УДК 61:51-7(571.17) https://doi.org/10.21603/-I-IC-20

КОМПЛЕКСНАЯ ПРОГРАММА ПРОГНОЗИРОВАНИЯ РИСКА РАЗВИТИЯ РАКА ЛЕГКОГО У РАБОЧИХ УГОЛЬНЫХ ШАХТ КУЗБАССА

Н. А. Васильев[*], Ю. А. Степанов[*], М. Л. Баканова[*,**], В. И. Минина [*,**]
[*]Федеральное государственное бюджетное образовательное учреждение высшего образования «Кемеровский государственный университет», Кемерово, Россия
[**]Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр угля и углехимии Сибирского отделения Российской академии наук», Кемерово, Россия

Аннотация

Обследованы 214 рабочих угольных шахт Кузбасса больных раком легкого и 300 здоровых рабочих угольных шахт Кузбасса. Выявлена взаимосвязь полиморфизмов генов *XPD rs13181, TGFb1 rs1800471, CHEK2 rs555607708, EGFR rs2227984, EPHX1 rs1051740, APEX1 rs1130409, TNFα rs1800629, IL1b rs16944* с риском развития рака легкого. Разработана компьютерная программа прогнозирования риска развития злокачественные новообразования легких у рабочих угольных шахт Кузбасса.

Цель: разработать комплексную программу прогнозирования риска развития рака легкого у рабочих угольных шахт Кузбасса.

Ключевые слова: машинное обучение, информационные технологии, анализ данных, шахтеры, генетический полиморфизм, рак легкого

Воздействие комплекса факторов производственной среды предприятий угольного цикла, может привести к развитию различных легочных заболеваний, таких как злокачественные новообразования легких [1]. Варианты генов, осуществляющий контроль биотрансформации ксенобиотиков, репарации ДНК, контроля клеточного цикл и апоптоза, системы антиоксидантной защиты, трансмембранного рецептора семейства рецепторных тирозинкиназ ErbBC нередко ассоциируют с формированием рака легкого (РЛ). Поиск специфичных маркеров предрасположенности людей к онкологическим заболеваниям, сегодня являются основной предпосылкой для разработки программы прогнозирования повышенного риска онкозаболеваний.

Материалы и методы. Обследованы 214 рабочих угольных шахт Кузбасса больных РЛ и 300 здоровых рабочих угольных шахт Кузбасса, которые составили группу сравнения. Испытуемые были подобранны с учетом пола, возраста и статуса курения, и у каждого было взято согласие на исследование. Варианты генов *hOGG1 (rs1052133), PARP1 (rs1136410), APEX1 (rs1130409), XPD (rs13181), XRCC1(rs25489, rs25487, rs1799782), XPG (rs17655), XPC(rs2228001), ATM (rs1801516), NBS1(rs1805794), XRCC4(rs2075685), Ligase IV (rs1805389), CAT (rs1001179), TGFβ (rs1800469), EPHX1(rs1051740)* изучали методом аллель-специфической ПЦР (НПФ «Литех», г.Москва), а варианты генов *CYP1A1(rs4646903), CYP1A2(rs762551), GSTM1(del), GSTT1(del), TP53(rs1042522), TERT(rs2736100) CYP2D6, (rs35742686), CYP2E1 (rs2031920) GSTP1(rs1138272) (rs1695) XRCC2, (rs3218536) XRCC3(rs861539) MTHFR, (rs1801133), MTR (rs1805087), SOD2(rs4880), GPx1(rs1050450), TNFα, (rs1800629), IL1b (rs16944)* - Real-time PCR (ООО «СибДНК», г.Новосибирск). Статистическая обработка материала проводилась с помощью SNPstats (http://bioinfo.iconcologia.net/SNPstats), «Statistica 10.0» (StatSoft, Inc., USA). Для программного обеспечения, позволяющего

прогнозировать риск появления РЛ на основе вариантов генов человека, была использована классификационная модель машинного обучения «Случайный лес» (от анг. «Random Forest) [2].

Результаты. В результате анализа генетических факторов риска с помощью SNPStats (http://bioinfo.iconcologia.net/SNPstats) были определены 8 наиболее значимых полиморфных локусов: *XPD rs13181, TGFb1 rs1800471, CHEK2 rs555607708, EGFR rs2227984, EPHX1 rs1051740, APEX1 rs1130409, TNFα rs1800629, IL1b rs16944.*

Далее с помощью классификационной модели машинного обучения «Случайный лес» (от анг. «Random Forest) была разработана компьютерная программа прогнозирования риска развития РЛ у рабочих угольных шахт Кузбасса.

Выбор именно этой модели машинного обучения был сделан по причине того, что данная модель классификации является представителем отличного семейства базовых классификаторов, поскольку модели семейства достаточно сложны и могут достигать нулевой ошибки на любой выборке.

Программа при прогнозировании результата позволяет выбрать лишь некоторые из общего числа предусматриваемых генотипов, которые и будут участвовать в прогнозе, но делать это не рекомендуется, так как при этом модель будет иметь недостаточно объясняющих признаков во время обучения, что повлечёт изменение информационной энтропии в сторону увеличения неопределенности системы.

Используемая в программе модель проходит тестовую выборку, которая равна 30 процентам от общего количества данных, с точностью в 91.5 процента, что является достаточно высоким показателем, учитывая имеющийся объём данных. Для того, чтобы результат работы модели был высоким, необходима настройка данной модели, регулирование гиперпараметров, которые отвечают за внутреннюю структуру модели, количество деревьев, глубину деревьев, случайная мера, число признаков для расщепления, минимальное число объектов на листьях и использование для построения деревьев подвыборки с возвращением. Первоначально, базовая модель могла предугадать результат с точностью в 75 процентов, но после глубокого анализа и тонкой настройки гиперпараметров модели соответствующим образом, показатель был улучшен до 91.5 процента.

При запуске программы пользователя встречает главное окно (рис. 1). Перед тем, как начать прогнозирование, нужно указать такие параметры человека, как:

1. ФИО человека
2. Возраст
3. Место жительства
4. Место работы, должность
5. Является ли человек курильщиком
6. Имеются ли хронические заболевания лёгких
7. Генотипы

1. IL1b
2. TNF
3. APEX1
4. XPD
5. EGFR
6. CHEK2
7. THFb1
8. EPHX1

После того, как все необходимые поля введены, а генотипы выбраны, следует нажать кнопку «Спрогнозировать» для получения результата. После нажатия на кнопку, модель заново обучается по выбранным полям генотипов и выдаёт окно результата прогнозирования (рис. 2).

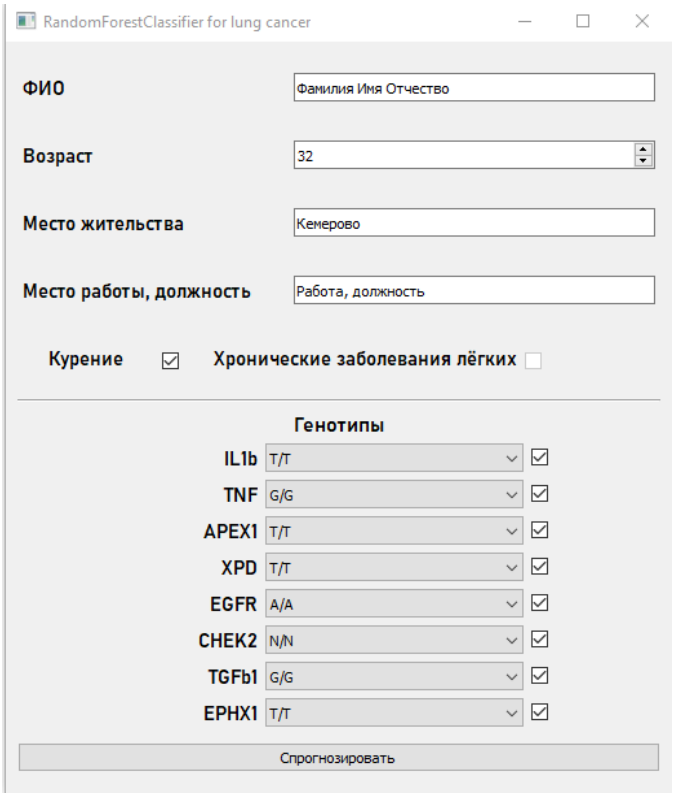

Рис. 1. Главное окно программы

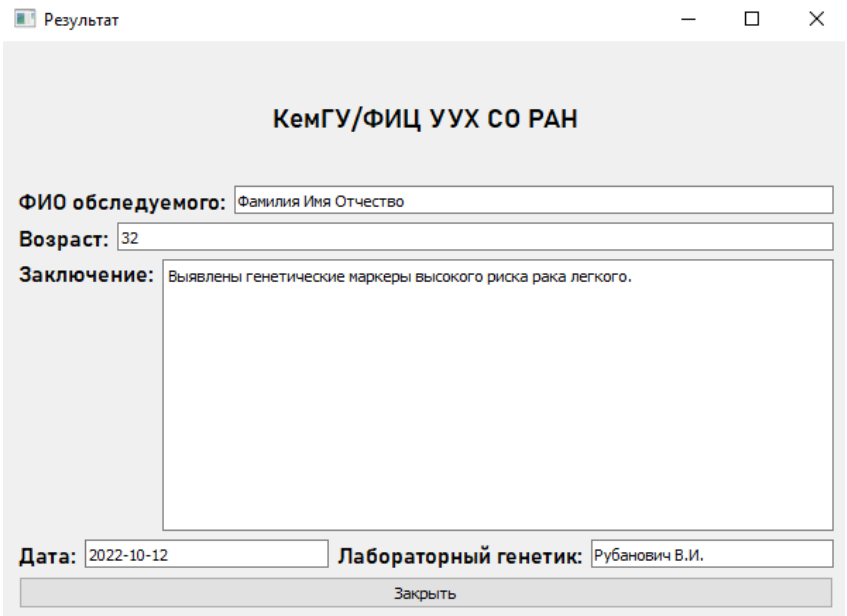

Рис. 2. Окно результатов

Список литературы

1.Taeger, D. Lung cancer among coal miners, ore miners and quarrymen: smoking-adjusted risk estimates from the synergy pooled analysis of case-control studies / D.Taeger, B.Pesch, B. Kendzia, et al. // Scand J Work Environ Health. 2015;41(5):467-477. doi:10.5271/sjweh.3513

2. sklearn.ensemble.RandomForestClassifier [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. – Дата доступа: 14.10.2022.

INTEGRATED PROGRAM FOR PREDICTION OF THE RISK OF LUNG CANCER DEVELOPMENT IN WORKING COAL MINES OF KUZBASS

N. A. Vasilev[*], Y.A. Stepanov[*], Bakanova M.L.[*,**], Minina V.I. [*,**]

[*] Kemerovo State University, Kemerovo, Russia [**] Federal State Budget Scientific Institution «The Federal Research Center of Coal and Coal Chemistry of Siberian Branch of the Russian Academy of Sciences», Kemerovo, Russia

Abstract

Polymorphic variants of genes of 214 miners of coal mines in Kuzbass in lung cancer patients, and 300 men of the control group were analyzed. A computer program has been developed for predicting the risk of developing lung cancer in the working coal mines of Kuzbass.

Objective: to develop a comprehensive program for predicting the risk of developing lung cancer in coal mine workers in Kuzbass.

Keywords: machine learning, information technology, data analysis, lung cancer, coal mines, genetic polymorphism

References

1. Taeger, D. Lung cancer among coal miners, ore miners and quarrymen: smoking-adjusted risk estimates from the synergy pooled analysis of case-control studies / D.Taeger, B.Pesch, B. Kendzia, et al. // Scand J Work Environ Health. 2015;41(5):467-477. doi:10.5271/sjweh.3513

2. sklearn.ensemble.RandomForestClassifier [Electronic resource]. – Access mode: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. – Access date: 14.10.2022.