УДК 004.896

РАЗРАБОТКА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ РИСКА ИНФАРКТА МИОКАРДА НА ОСНОВЕ ДАННЫХ МАССОВОГО МЕДИЦИНСКОГО ОПРОСА BRFSS 2022 UPDATE

Е. С. Палто

Брестский государственный технический университет,
Брест, Беларусь, alena_sokolovskau@mail.ru
*Научный руководитель – **В. А. Головко**, заведующий кафедрой интеллектуальных информационных технологий БрГТУ, доктор технических наук, профессор*
vladimir.golovko@gmail.com

В данной работе рассматриваются различные методы машинного обучения, включая как классические методы, так и нейронные сети, для прогнозирования риска инфаркта миокарда на основе данных BRFSS 2022. Также рассмотрены разнообразные методы предобработки данных. Эксперименты показали высокую эффективность моделей LightGBM и FCNN с точностью 0,948 и 0,943 соответственно, что подтверждает важность применения машинного обучения для ранней диагностики инфаркта миокарда.

Ключевые слова: прогнозирование инфаркта миокарда; машинное обучение; нейронные сети.

DEVELOPMENT AND COMPARATIVE ANALYSIS OF MLMs FOR MYOCARDIAL INFARCTION RISK PREDICTION BASED ON BRFSS 2022 DATA

E. S. Palto

Brest State Technical University,
Brest, Belarus, alena_sokolovskau@mail.ru
*Scientific supervisor – **V. A. Golovko**, head of the Intelligence Information Technologies Department, Brest STU, Doctoral science degree in Computer Science, Professor*
vladimir.golovko@gmail.com

This paper examines various machine learning methods, including both classical algorithms and neural networks, for predicting the risk of myocardial infarction based on BRFSS 2022 data. Additionally, various data preprocessing techniques are explored. The experiments demonstrated high effectiveness of the LightGBM and FCNN models, achieving accuracies of 0.948 and 0.943, respectively, which underscores the significance of applying machine learning for the early diagnosis of myocardial infarction.

Keywords: myocardial infarction prediction; machine learning; neural networks.

1. Введение

Современные достижения в области машинного обучения значительно расширяют возможности обработки и анализа различного рода данных, что особенно актуально в медицине, где объемы информации достигают сотен тысяч единиц и крайне важно учесть все особенности каждого пациента. Быстрое развитие методов машинного обучения способствует созданию высокоэффективных инструментов для ранней диагностики и прогнозирования различных заболеваний. В настоящее время наблюдается возрастающая тенденция разработки систем искусственного интеллекта, основанных на нейронных сетях, что открывает новые перспективы для более точных, адаптивных и персонализированных решений в медицинской практике.

Одним из наиболее распространённых и клинически значимых заболеваний является инфаркт миокарда, также известный как сердечный приступ. Это заболевание остаётся одной из ведущих причин глобальной смертности, при этом около 30% случаев инфаркта миокарда заканчиваются летальным исходом до госпитализации. Высокая смертность от этого заболевания подчеркивает актуальность разработки методов ранней диагностики и прогнозирования.

Целью настоящего исследования является разработка и всесторонняя оценка моделей машинного обучения, направленных на прогнозирование риска возникновения инфаркта миокарда у пациентов. Для этого будут использованы данные опросов, собранных в рамках Behavioral Risk Factor Surveillance System (BRFSS) [1], ежегодного эпидемиологического исследования, проводимого Центром по контролю и профилактике заболеваний США. Данные BRFSS 2022 [2] года включают информацию о более чем 400 000 респондентов и содержат демографические, поведенческие и медицинские параметры, такие как состояние здоровья, наличие хронических заболеваний и поведенческие факторы, что открывает широкие возможности для применения методов анализа данных и прогнозирования различных заболеваний у респондентов на ранних стадиях.

Важно отметить, что различные исследования подтвердили высокую эффективность обучения современных моделей машинного обучения на основе данных BRFSS. Например, авторы работы, опубликованной на PubMed в 2025 году, предложили архитектуру Dual-Path ANN (DP-ANN), которая объединила трёхкритериальный отбор признаков и новую схему Minority-Weighted Sampling. Модель показала AUROC 0,895 на данных BRFSS и NHIS, с практически симметричными значениями специфичности (80%) и чувствительности (82%) [3].

Другое исследование также подтвердило эффективность методов машинного обучения на данных BRFSS. Анализ 438 693 анкет BRFSS-2020 показал, что алгоритмы XGBoost и AdaBoost достигли AUROC 0,91, что на 10% выше результатов базовой модели логистической регрессии. Эти результаты подтверждают, что более сложные модели способны значительно улучшить прогнозирование риска заболеваний, включая инфаркт миокарда [4].

В рамках данного научного исследования решена задача бинарной классификации, целью которой является прогнозирование наличия перенесённого инфаркта миокарда, представленного в виде бинарного индикатора. На основе проведённых экспериментов были получены высокие результаты с использованием различных моделей машинного обучения. Наибольшую эффективность продемонстрировали модели LightGBM, FCNN и DummyClassifier, достигнувшие значений точности 0,948, 0,944 и 0,943 соответственно. Эти результаты подтверждают высокую применимость данных моделей для решения задач прогнозирования в области кардиологии и подчеркивают значимость машинного обучения в разработке инструментов для ранней диагностики инфаркта миокарда.

2. Анализ данных

Набор данных содержит информацию о более чем 400 000 респондентов и включает в себя 40 признаков. Целевой переменной является HadHeartAttack — бинарный признак, указывающий, был ли у респондента инфаркт миокарда. Описание параметров опроса BRFSS 2022 [2] представлено в табл. 1.

Таблица 1

Параметры опроса BRFSS 2022

Название параметра	Значение параметра
PhysicalHealthDays	Количество дней за последний месяц, когда у респондента были физические проблемы со здоровьем.
MentalHealthDays	Количество дней за последний месяц, когда у респондента были проблемы с психическим здоровьем.
SleepHours,Среднее	Среднее количество часов сна в сутки.
HeightInMeters	Рост респондента в метрах.
WeightInKilograms	Вес респондента в килограммах.
BMI	Индекс массы тела респондента.
State	Штат США, в котором проживает респондент.
Sex	Пол респондента.
GeneralHealth	Общее состояние здоровья.
LastCheckupTime	Когда в последний раз респондент проходил медицинский осмотр.
PhysicalActivities	Занимается ли респондент физической активностью в свободное время.

Окончание табл. 1

Название параметра	Значение параметра
RemovedTeeth	Количество удалённых зубов.
HadHeartAttack	Был ли у респондента инфаркт миокарда — целевой признак.
HadAngina	Страдал ли респондент стенокардией.
HadStroke	Был ли у респондента инсульт
HadAsthma	Есть ли у респондента астма
HadSkinCancer	Болел ли респондент раком кожи.
HadCOPD	Болен ли хронической обструктивной болезнью лёгких.
HadDepressiveDisorder	Диагностировано ли у респондента депрессивное расстройство.
HadKidneyDisease	Есть ли заболевание почек.
HadArthritis	Болен ли респондент артритом.
HadDiabetes	Болен ли респондент сахарным диабетом.
DeafOrHardOfHearing	Имеет ли респондент проблемы со слухом.
BlindOrVisionDifficulty	Имеет ли проблемы со зрением.
DifficultyConcentrating	Испытывает ли трудности с концентрацией, памятью или принятием решений.
DifficultyWalking	Испытывает ли трудности при ходьбе или подъеме по лестнице.
DifficultyDressingBathing	Испытывает ли трудности с одеванием или купанием.
DifficultyErrands	Испытывает ли трудности с выполнением повседневных дел.
SmokerStatus	Статус курильщика. Т. е. курит/курил ли респондент.
ECigaretteUsage	Использует ли респондент электронные сигареты.
ChestScan	Проходил ли сканирование грудной клетки (например, рентген, КТ, флюорография).
RaceEthnicityCategory	Расовая и этническая принадлежность респондента.
AgeCategory	Возрастная категория.
AlcoholDrinkers	Употребляет ли респондент алкоголь.
HIVTesting	HIVTesting,Проходил ли тестирование на ВИЧ.
FluVaxLast12	Делал ли респондент прививку от гриппа за последние 12 месяцев.
PneumoVaxEver	Получал ли прививку от пневмококковой инфекции.
TetanusLast10Tdap	Делал ли прививку от столбняка за последние 10 лет.
HighRiskLastYear	Относился ли респондент к группе повышенного риска (по здоровью) в прошлом году.

Примечание. Параметры, приведённые в опросе BRFSS 2022 [2].

3. Предобработка данных

Первым этапом предобработки данных является анализ баланса класса целевой переменной для обеспечения корректности последующего построения и оценки моделей машинного обучения. Результаты проверки представлены в табл. 2.

Таблица 2

Распределение классов целевой переменной HadHeartAttack

Значение HadHeartAttack	Доля наблюдений, %
No	94,3204
Yes	5,69760

Примечание. Распределение классов целевой переменной HadHeartAttack.

Как видно из табл. 2, распределение целевой переменной существенно несбалансированно.

Далее данные были проверены на наличие анкет с пропущенными параметрами: их доля оказалась незначительной, поэтому записи с отсутствующими значениями были удалены. Все категориальные переменные идентифицированы и преобразованы в числовой формат посредством one-hot, label-encoding или target-encoding; метод способа преобразование категориальных данных в числовые был выбран индивидуально по типу переменной.

Для устранения выраженного дисбаланса целевой переменной были апробированы два подхода.

1. Upsampling – метод случайного дублирования наблюдений миноритарного класса. В результате объём выборки увеличен до 568 918 записей, что обеспечило паритет между классами.

2. Downsampling – метод случайного удаление наблюдений мажоритарного класса. Итоговый размер выборки составил 34 242 записи, сохраняя равное представление классов при минимальной избыточности данных.

В качестве методов понижения размерности были выбраны два метода.

1. Principal Component Analysis (PCA) — линейное ортогональное преобразование, позволяющее сохранить максимальную дисперсию в главных компонентах и тем самым сократить размер признакового пространства без значимой потери информации.

2. Correlation-based Feature Selection (CFS) — отбор самых информативных признаков посредством построения и анализа тепловой корреляционной матрицы. Корреляционный анализ показал, что для повышения информативности и уменьшения размерности признакового пространства следует оставить лишь переменные с наибольшей связью с целевой переменной HadHeartAttack. К числу наиболее переменных с наибольшей связью с целевой переменной относятся: HadAngina, GeneralHealth, HadStroke, RemovedTeeth, AgeCategory, HadDiabetes, PhysicalHealthDays, HadCOPD, HadArthritis, HadKidneyDisease, ChestScan, DifficultyWalking, PhysicalActivities, SmokerStatus, Sex, LastCheckupTime, DeafOrHardOfHearing, HadSkinCancer, AlcoholDrinkers, PneumoVaxEver и State. Признаки с низкой корреляцией с целевой переменной были

исключены, что снижает риск переобучения и упрощает последующее моделирование без потери существенной информации.

4. Описание исследуемых моделей машинного обучения

Для выбора оптимальной модели был испытан широкий спектр алгоритмов машинного обучения: от классической логистической регрессии, RandomForest и др. до современных решений — LightGBM, FCNN разных архитектур.

Для повышения эффективности моделирования был применён метод Grid Search в сочетании с кросс-валидацией, что обеспечило систематическую оптимизацию гиперпараметров и выбор конфигураций, демонстрирующих наилучшую обобщающую способность на независимых данных.

Представленный комплексный подход обеспечил всестороннюю оценку как традиционных, так и современных архитектур машинного обучения.

Используются следующие архитектуры FCNN.

1. Первая архитектура:
    i. Полносвязный слой на 128 нейронов, функция активации ReLU;
    ii. Полносвязный слой на 64 нейрона, функция активации ReLU;
    iii. Полносвязный слой на 32 нейрона, функция активации ReLU;
    iv. Выходной полносвязный слой на 1 нейрон с функцией активации Sigmoid.
2. Вторая архитектура:
    i. Входной полносвязный слой на 128 нейронов с функцией активации LeakyReLU (alpha=0,01);
    ii. Слой пакетной нормализации (BatchNormalization);
    iii. Слой Dropout с коэффициентом 0,3;
    iv. Первый скрытый полносвязный слой на 64 нейрона с функцией активации LeakyReLU (alpha=0,01);
    v. Слой пакетной нормализации (BatchNormalization);
    vi. Слой Dropout с коэффициентом 0,3;
    vii. Второй скрытый полносвязный слой на 32 нейрона с функцией активации LeakyReLU (alpha=0,01);
    viii. Слой пакетной нормализации (BatchNormalization);
    ix. Слой Dropout с коэффициентом 0,3;
    x. Выходной полносвязный слой на 1 нейрон с функцией активации Sigmoid.

Архитектура SSD-Convolutional:

i. Свёрточный слой Conv1D на 32 фильтра;
ii. Слой пакетной нормализации (BatchNormalization);

iii. Слой максимального подвыборки (MaxPooling1D);
iv. Свёрточный слой Conv1D на 64 фильтра;
v. Слой пакетной нормализации BatchNormalization;
vi. Слой максимального подвыборки MaxPooling1D;
vii. Свёрточный слой Conv1D на 128 фильтров;
viii. Слой пакетной нормализации BatchNormalization;
ix. Слой глобального усреднения по временной оси (GlobalAveragePooling1D);
x. Полносвязный слой на 64 нейрона, функция активации ReLU;
xi. Слой Dropout для регуляризации (параметров 0);
xii. Выходной полносвязный слой на 1 нейрон с функцией активацией Sigmoid.

5. Анализ результатов работы моделей машинного обучения

Проведён систематический подбор гиперпараметров комбинированной модели PCA и логистическая регрессия с использованием GridSearch и 5-фолдной кросс-валидации. Проведён сравнительный анализ 81 конфигурации модели, включавших различные значения доли объяснённой дисперсии в методе главных компонент (PCA) — 0,85, 0,90 и 0,9; решателей — lbfgs, newton-cg, liblinear, saga; типов регуляризации — L1, L2, elasticnet; и значений параметра регуляризации C — 0,1, 1, 10, включая перебор l1_ratio — 0,3, 0,5, 0,7 для elasticnet. Важно отметить, что всего 81 валидная конфигурация, т. к. не все комбинации решателей и типов регуляризации совместимы.

Таким образом была найдена оптимальная конфигурация из исследуемых конфигураций:

pca__n_components: 0,95;
clf__solver: newton-cg;
clf__penalty: l2;
clf__C: 10.

Модель с данной конфигурацией на тестовой выборке достигла точности 0,805. Дополнительно были протестированы модели с данной конфигурацией, но с различными значениями параметра pca__n_components: сохраняя 95% дисперсии, сохраняя 50% дисперсии, сохраняя 20 параметров.

Также для моделей Logistic Regression, CatBoostClassifier, LightGBM, RandonForestClassifier с помощью алгоритма Grid Search были подобраны оптимальные гиперпараметры.

Модели с наилучшими показателями точности на независимой тестовой выборке представлены в табл. 3.

Таблица 3

Сравнительные показатели точности моделей машинного обучения при различных схемах балансировки классов и сокращения размерности

Название модели машинного обучения	Метод сокращения размерности	Метод балансировки классов	Accuracy
LightGBM	CFS	Downdampling	0,948
DummyClissifer	CFS	Upsampling	0,943
DummyClissifer	CFS	Downdampling	0,943
RandonForestClassifier	CFS	Downdampling	0,940
FCNN с первой архитектурой	CFS	Downdampling	0,930
FCNN с первой архитектурой	CFS	Upsampling	0,925
RandonForestClassifier	CFS	Upsampling	0,850
LightGBM	CFS	Upsampling	0,843
CatBoostClassifier	CFS	Upsampling	0,841
CatBoostClassifier	CFS	Downdampling	0,841
Logistic Regression	CFS	Upsampling	0,835
Logistic Regression	CFS	Downdampling	0,835
FCNN со второй архитектурой	CFS	Downdampling	0,831
Logistic Regression	PCA, сохраняя 95% дисперсии	Downdampling	0,805
Logistic Regression	PCA, сохраняя 20 параметров	Downdampling	0,801
FCNN с первой архитектурой	PCA, сохраняя 95% дисперсии	Downdampling	0,794
Logistic Regression	PCA, сохраняя 50% дисперсии	Downdampling	0,782
SSD-Convolutional	PCA, сохраняя 95% дисперсии	Downdampling	0,537

Примечание. Accuracy рассчитана на независимой тестовой выборке. PCA — *Principal Component Analysis* (метод главных компонент); CFS — *Correlation-based Feature Selection* (корреляционный отбор признаков). NN — полносвязная нейронная сеть.

Изучив табл. 3, можно сделать вывод о важности роли тщательной предобработки данных для повышения эффективности работы моделей машинного обучения.

С точки зрения алгоритмических архитектур моделей машинного обучения, представленных в данной работе, то LightGBM, нейронные сети различных архитектур и Dummy Classifier продемонстрировали высокую способность извлекать релевантные закономерности из многомерных медицинских данных и довольно точно предсказывать наличия инфаркта миокарда у респондентов. Их устойчивые показатели точности

подтверждают релевантность данных моделей для надёжного прогнозирования клинических исходов.

Таким образом, изучив результаты работы моделей машинного обучения, можно сделать вывод, что современный инструментарий машинного обучения способен не только дополнять, но и значительно усиливать традиционные методы анализа, открывая путь к более точным и персонализированным решениям в медицине.

Библиографические ссылки

1. Annual Survey Data. URL: https://www.cdc.gov/brfss/annual_data/annual_data.htm (date of access: 12.08.2025).

2. 2022 BRFSS Survey Data and Documentation. URL: https://www.cdc.gov/brfss/annual_data/annual_2022.html (date of access: 12.08.2025).

3. Fair and explainable Myocardial Infarction (MI) prediction: Novel strategies for feature selection and class imbalance correction / S. B. Akter [et al.] // Computational Biology and Medicine. 2025. Vol. 184. Article no. 109413. DOI: 10.1016/j.compbiomed.2024.109413.

4. Identification of Myocardial Infarction (MI) Probability from Imbalanced Medical Survey Data: An Artificial Neural Network (ANN) with Explainable AI (XAI) Insights / S. B. Akter [et al.] // medRxiv. 2024. URL: https://www.medrxiv.org/content/10.1101/2024.02.28.24303497v1 (date of access: 12.08.2025).