DOI 10.37539/2949-1991.2026.40.5.005

Поляков Иван Алексеевич, студент,
Белгородский государственный национальный исследовательский университет

Михелев Владимир Михайлович, к.т.н.,
Белгородский государственный национальный исследовательский университет

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ДИАГНОСТИКИ САХАРНОГО ДИАБЕТА НА ОСНОВЕ АЛГОРИТМА СЛУЧАЙНОГО ЛЕСА

Аннотация. В статье рассматривается эффективность алгоритма случайного леса для бинарной классификации сахарного диабета по данным медицинских анализов. Проанализирован датасет из 100.000 записей пациентов с восьмью клиническими признаками, проведено сравнение трех алгоритмов машинного обучения. Оценка метрик качества выполнена на основе матрицы ошибок, ROC-AUC и 5-кратной кросс-валидации.

Ключевые слова: Сахарный диабет, искусственный интеллект, машинное обучение, медицина, интеллектуальная система, случайный лес.

Введение

По последней информации Всемирной организации здравоохранения, число новых случаев заболеваний сахарным диабетом в Российской Федерации, на 2023 год достигло 427749 [1], что является рекордным показателем, с момента начала ведения статистики – 1984 года. Тема сахарного диабета крайне актуальная на сегодняшний день, так как данное заболевание является самым распространенным среди эндокринных.

Стандартные методы определения сахарного диабета, основанные на лабораторной диагностике, включающей в себя различные виды анализов обладают рядом ограничений. Для определения всевозможных факторов заболевания, конкретного метода лечения и индивидуальных показателей пациента, применение методов машинного обучения и искусственного интеллекта является все более предпочтительным в рамках современной медицины. Такие технологии помогают выявлять индивидуальный риск, выступая в качестве инструмента поддержки принятия врачебных решений. Эти тенденции определяют актуальность разработки, адаптированной и практичной интеллектуальной системы диагностики сахарного диабета.

Описание предметной области

Сахарным диабетом называют группу заболеваний обмена веществ, при которых возникает неконтролируемое увеличение уровня глюкозы в крови. Гормон инсулин отвечает за усвоение глюкозы клетками и обмен углеводов в организме. Различные патологические состояния могут обуславливать недостаточное выделение инсулина – диабет 1 типа или невосприимчивость клеток к этому гормону – диабет 2 типа. Сахарный диабет относится к самым распространенным заболеваниям эндокринной системы. Разные типы диабета диагностируются примерно у 8% людей в течение жизни, причем повсеместные особенности питания с каждым годом увеличивают число больных [2].

При таком заболевании необходимо вовремя провести диагностику и правильно установить тип болезни. Основные методы обследования для определения сахарного диабета:

- определение концентрации глюкозы в крови;
- оценка количества гликированного гемоглобина;

- исследование концентрации глюкозы в моче;
- глюкозотолерантный тест – вспомогательный метод, который используют при неоднозначных результатах предыдущих исследований, помогает выявить преддиабет [3].

Существующие системы диагностики сахарного диабета

В современной медицине набирает популярность создание комплексных платформ, направленных на объединение данных из разных источников, которые с помощью современных технологий помогают анализировать большие данные, определять закономерности и повышать точность определения заболеваний.

В первую очередь в 2022 году была приведена модель машинного обучения, основанная на алгоритме градиентного бустинга для выявления случаев сахарного диабета 1 типа среди пациентов, которым уже установили сахарный диабет 2 типа. Был проведен анализ всевозможных данных и алгоритм выявил наиболее значимые прогнозы ошибочного определения 2 типа сахарного диабета. Этот алгоритм был предложен для внедрения в медицинскую информационную систему в США в качестве Системы поддержки принятия врачебных решений (СППВР) [4].

Среди самых актуальных решений для работы с сахарным диабетом выделяется MiniMed Go Smart MDI – умная система для управления множественными ежедневными инъекциями, разработанная компанией Medtronic. Разрешение на публичное использование было получено в январе 2026 года. Данная система предназначена для пациентов с инсулинозависимым диабетом 1 и 2 типов. С помощью алгоритмов, основанных на принципах машинного обучения реализован следующий функционал:

- автоматическое отслеживание доз инсулина и уровня глюкозы в одном интерфейсе;
- оповещения о пропущенных дозах, которые помогают избежать состояния гипергликемии;
- калькулятор доз, ведет учет текущего уровня глюкозы в организме и планируемое потребление углеводов;
- отчеты для врача, которые сразу отправляются ему, что облегчает анализ данных и корректировку терапии [5].

Искусственный интеллект в современной медицине

Искусственный интеллект (ИИ) – наука и технология создания интеллектуальных систем, то есть систем, способных выполнять функции, ранее свойственные только человеку: в их числе способность правильно интерпретировать внешние данные, извлекать уроки из таких данных и использовать полученные знания для достижения конкретных целей и задач при помощи глубокой адаптации [6].

ИИ в современных реалиях является неотъемлемой частью ряда областей, в которых он применяется. Как было сказано раннее, искусственный интеллект в медицине занимает крайне важное место. Выделим несколько первичных областей работы ИИ:

- лечение: помогает персонализировать подходы к лечению, учитывая личные параметры каждого пациента;
- диагностика: ИИ помогает врачам быстро и точно определять заболевания, анализируя различные медицинские изображения или снимки;
- психическое здоровье: дает возможность определять признаки депрессии на начальных этапах, позволяя начать лечение на ранних этапах;
- медицинское оборудование: ИИ встраивается в всевозможное медицинское оборудование для облегчения настройки и повышения точности лечения [7].

Обзор методов машинного обучения

Машинное обучение (МО) – это подвид ИИ, раздела информатики, который занимается созданием компьютерных систем, способных решать задачи, предназначенные для человеческого интеллекта [8].

В контексте будущей интеллектуальной системы диагностики сахарного диабета, необходимо сделать акцент на разновидности МО «с учителем». Обучение с учителем решает задачу бинарной классификации, где необходимо по исходным данным анализов пациента соотнести его к одной из двух категорий:

- категория 0: диабет не выявлен;
- категория 1: высокий риск или наличие диабета.

Для решения задач бинарной классификации также важен выбор конкретного алгоритма для работы модели МО. Алгоритм должен обеспечивать:

- высокую точность прогноза;
- возможность переобучения;
- работу с разными типами данных;
- корректные и подробные результаты работы.

Для выбора оптимального алгоритма был проведен сравнительный анализ трех популярных решений, что отображено в таблице 1.

Таблица 1

Сравнительный анализ алгоритмов МО для задачи бинарной классификации

Критерий	**Логистическая регрессия**	**Случайный лес (Random Forest)**	**Градиентный бустинг**
Принцип работы алгоритма	Преобразование линейной комбинации признаков в вероятности от 0 до 1	Построение множества отдельных решающих деревьев с помощью ансамблей	Последовательное построение деревьев, исправляющих ошибки предыдущих
Основное преимущество алгоритма	Высокая скорость и низкий риск переобучения	Баланс точности и устойчивости алгоритма	Максимальная точность на структурированных данных
Главный недостаток	Невозможность создавать сложные нелинейные зависимости между признаками	Меньшая интерпретируемость и требует больше ресурсов	Сильная склонность к переобучению, требует долгой настройки
Интерпретируемость	**Очень высокая.** Ясная связь признаков и прогноза	**Высокая.** Корректная оценка важности признаков для модели в целом	**Средняя.** Корректная оценка важности признаков, но последовательность усложняет анализ

Исходя из информации, приведенной в таблице 1 можно сделать вывод, что алгоритм «Случайного леса» является наиболее сбалансированным решением для разработки интеллектуальной системы.

Выбор данных медицинских анализов

Основной целью при выборе медицинских анализов является исследования ряда факторов, связанных со здоровьем и их взаимосвязи для точной классификации диабета.

В качестве выбранных данных для обучения модели в работе используется набор данных «Diabetes Prediction Dataset» с платформы Kaggle. Выбранный датасет содержит записи о 100.000 пациентов, каждая из которых описывается восьмью признаками. Каждый признак является важным фактором при определении сахарного диабета и именно комплекс таких данных позволяет верно определить диагноз. Описание факторов:

- возраст: является важным фактором при прогнозировании риска диабета. По мере старения человека риск развития диабета возрастает. Это отчасти связано с такими факторами, как снижение физической активности, изменение гормонального уровня и повышенная вероятность развития других заболеваний, которые могут способствовать возникновению диабета;
- пол: может играть роль в риске развития диабета, хотя его влияние может различаться. Некоторые исследования показывают, что у мужчин риск диабета может быть несколько выше, чем у женщин;
- индекс массы тела (ИМТ): показатель, основанный на соотношении роста и веса человека, который используется для оценки количества жира в организме. Он часто применяется как индикатор общего весового статуса и может быть полезен для прогнозирования риска диабета. Более высокий ИМТ ассоциируется с повышенной вероятностью развития диабета 2 типа;
- гипертония или высокое кровяное давление: состояние, которое часто сопутствует диабету. Оба состояния имеют общие факторы риска и могут способствовать развитию друг друга. Наличие гипертонии увеличивает риск развития диабета 2 типа, и наоборот;
- сердечные заболевания: болезни сердца, включая ишемическую болезнь сердца и сердечную недостаточность, связаны с повышенным риском диабета. Взаимосвязь между сердечными заболеваниями и диабетом является двунаправленной: наличие одного состояния увеличивает риск развития другого;
- история курения: курение является изменяемым фактором риска развития диабета. Установлено, что курение сигарет увеличивает риск развития диабета 2 типа;
- уровень HbA1c: HbA1c (гликированный гемоглобин) – это показатель среднего уровня глюкозы в крови за последние 2 -3 месяца. Он отражает долгосрочный контроль уровня сахара в крови. Более высокие уровни HbA1c связаны с повышенным риском развития диабета и его осложнений;
- уровень глюкозы в крови: количество глюкозы, присутствующее в крови в конкретный момент времени. Повышенный уровень глюкозы в крови, особенно натощак или после употребления углеводов, может указывать на нарушение регуляции глюкозы и повышать риск развития диабета. Регулярный мониторинг уровня глюкозы в крови важен для диагностики и лечения диабета.

Совместный анализ этих признаков с помощью соответствующих статистических методов и методов машинного обучения может помочь в прогнозировании индивидуального риска развития диабета.

Метрики качества для оценки работы системы

Для оценки качества работы интеллектуальной системы диагностики сахарного диабета воспользуемся метриками бинарной классификации. Основной для их вычисления выступает матрица ошибок, которая включает в себя следующие значения:

- TP (True Positive) – верно предсказанные положительные случаи;

- TN (True Negative) – верно предсказанные отрицательные случаи;
- FP (False Positive) – ложноположительные ошибки;
- FN (False Negative) – ложноотрицательные ошибки.

На основе этих величин будут производиться расчеты для следующих метрик:

1. Точность (Accuracy) – доля правильных ответов модели, которая вычисляется по формуле:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Метрика позволяет обнаружить баланс классов и определить качество предсказания модели МО.

2. Точность положительных предсказаний (Precision) – показывает, насколько верны положительные прогнозы модели. Вычисляется по формуле:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Высокое значение Precision означает, что если модель поставила диагноз сахарный диабет, то в большинстве случаев он верен.

3. Полнота (Recall) – отражает способность модели обнаруживать заболевание по следующей формуле:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

Высокий показатель полноты важен в рамках медицины, так как пропуск болезни может иметь серьезные последствия для пациента.

4. F1-мера (F1-score) – гармоническое среднее между точностью положительных предсказаний и полнотой. Показатель, который позволяет оценить баланс между этими двумя метриками, рассчитывается по формуле:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

F1-мера близка к 1, если обе метрики имеют высокие показатели.

5. ROC-кривая и площадь под кривой (AUC) – определяются на матрице ошибок в координатах полей «доля ложноположительных ответов» (FPR) и «доля истинно положительных ответов» (TPR). Вычисляются по формулам:

$$\text{FPR} = \frac{FP}{FP + TN} \quad (5)$$

$$\text{TPR} = \frac{TP}{TP + FN} \quad (6)$$

Площадь под ROC-кривой является интегральной характеристикой, то есть чем ближе AUC к 1, тем лучше модель МО различает классы.

Данные метрики помогут в полной мере оценить качество работы интеллектуальной системы диагностики сахарного диабета.

Обучение модели машинного обучения

Модель машинного обучения строилась на ранее выбранном алгоритме «случайного леса». Для числовых признаков используется метод стандартизации и масштабирования. Эти методы позволяют избежать искажения обучения модели МО и задать одинаковую значимость всем признакам.

Для правильной бинарной классификации сахарного диабета в датасете есть целевая переменная «Наличие сахарного диабета». Но в ней наблюдается дисбаланс данных в сторону

здоровых пациентов. Для обучения на сбалансированной выборке данных используем специальную технику, чтобы уравнять количество здоровых и больных пациентов. SMOTE (Synthetic Minority Oversampling Technique) – техника пересэмплирования синтетического меньшинства – метод подготовки несбалансированного датасета к загрузке в модель МО, предполагающий дублирование класса, представителей которого в наборе меньше остальных [9].

Для дополнительной проверки стабильность модели МО выполняется пяти кратная кросс-валидация, которая демонстрирует среднюю точность и стандартное отклонение. Кросс-валидация – это один из ключевых методов оценки качества моделей машинного обучения. Этот метод помогает избежать переобучения и обеспечивает более точную оценку моделей [10].

Результаты работы

Матрица ошибок показана в качестве тепловой карты, где по горизонтали – предсказанные классы «Диабет» или «Нет диабета», а по вертикали – истинные классы. Верхняя левая ячейка с результатом в 16711 – это истинно отрицательные случаи, которые показывают пациентов, у которых действительно нет сахарного диабета. Верхняя правая ячейка – ложноположительные срабатывания системы, это те пациенты, у которых на самом деле нет сахарного диабета, но система ошибочно отнесла их к классу «Диабет». Нижняя левая ячейка показывает ложноотрицательные случаи. Здесь те случаи, когда пациенты действительно больны диабетом, но система не смогла их распознать – их количество 302 случая. Последняя, нижняя правая ячейка – это истинно положительные 1394 случая. Пациенты с подтвержденным диабетом, которых модель корректно распознала. Результаты представлены на рисунке 1.

Рисунок 1. Матрица ошибок

Далее по итогам работы интеллектуальной системы был получен график ROC-кривой, показывающий зависимость доли истинно положительных случаев от доли ложно положительных случаев при различных вариациях порога принятия решения. Кривая плавно поднимается от левого нижнего угла к верхнему левому, что говорит о хорошем разделении классов моделью МО. Также полученное при расчетах высокое значение AUC говорит об высоком качестве разделения классов. ROC-кривая представлена на рисунке 2.

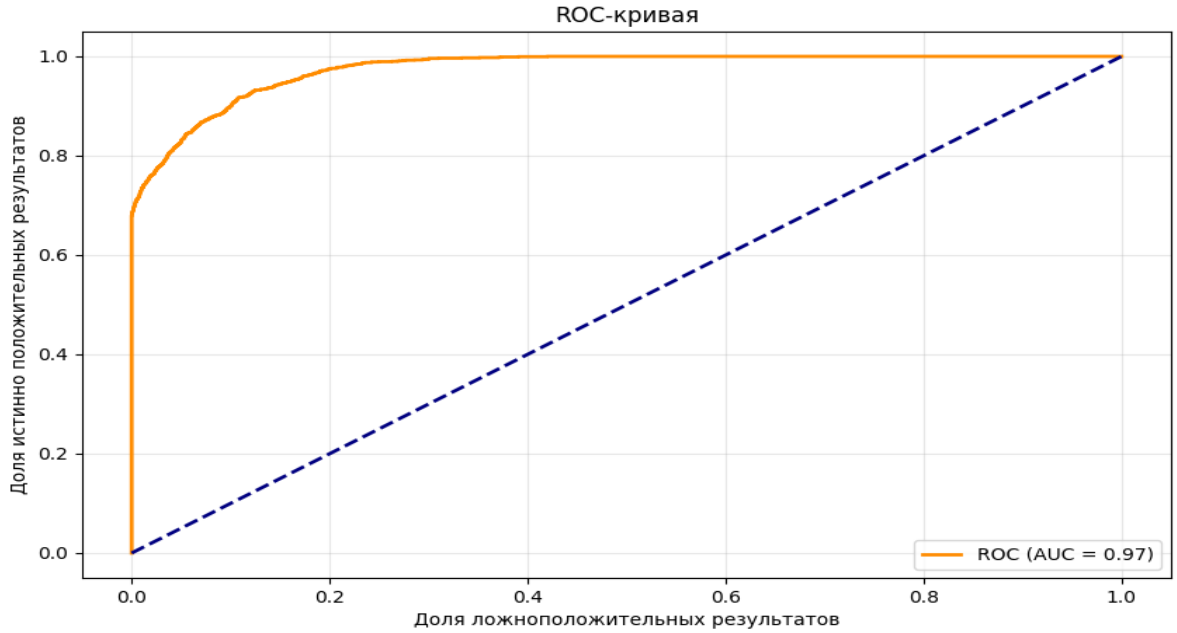

Рисунок 2. ROC – кривая

Столбчатая диаграмма кросс-валидации показывает высокий средний уровень точности интеллектуальной системы. Близость всех столбцов к среднему значению показывает отсутствие переобучения модели. Диаграмма изображена на рисунке 3.

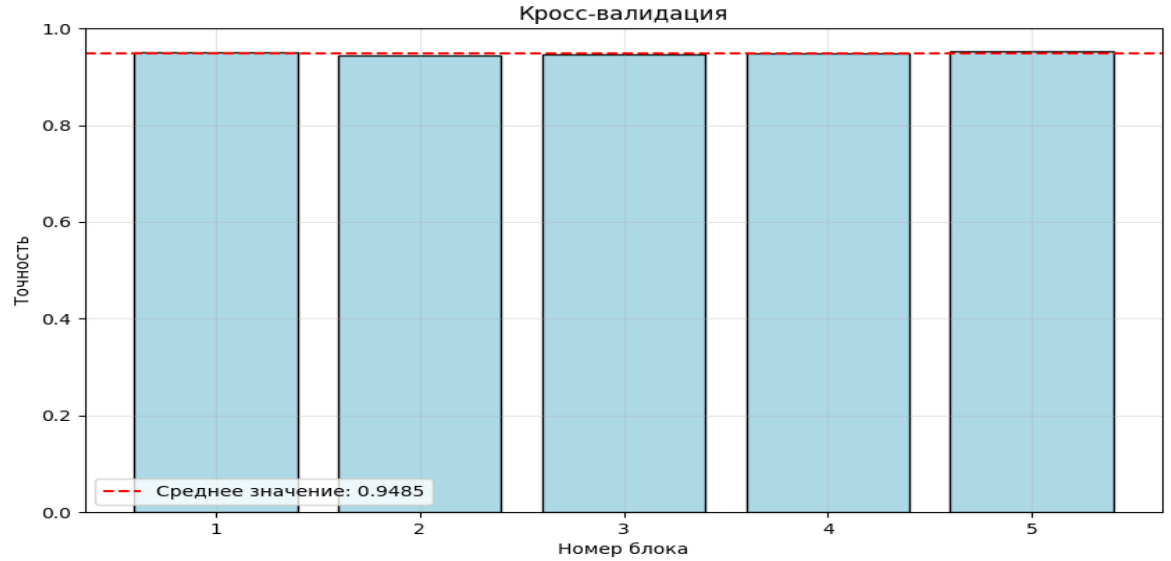

Рисунок 3. Столбчатая диаграмма кросс-валидации

Заключение

Разработанная интеллектуальная система диагностики сахарного диабета на основе алгоритма случайного леса показала высокий уровень бинарной классификации сахарного диабета. Итоговая точность составила 94,17 %, показатель полноты – 82,19 %. Площадь под ROC-кривой составила 0,9746, при доле ложноположительных ответов соответствующей всего 4,67 %. Пятикратная кросс-валидация наглядно показала и подтвердила стабильность модели, за счет высокого показателя средней точности равного 94,85 %. Наибольший вклад в прогноз внесли два показателя – это уровень гликированного гемоглобина и уровень глюкозы в крови, что соответствует клиническим представлениям о сахарном диабете. Полученные результаты позволяют рекомендовать разработанную интеллектуальную систему в качестве инструмента поддержки принятия врачебных решений.

Список литературы:

1. Распространенность сахарного диабета (%) [Электронный ресурс] – URL: https://gateway.euro.who.int/ru/indicators/hfa_379-2370-prevalence-of-diabetes-mellitus/#id=19310 (дата обращения: 25.04.2026)
2. Сахарный диабет [Электронный ресурс] – URL: https://www.smclinic.ru/diseases/sakharnyy-diabet/ (дата обращения: 25.04.2026)
3. Сахарный диабет: причины, симптомы, лечение [Электронный ресурс] – URL: https://cmed72.ru/information/articles/sakharnyy-diabet-prichiny-simptomy-lechenie/ (дата обращения: 25.04.2026)
4. Predicting misdiagnosed adult-onset type 1 diabetes using machine learning [Электронный ресурс] – URL: https://www.diabetesresearchclinicalpractice.com/article/S0168-8227 (22)00843-9/fulltext (дата обращения: 26.04.2026)
5. What is a Smart MDI System? [Электронный ресурс] – URL: https://www.medtronic-diabetes.com.au/products/smart-mdi-system (дата обращения: 26.04.2026)
6. Искусственный интеллект в диабетологии [Электронный ресурс] – URL: https://www.dia-endojournals.ru/jour/article/view/12665?locale=ru_RU (дата обращения: 26.04.2026)
7. Сферы применения искусственного интеллекта [Электронный ресурс] – URL: https://sbermed.ai/sfery-primeneniya-iskusstvennogo-intellekta (дата обращения: 27.04.2026)
8. Методы машинного обучения в дифференциальной диагностике сложно классифицируемых типов сахарного диабета [Электронный ресурс] – URL: https://www.dia-endojournals.ru/jour/article/view/13070 (дата обращения: 27.04.2026)
9. SMOTE в Машинном обучении простыми словами [Электронный ресурс] – URL: https://dzen.ru/a/YXz5mkT6EgivDfvB (дата обращения: 28.04.2026)
10. Что такое кросс-валидация [Электронный ресурс] – URL: https://evmservice.ru/blog/chto-takoe-kross-validaciya/ (дата обращения: 28.04.2026).