*М. М. МИЩЕНКО, А. Д. КАРПЕНКО*

**ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОЦЕНКИ БИОТОКСИЧНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ ДЛЯ ИДЕНТИФИКАЦИИ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ BCR-ABL ТИРОЗИНКИНАЗЫ**

*ГУО «Университет НАН Беларуси», 220070 г. Минск, ул. Радиальная 38Б, Беларусь*

*Объединенный институт проблем информатики НАН Беларуси, 220012 г. Минск, ул. Сурганова 6, Беларусь*

*ГУО «Белорусский государственный университет», Факультет прикладной математики и информатики, 220030, г.Минск, пр.Независимости 4, Беларусь*

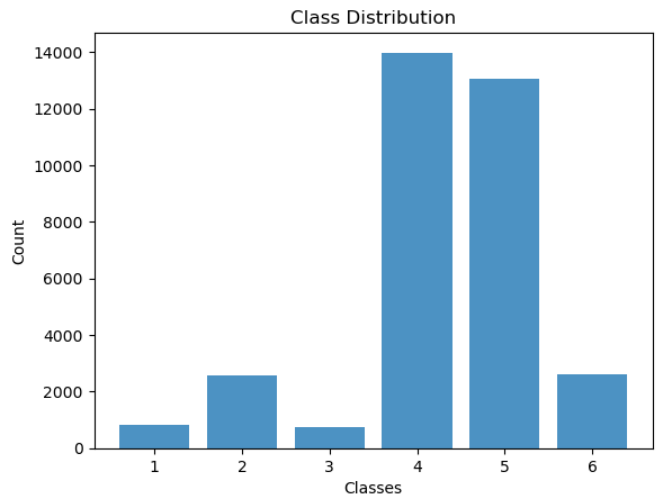
*Email:* [*mischenko.maria2004@gmail.com*](mailto:mischenko.maria2004@gmail.com)*, rfe.karpenko@gmail.com*

**Введение.** Токсичность и сопутствующие побочные эффекты представляют собой основные факторы, способствующие неэффективности потенциальных фармацевтических препаратов. В силу токсических свойств количество новых лекарств, выходящих на рынок в рамках традиционного процесса их разработки, постепенно сокращается на протяжении многих лет. На данный момент одним из актуальных направлений в области разработки новых лекарственных средств и материалов является оценка биотоксичности низкомолекулярных соединений. Этот подход позволяет на ранних стадиях разработки исключать потенциально опасные вещества и сосредоточить ресурсы на наиболее перспективных соединениях. Модели машинного обучения, предназначенные для прогнозирования токсичности, способны существенно сократить затраты и время, затрачиваемые на экспериментальные исследования. В то же время они решают этические проблемы, сводя к минимуму необходимость проведения испытаний на животных и клинических испытаний, что делает их особенно привлекательными для применения в современных научных изысканиях.

**Материалы и методы.** В данной работе стояла задача оценки биотоксичности ингибиторов BCR-ABL тирозинкиназы (nilotinib, imatinib, dasatinib, bosutinib, ponatinib, asciminib) с помощью применения одного из методов машинного обучения (в данной работе был выбран метод Random Forest). В качестве исходных данных был использован набор из 143335 соединений с уже известным классом токсичности для каждого из них, представленных в формате канонического SMILES. В качестве тренировочной выборки было взято 70% соединений из набора, в качестве тестовой - 30%. Для работы с соединениями необходимо было привести их в вид, приемлимый для обучения модели и работы с ними. Для этого для каждого соединения из SMILES при помощи функции **MolFromSmiles**  модуля **Chem** (https://www.rdkit.org/docs/source/rdkit.Chem.rdchem.html) библиотеки **rdkit (**https://www.rdkit.org/**)** была получена структура молекулы, затем с помощью функции модуля AllChem той же библиотеки был получен Моргановский фингерпринт, который на следующем шаге был переведён в вид строки из нулей и единиц. Далее на тренировочной выборке была обучена ранее упомянутая модель, а затем с помощью уже обученной модели были предсказаны классы токсичности для соединений тестовой выборки, а так же для требуемых ингибиторов BCR-ABL тирозинкиназы. Результатом каждого предсказания мог являться один из шести возможных классов токсичности в зависимости от степени токсичности соединения:

* Класс 1 (A): смертельно при проглатывании (LD50 ≤ 5);
* Класс 2 (B): смертельно при проглатывании (5 < LD50 ≤ 50);
* Класс 3 (C): токсичен при проглатывании (50 < LD50 ≤ 300);
* Класс 4 (D): вреден при проглатывании (300 < LD50 ≤ 2000);
* Класс 5 (E): может быть вреден при проглатывании (2000 < LD50 ≤ 5000);
* Класс 6 (F): нетоксичный (LD50 > 5000).

**Результаты и выводы.** В результате был получено следующее классовое распределение



*Рис.1 – Классовое распределение тестовой выборки*

Оценка качества модели была проведена с помощью функции потерь Multiclass Cross-Entropy, вычисляемой по следующей формуле:

где p(x) – реальное распределение вероятностей каждого из шести ингибиторов BCR ABL тирозинкиназы

q(x) – распределение вероятностей каждого из шести ингибиторов BCR ABL тирозинкиназы, полученное моделью

В дальнейших исследованиях рассматривается перспектива улучшения модели для предсказания биотоксичности путём комбинирования различных методов машинного обучения с целью выявления наиболее эффективной комбинации и улучшения точности получаемых прогнозов.