-0.85	–	808 пациент, Пекин; ерте сатыдағы ГЦК
Liao H. et al., 2020 [11]	Random Forest (гистопатология)	0.886-0.988	–	TCGA + Батыс Қытай ауруханасы
Morshid A. et al., 2019 [12]	Random Forest (КТ радиомикасы)	0.73	–	105 пациент, MD Anderson, АҚШ
Johnson P.J. et al., 2024 [13]	Random Forest (GALAD-RF)	0.91	–	3 473 пациент, Жапония, бақылау бағдарламасы

Ескерту: AUROC – қысық астындағы аудан; «–» – зерттеуде көрсетілмеген; ГЦК – гепатоцеллюлярлық карцинома; TACE – транскатетерлік артериялық химиэмболизация.

Зерттеу әдістері

Зерттеуде Kaggle платформасынан алынған ашық қолжетімді деректер жиынтығы пайдаланылды.[14] Онда 5000 пациентті қамтиды. Мақсатты айналымы- Liver_cancer (0-ауру жоқ, 1- ауру бар). Ісік жоқ тобы – 78,22% , ісік бар тобы – 21,78% құрайды,

Зерттеуде қолданылған кіріс айнымалыларға пациенттің жасы, дене салмағының индексі, В және С вирустық гепатиттерінің болуы, бауыр функциясының интегралды көрсеткіші, альфа-фетопротеин деңгейі, бауыр циррозының анамнезі, онкологиялық аурулардың отбасылық анамнезі және қант диабетінің болуы жатады. Аталған айнымалылардың сипаттамасы, типі және мәндер диапазоны (кесте 2) толық көрсетілген.

Кесте 2 – Зерттеуде қолданылған кіріс айнымалылардың сипаттамасы

Айнымалы атауы	Сипаттамасы	Типі	Мәндер диапазоны / кодталуы
Age	Пациенттің жасы	Сандық	30 – 84 жас
BMI	Дене салмағының индексі	Сандық	8.1 – 43.7
Hepatitis B	В гепатитінің болуы	Категориялық (бинарлы)	0 – жоқ, 1 – бар
Hepatitis C	С гепатитінің болуы	Категориялық (бинарлы)	0 – жоқ, 1 – бар
Liver function score	Бауыр функциясының интегралды көрсеткіші	Сандық	5.5 – 115.7
Alpha-fetoprotein (AFP)	Альфа-фетопротеин деңгейі	Сандық	0.0 – 136.8
Cirrhosis history	Бауыр циррозының анамнезі	Категориялық (бинарлы)	0 – жоқ, 1 – бар
Family history of cancer	Онкологиялық аурулардың отбасылық анамнезі	Категориялық (бинарлы)	0 – жоқ, 1 – бар
Diabetes	Қант диабетінің болуы	Категориялық (бинарлы)	0 – жоқ, 1 – бар
Target (Liver cancer)	Бауыр қатерлі ісігінің болуы	Бинарлы мақсатты айнымалы	0 – жоқ, 1 – бар

Деректерді модельдерге беру алдында бірнеше өңдеу қадамы орындалды. Алдымен жиынтықтың толықтығы тексерілді- жоғалған міндер анықталмады. Сандық айнымалылар (жасы, ИДС,AFP, бауыр функциясы) стандарттау арқылы нормаланды. Әрбір белгінің орташа мәні 0-ге, стандартты ауытқуы-1-ге келтірілді. Бұл қадам әсіресе қашықтыққа сезімтал алгоритмдер үшін – kNN мен SVM аса маңызды себебі масштаб айырмасы болжам сапасын бұрмалауы мүмкін. Категориялық белгілер бастапқыда бинарлы форматта берілгендіктен қосымша кодтауды қажет етпеді. Деректер 70-те 30 қатынасында оқу және тест жиынтықтарына бөлінді.

Салыстырмалы талдауға жұмыс принциптері әртүрлі бес модель енгізілді. Logistic Regression – ең қарапайым және түсінікті базалық модель, нәтижені ықтималдық түрінде береді; зерттеуде қалған модельдермен салыстыру үшін бастапқы эталон ретінде алынды. SVM (тірек векторлар әдісі) – деректерді кеңістікте ең оңтайлы шекара арқылы бөлетін модель; бұл жерде RBF ядросы қолданылды, себебі ол сызықтық емес, күрделі байланыстарды да дәл ажырата алады. kNN (k ең жақын көршілер) – жаңа пациентті оған ең ұқсас 5 нүктенің көпшілігіне қарап жіктейтін алгоритм; қарапайым болғанымен деректер құрылымына өте сезімтал. Random Forest – 200 шешім ағашының дауысын біріктіретін ансамбльдік модель; артық үйрену қаупі аз, нәтижесі тұрақты. Gradient Boosting – алдыңғы модельдің қателерін келесісі түзете отырып, кезең-кезеңімен жетілдіретін ансамбль; күрделі байланыстарды меңгеруде ең тиімді алгоритм ретінде танылады. Барлық бес модель бірдей деректерде және бірдей шарттарда іске қосылды, бұл нәтижелерді тікелей салыстыруға мүмкіндік берді.

Класстар теңгерімсіз болғандықтан тек Ассурасу-мен шектелу жеткіліксіз. Сол себепті нәтижелер алты метрика арқылы бағаланды: ROC-AUC – кластарды ажырату қабілетінің жалпы өлшемі; Ассурасу – дұрыс жіктелген жағдайлардың жалпы үлесі; Precision – «ауру бар» деп болжанғандардың ішіндегі нақты ауру жағдайларының үлесі; Reremember/Sensitive – нақты ауру жағдайларын дұрыс анықтау қабілеті; Specificity – ауру жоқ жағдайларды дұрыс ажырату қабілеті; F1-score – нақтылық пен сезімталдықтың теңгерімді орташасы. Медициналық диагностикада жалған теріс нәтиженің – яғни ауру бар пациентті сау деп жіберіп

алудың – салдары ауыр болуы мүмкін, сондықтан Sensitivity мен F1-ге ерекше назар аударылды.

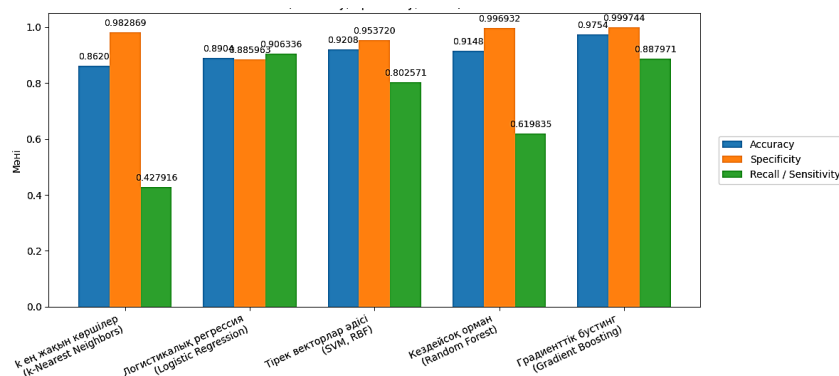
Зерттеу нәтижелері

Барлық бес модель бірдей алдын ала өңделген деректер жиынтығында оқытылып, бірдей тест үлгісінде бағаланды. Модельдің негізгі диагностикалық көрсеткіштері 3- кестеде жинақталған.

Кесте 3 – Машиналық оқыту модельдерінің диагностикалық нәтижелерін салыстыру

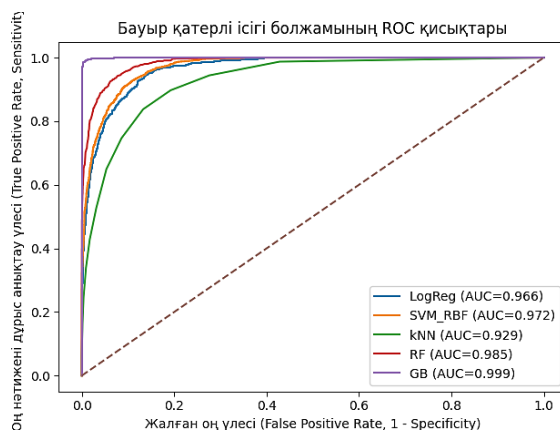
Модель	ROC_AUC	Accuracy	Precision	Recall (Sensitivity)	Specificity	F1
k-Nearest Neighbors	0.928641	0.8620	0.874296	0.427916	0.982869	0.574599
Logistic Regression	0.965860	0.8904	0.688765	0.906336	0.885963	0.782712
SVM (RBF)	0.971744	0.9208	0.828436	0.802571	0.953720	0.815299
Random Forest	0.984706	0.9148	0.982533	0.619835	0.996932	0.760135
Gradient Boosting	0.999498	0.9754	0.998967	0.887971	0.999744	0.940204

Модельдің Accuracy, Sensitivity және Specificity бойынша визуалды салыстыруы сурет 1 – берілген.



Сурет 1 – Бауыр қатерлі ісігін болжау модельдерінің Accuracy, Specificity және Recall метрикалары бойынша салыстырмалы талдауы

Дискриминациялық қабілетті бағалау үшін барлық модельдердің ROC қисықтары тұрғызылды (2-сурет). Gradient Boosting моделінің практикалық қолдану мүмкіндігін көрсету мақсатында тест жиынтығынан кездейсоқ таңдалған бес пациент үшін қатерлі ісіктің даму ықтималдығы жеке есептелді.



Сурет 2 – ML модельдерінің ROC қисықтары

Зерттеу нәтижелерін талқылау

Салыстырмалы талдау нәтижесінде модельдер арасында диагностикалық тиімділік тұрғысынан біраз айырмашылықтар анықталды. kNN алгоритмі сезімталдық тұрғысынан

барлық модельдер ішінде ең төменгі нәтиже көрсетті. Бұл нақты науқастардың жартысынан астамының анықталмай өтуіне алып келеді. Клиникалық скрининг контекстінде жалған теріс нәтиже – науқасты сау деп жіктеу ауыр медициналық салдарға апарады. Сол себепті бұл алгоритмнің диагностикалық құрал ретіндегі мүмкіндіктері шектеулі деп бағаланады.

Logistic Regression нақты ауру жағдайларының басым бөлігін дұрыс анықтағанымен нақтылық көрсеткішінің төмендігі жалған оң нәтижелердің жиілігін арттырады. Бұл оны бастапқы сатысында пайдалы етсе де, дербес диагностикалық шешім ретінде қолдану үшін қосымша растау қажеттілігін туындатады.

SVM (RBF) моделі сезімталдық пен ерекшелік арасында тепе-теңдік сақтады, бұл оны клиникалық жағдайда тұрақты нұсқа ретінде қарастыруға мүмкіндік береді. Бұл алгоритм жоғары нақтылық пен ерекшелік көрсеткенімен, сезімталдығының салыстырмалы төмендігі нақты ауру жағдайларының үлкен бөлігін өткізіп жіберуіне себеп болды – бұл скрининг мақсатында маңызды кемшілік болып табылады.

Gradient Boosting моделі барлық алты метрика бойынша үздік нәтиже берді. Бұл нәтиже халықаралық зерттеулермен үндеседі: Sato M. et al. [7] осы алгоритм үшін AUROC=0.940, сезімталдық 93.3% алса, Hashem S. et al. [8] балама ансамбльдік модельде AUROC 0.955-0.99 аралығын тіркеген. Алайда біздің зерттеудегі өте жоғары мәндер деректер жиынтығының синтетикалық сипатымен және шектеулі белгілер кеңістігімен түсіндіріледі – нақты клиникалық популяцияда көрсеткіштердің төмендеуі ықтимал, сондықтан сыртқы валидация қажет.

Қорытынды

Бұл жұмыста Kaggle платформасынан алынған ашық клиникалық деректер жиынтығы (N=5000) негізінде бауыр қатерлі ісігін болжауға арналған бес машиналық оқыту алгоритмі бірыңғай эксперименттік шарттарда салыстырылды. Нәтижелер модель таңдауының диагностикалық нәтижеге тікелей әсер ететінін айқын көрсетті.

Жалпы көрсеткіштер бойынша Gradient Boosting үздік нәтиже берді: ROC-AUC=0.9995, Дәлдік=0.9754, F1=0.940. Ансамбльдік алгоритмдер қателерді кезең-кезеңмен азайта отырып үйренетіндіктен, күрделі клиникалық деректерде жоғары тиімділік көрсетеді. Халықаралық зерттеулердің қорытындылары да осыны растайды [4, 15, 16].

Дегенмен барлық жағдайда Gradient Boosting ең дұрыс таңдау емес. Егер скрининг мақсатында ауруды жіберіп алмау басты міндет болса, Logistic Regression моделінің жоғары сезімталдығы (0.906) назар аударады. Ал kNN моделі сезімталдық бойынша (0.428) барлық алгоритмдерден айтарлықтай артта қалды – бауыр қатерлі ісігі сияқты жауапты диагностикада науқастардың жартысынан астамын анықтамай өткізу қолайсыз нәтиже болғандықтан, kNN осы мақсатқа жарамсыз деп есептеледі.

Зерттеудің шектеулерін де атап өту қажет. Деректер жиынтығында сыныптар арасындағы үлестік алшақтық бар – 78.22% және 21.78%, бұл кейбір модельдердің көпшілік топқа бейімделуіне әкелуі мүмкін. Деректердің синтетикалық сипатына байланысты Gradient Boosting-тің жоғары мәндері нақты клиникалық жағдайда төмендеуі ықтимал. Барлық модель тек ішкі тест деректерінде тексерілді – сыртқы валидация жүргізілмеді.

Болашақ зерттеулерде нақты аурухана деректерінде сыртқы валидация жүргізу, SHAP немесе LIME әдістері арқылы модель шешімдерінің түсіндірілімділігін арттыру, сондай-ақ SMOTE сияқты теңестіру әдістерін қолдану арқылы модельдің клиникалық практикаға жарамдылығын нығайту ұсынылады.

Жалпы алғанда, зерттеу нәтижелері машиналық оқыту алгоритмдерінің, әсіресе ансамбльдік әдістердің, бауыр қатерлі ісігін ерте болжауда практикалық қолданысқа жарамды екенін дәлелдейді. Алынған нәтижелер осы бағыттағы одан әрі клиникалық зерттеулер үшін негіз бола алады.

Әдебиеттер тізімі

1. Global Cancer Observatory (GCO). Liver and intrahepatic bile ducts: Fact sheet. GLOBOCAN 2022. – Lyon: International Agency for Research on Cancer (IARC), 2024. <https://gco.iarc.who.int/media/globocan/factsheets/cancers/11-liver-and-intrahepatic-bile-ducts-fact-sheet.pdf> (accessed date: 21.02.2026)
2. Hepatocellular Carcinoma: A Comprehensive Review / N. Amin et al // Diseases. – 2025. – Vol. 13, № 7. – Art. 207. <https://doi.org/10.3390/diseases13070207>.

3. European Association for the Study of the Liver. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma // Journal of Hepatology. – 2018. – Vol. 69, № 1. – P. 182-236. <https://doi.org/10.1016/j.jhep.2018.03.019>.
4. Current Status and Analysis of Machine Learning in Hepatocellular Carcinoma / S. Feng et al // Journal of Clinical and Translational Hepatology. – 2023. – Vol. 11, № 5. – P. 1184-1191. <https://doi.org/10.14218/JCTH.2022.00077S>.
5. Novel machine learning models outperform risk scores in predicting hepatocellular carcinoma in patients with chronic viral hepatitis / G.L.-H. Wong et al // JHEP Reports. – 2022. – Vol. 4, № 3. – Art. 100441. <https://doi.org/10.1016/j.jhepr.2022.100441>.
6. An artificial intelligence model to predict hepatocellular carcinoma risk in Korean and Caucasian patients with chronic hepatitis B / H.Y. Kim et al // Journal of Hepatology. – 2022. – Vol. 76, № 2. – P. 311-318. <https://doi.org/10.1016/j.jhep.2021.09.025>.
7. Machine-learning Approach for the Development of a Novel Predictive Model for the Diagnosis of Hepatocellular Carcinoma / M. Sato et al // Scientific Reports. – 2019. – Vol. 9, № 1. – Art. 7704. <https://doi.org/10.1038/s41598-019-44022-8>.
8. Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease / S. Hashem et al // Computer Methods and Programs in Biomedicine. – 2020. – Vol. 196. – Art. 105551. <https://doi.org/10.1016/j.cmpb.2020.105551>.
9. Feature reduction for hepatocellular carcinoma prediction using machine learning algorithms / G. Mostafa et al // Journal of Big Data. – 2024. – Vol. 11. – Art. 88. <https://doi.org/10.1186/s40537-024-00944-3>.
10. Machine learning model using immune indicators to predict outcomes in early liver cancer / Y. Zhang et al // World Journal of Gastroenterology. – 2025. – Vol. 31, № 5. – Art. 101722. <https://doi.org/10.3748/wjg.v31.i5.101722>.
11. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma / H. Liao et al // Clinical and Translational Medicine. – 2020. – Vol. 10, № 2. – Art. e102. <https://doi.org/10.1002/ctm2.102>.
12. A machine-learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization / A. Morshid et al // Radiology: Artificial Intelligence. – 2019. – Vol. 1, № 5. – Art. e180021. <https://doi.org/10.1148/ryai.2019180021>.
13. Serologic Detection of Hepatocellular Carcinoma: Application of Machine Learning and Implications for Diagnostic Models / P.J. Johnson et al // JCO Clinical Cancer Informatics. – 2024. – Vol. 8. – Art. e2300199. <https://doi.org/10.1200/CCI.23.00199>.
14. miadul. Predict Liver Cancer from & Clinical Features [Electronic resource] – Kaggle Datasets, 2024. – URL: <https://www.kaggle.com/datasets/miadul/predict-liver-cancer-from-and-clinical-features> (accessed date: 20.02.2026).
15. Reporting of prognostic clinical prediction models based on machine learning methods in oncology: a systematic review / P. Dhiman et al // Diagnostic and Prognostic Research. – 2022. – Vol. 6, № 1. – Art. 13. <https://doi.org/10.1186/s41512-022-00126-w>.
16. Development of an Early Lung Cancer Diagnosis Method Based on a Neural Network / I. Karymsakova et al // Computers. – 2025. – Vol. 14. – Art. 397. <https://doi.org/10.3390/computers14090397>.

Д.Ж. Шырынханова¹, И.Б. Карымсакова², Г. Жомартқызы¹, Д.О. Кожахметова², Т.А. Устинова²

¹Восточно-Казахстанский технический университет имени Д. Серикбаева,
070004, Республика Казахстан, ул. Серикбаева, 19, г. Усть-Каменогорск

²Шәкәрім университет,
071412, Республика Казахстан, г. Семей, ул. Глинки, 20 А

*e-mail: karymsakova.indira@mail.ru

ПРОГНОЗИРОВАНИЕ РАКА ПЕЧЕНИ С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ (ML)

На основе открытого набора структурированных клинических данных выполнена сравнительная оценка моделей машинного обучения для прогнозирования риска рака печени. В исследовании использован датасет платформы Kaggle (N=5000): целевая переменная Liver_cancer, 13 признаков (9 числовых, 4 категориальных), классы несбалансированы (0: 78,22%,

1: 21,78%). Этапы предобработки данных, обучения моделей и оценивания выполнены в рамках единого экспериментального протокола. Числовые переменные нормализованы посредством стандартизации; категориальные переменные закодированы в бинарном формате.

В сравнительный анализ включены логистическая регрессия (Logistic Regression), метод опорных векторов (Support Vector Machine с RBF-ядром), k -ближайших соседей (k -Nearest Neighbors), случайный лес (Random Forest) и градиентный бустинг (Gradient Boosting). Все модели оценивались по метрикам точности (accuracy), прецизионности (precision), чувствительности (recall/sensitivity), специфичности (specificity), F1-меры (F1-score) и площади под ROC-кривой (ROC-AUC).

По результатам эксперимента градиентный бустинг (Gradient Boosting) продемонстрировал наивысшую совокупную эффективность: ROC-AUC=0.999498, Accuracy=0.9754, Precision=0.998967, Sensitivity=0.887971, Specificity=0.999744, F1=0.940204. Хотя SVM (RBF) и Random Forest обеспечивали стабильную дискриминацию, по чувствительности и F1 они уступали Gradient Boosting. Полученные результаты показывают, что при выборе модели на основе структурированных клинических признаков необходим комплексный анализ метрик; в качестве ограничений отмечены несбалансированность классов и отсутствие внешней валидации.

Ключевые слова: рак печени, структурированные клинические данные, машинное обучение, градиентный бустинг, сравнительный анализ, ROC-AUC, чувствительность, специфичность.

D.Zh. Shyrynkhanova¹, I.B. Karymsakova*, G. Zhomartkyzy¹, D.O. Kozhakhmetova², T.A. Ustinova²

¹D. Serikbayev East Kazakhstan Technical University,
070004, Republic of Kazakhstan, Ust-Kamenogorsk, 19 Serikbayeva Street,

²Shakarim University,
071412, Republic of Kazakhstan, Semey, 20 A Glinka Street

*e-mail: karymsakova.indira@mail.ru

LIVER CANCER PREDICTION USING MACHINE LEARNING (ML) MODELS

A comparative evaluation of machine learning models for predicting liver cancer risk was performed using an open structured clinical dataset. A Kaggle dataset was used in the study (N=5000): target variable Liver_cancer, 13 features (9 numerical, 4 categorical), with class imbalance (0: 78.22%, 1: 21.78%). Data preprocessing, model training and evaluation were carried out within a unified experimental protocol. Numerical variables were normalized via standardization; categorical variables were encoded in binary format.

The comparative analysis included Logistic Regression, Support Vector Machine with an RBF kernel, k -Nearest Neighbors, Random Forest, and Gradient Boosting. All models were evaluated using accuracy, precision, recall (sensitivity), specificity, F1-score, and the area under the ROC curve (ROC-AUC).

Gradient Boosting achieved the highest overall performance: ROC-AUC=0.999498, Accuracy=0.9754, Precision=0.998967, Sensitivity=0.887971, Specificity=0.999744, F1=0.940204. Although SVM (RBF) and Random Forest demonstrated stable discrimination, they underperformed compared to Gradient Boosting in terms of sensitivity and F1. The results indicate that selecting a model based on structured clinical features requires a comprehensive multi-metric assessment; key limitations include class imbalance and the absence of external validation.

Key words: liver cancer, structured clinical data, machine learning, gradient boosting, comparative analysis, ROC-AUC, sensitivity, specificity.

Авторлар туралы мәліметтер

Динара Жаксылыковна Шырынханова – Цифрлық технологиялар және жасанды интеллект мектебінің докторанты, Д. Серикбаев атындағы Шығыс Қазақстан Техникалық Университеті, Өскемен қаласы, Қазақстан; e-mail: d.shyrynkhanova@shakarim.kz. ORCID: <https://orcid.org/0009-0007-5311-1221>.

Индира Бекеновна Карымсакова* – «Автоматтандыру және ақпараттық технологиялар» кафедрасының PhD докторы, Шәкәрім университеті, Семей қаласы, Қазақстан; e-mail: karymsakova.indira@mail.ru ORCID: <https://orcid.org/0000-0002-1496-3188>.

Гүлнәз Жомартқызы – Цифрлық технологиялар және жасанды интеллект мектебінің PhD докторы, қауымдастырылған профессоры, Д. Серикбаев атындағы Шығыс Қазақстан Техникалық Университеті, Өскемен қаласы, Қазақстан; e-mail: GZhomartkyzy@edu.ektu.kz. ORCID: <https://orcid.org/0000-0003-1465-3451>.

Динара Ошановна Кожаметова – «Автоматтандыру және ақпараттық технологиялар» кафедрасының PhD докторы, қауымдастырылған профессоры, Шәкәрім университеті, Семей қаласы, Қазақстан; e-mail: d.kojahmetova@shakarim.kz. ORCID: <https://orcid.org/0000-0002-4327-3899>.

Татьяна Анатольевна Устинова – «Автоматтандыру және ақпараттық технологиялар» кафедрасының аға оқытушысы, Шәкәрім университеті, Семей қаласы, Қазақстан; e-mail: ustinova-t-a@yandex.ru. ORCID: <https://orcid.org/0009-0001-8244-6043>.

Сведения об авторах

Динара Жаксылыковна Шырынханова – докторант Школы цифровых технологий и искусственного интеллекта, Восточно-Казахстанский технический университет имени Д. Серикбаева, г. Усть-Каменогорск, Казахстан; e-mail: d.shyrynkhanova@shakarim.kz. ORCID: <https://orcid.org/0009-0007-5311-1221>.

Индира Бекеновна Карымсакова* – PhD доктор кафедры «Автоматизация и информационные технологии», Шәкәрім университет, г. Семей, Казахстан; e-mail: karymsakova.indira@mail.ru. ORCID: <https://orcid.org/0000-0002-1496-3188>.

Гүлназ Жомартқызы – PhD доктор, ассоциированный профессор Школы цифровых технологий и искусственного интеллекта, Восточно-Казахстанский технический университет имени Д. Серикбаева, г. Усть-Каменогорск, Казахстан; e-mail: GZhomartkyzy@edu.ektu.kz. ORCID: <https://orcid.org/0000-0003-1465-3451>.

Динара Ошановна Кожаметова – PhD доктор, ассоциированный профессор кафедры «Автоматизация и информационные технологии», Шәкәрім Университет, г. Семей, Казахстан; e-mail: d.kojahmetova@shakarim.kz. ORCID: <https://orcid.org/0000-0002-4327-3899>.

Татьяна Анатольевна Устинова – старший преподаватель кафедры «Автоматизация и информационные технологии», Шәкәрім университет, г. Семей, Казахстан; e-mail: ustinova-t-a@yandex.ru ORCID: <https://orcid.org/0009-0001-8244-6043>.

Information about the authors

Dinara Zhaksylykovna Shyrynkhanova – Doctoral (PhD) student, School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan; e-mail: d.shyrynkhanova@shakarim.kz. ORCID: <https://orcid.org/0009-0007-5311-1221>.

Indira Bekenovna Karymsakova* – PhD, Department of Automation and Information Technologies, Shakarim University, Semey, Kazakhstan; e-mail: karymsakova.indira@mail.ru. ORCID: <https://orcid.org/0000-0002-1496-3188>.

Gulnaz Zhomartkyzy – PhD, Associate Professor, School of Digital Technologies and Artificial Intelligence, D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan; e-mail: GZhomartkyzy@edu.ektu.kz. ORCID: <https://orcid.org/0000-0003-1465-3451>.

Dinara Oshanovna Kozhakhmetova – PhD, Associate Professor, Department of Automation and Information Technologies, Shakarim University, Semey, Kazakhstan; e-mail: d.kojahmetova@shakarim.kz. ORCID: <https://orcid.org/0000-0002-4327-3899>.

Tatyana Anatolyevna Ustinova – Senior Lecturer of the Department of Automation and Information Technologies, Shakarim University, Semey, Kazakhstan; e-mail: ustinova-t-a@yandex.ru. ORCID: <https://orcid.org/0009-0001-8244-6043>.

Редакцияға енуі 30.03.2026

Өңдеуден кейін түсуі 14.05.2026

Жариялауға қабылданды 18.05.2026

[https://doi.org/10.53360/2788-7995-2026-2\(22\)-10](https://doi.org/10.53360/2788-7995-2026-2(22)-10)

FTAXP: 28.15.19



Ә.Ж. Жамбулатова*, Қ. Әлімхан

Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
010000, Қазақстан Республикасы, Астана қ., Сәтпаев көшесі, 2

*e-mail: aigerimzhambulat@gmail.com

КОММУТАЦИЯЛАНАТЫН КЕШІГҮІ БАР СЫЗЫҚТЫ ЕМЕС ЖҮЙЕЛЕРДІҢ ТҰРАҚТЫЛЫҒЫ

Аннотация: Бұл жұмыста уақыт бойынша кешігуі бар коммутацияланатын сызықты емес жүйелердің тұрақтылығы мәселесі қарастырылады. Қарастырылатын жүйе динамикасы бірнеше ішкі жүйелер арасында ауысуды сипаттайтын коммутация сигналы арқылы анықталады және уақыттық кешігу әсері ескеріледі. Коммутация сигналы ерікті түрде өзгертін жағдайда жүйенің

©Ә.Ж. Жамбулатова, Қ. Әлімхан, 2026