УДК 004.855.5

***[1]Арын А. Б.[1*], [2]Мамырбаев О. Ж., [3]Павлов С. В., 2025.**

[1]Caspian University, 050000, Алматы, Казахстан;
[2]РГП на ПХВ «Института информационных и вычислительных технологий» МНВО РК, 050000, Алматы, Казахстан;
[3]Кафедра биомедицинской инженерии и оптико-электронных систем, Винницкий национальный технический университет, 21021, Винница, Украина.
*E-mail: arun_arai@mail.ru

КЛАССИФИКАЦИЯ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ, ИСПОЛЬЗУЯ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Арын Арай Болатқызы, магистр технических наук, сеньор-лектор, Институт инженерии Каспийского университета, проспект Достык, 85А, Алматы, 050000, Казахстан.
E-mail: arun_arai@mail.ru, ORCID 0000-0001-7023-0424
Мамырбаев Оркен Жумажанович, PhD, профессор, РГП на ПХВ «Институт информационных и вычислительных технологий» МНВО РК, ул. Курмангазы, 29, Алматы, 050000, Казахстан.
E-mail: morkenj@mail.ru, ORCID 0000-0001-8318-3794
Павлов Сергей Владимирович, доктор технических наук, профессор, Кафедра биомедицинской инженерии и оптико-электронных систем Винницкого национального технического университета, Хмельницкое шоссе 95, Винница 21021, Украина.
E-mail: psv@vntu.edu.ua, ORCID 0000-0002-0051-5560

Статья посвящена исследованию методов классификации рака молочной железы с использованием современных технологий машинного обучения, таких как многослойный персептрон (MLP, Multi-Layer Perceptron), метод опорных векторов (SVM, Support Vector Machine), логистическая регрессия (Logistic Regression), (математическая регрессия), метод К-ближайших соседей (KNN, K-Nearest Neighbors), экстремальное повышение градиента (XGBoost, Extreme Gradient Boosting), случайный лес (RF, Random Forest), машина для повышения светового градиента (LightGBM, Light Gradient Boosting Machine), Наивный байесовский классификатор (Naive Bayes (GaussianNB)). В ходе работы были проанализированы различные показатели на наборе данных по диагностике рака молочной железы в Висконсине, доступные в системе машинного обучения UCI Repository. Результаты экспериментов демонстрируют высокую точность, обеспечивая надежное выявление рака молочной железы. Полученные данные подтверждают эффективность использования алгоритмов машинного обучения в области медицинской диагностики.

Ключевые слова: *рак молочной железы, классификация, нейронные сети, диагностика рака молочной железы, алгоритм машинного обучения.*

***[1]Арын А. Б., [2]Мамырбаев Ө.Ж., [3]Павлов С. В., 2025.**

[1]Caspian University, 050000, Алматы, Қазақстан;
[2]Шаруашылық жүргізу құқығындағы республикалық мемлекеттік кәсіпорын «Ақпараттық және есептеу технологиялары институты», 050000, Алматы, Қазақстан;
[3]Биомедициналық инженерия және оптикалық-электронды жүйелер кафедрасы, Винница ұлттық техникалық университеті, 2102,1 Винница, Украина.
*E-mail: arun_arai@mail.ru

МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП СҮТ БЕЗІ ҚАТЕРЛІ ІСІГІН ЖІКТЕУ

Арын Арай Болатқызы, техника ғылымдарының магистрі, сеньор-лектор.
E-mail: arun_arai@mail.ru, ORCID 0000-0001-7023-0424
Мамырбаев Оркен Жумажанович, PhD, профессор.
E-mail: morkenj@mail.ru, ORCID 0000-0001-8318-3794
Павлов Сергей Владимирович, техника ғылымдарының докторы, профессор.
E-mail: psv@vntu.edu.ua, ORCID 0000-0002-0051-5560

Бұл жұмыс көп қабатты перцептрон (MLP, Multi-Layer Perceptron), тіректі вектор әдісі (SVM), логистикалық регрессия, k-жақын көршілер әдісі (KNN, K-Nearest Neighbors), экстремалды градиентті арттыру (XGBoost, Extreme Gradient Boosting), кездейсоқ орман (RF, Random Forest), жарық градиентін арттыру машинасы (LightGBM, Light Gradient Boosting Machine), Байес классификаторы (Naive Bayes (GaussianNB)) әдістерін зерттеуге арналған. Жұмыс барысында UCI Repository машиналық оқыту жүйесінде қол жетімді түрде Висконсиндегі сүт безі обырын диагностикалау деректер жинағындағы әртүрлі көрсеткіштер талданды. Эксперименттердің нәтижелері сүт безі обырын сенімді анықтауды қамтамасыз ететін жоғары дәлдікті көрсетті. Нәтижелер медициналық диагностика саласында машиналық оқыту алгоритмдерін қолданудың тиімділігін растайды.

Түйін сөздер: *сүт безі қатерлі ісігі, жіктеу, нейрондық желілер, сүт безі қатерлі ісігінің диагностикасы, машиналық оқыту алгоритмі.*

***[1]Aryn Aray, [2]Mamyrbayev Orken, [3]Pavlov Sergey, 2025.**

[1]Caspian University, 050000, Almaty, Kazakhstan;
[2]Institute of Information and Computing Technologies of the Ministry of Internal Affairs of the Republic of Kazakhstan, 050000, Almaty, Kazakhstan;
[3]Department of Biomedical Engineering and Optoelectronic Systems, Vinnytsia National Technical University, 21021, Vinnytsia, Ukraine.
*E-mail: arun_arai@mail.ru

BREAST CANCER CLASSIFICATION USING MACHINE LEARNING METHODS

Aryn Aray Bolatovna, Master of Technical Sciences, Senior Lecturer.
E-mail: arun_arai@mail.ru, ORCID 0000-0001-7023-0424
Mamyrbayev Orken Zhumazhanovich, PhD, Professor.
E-mail: morkenj@mail.ru, ORCID 0000-0001-8318-3794
Pavlov Sergey Vladimirovich, Doctor of Technical Sciences, Professor.
E-mail: psv@vntu.edu.ua, ORCID 0000-0002-0051-5560

This work is devoted to the study of breast cancer classification methods using modern machine learning technologies, such as multilayer Perceptron (MLP, Multi-Layer Perceptron), Support Vector

Machine (SVM, Support Vector Machine), Logistic Regression (mathematical Regression), K-nearest neighbors method (KNN, K-Nearest Neighbors), extreme gradient boosting (XGBoost, Extreme Gradient Boosting), Random Forest (RF, Random Forest), Light Gradient Boosting Machine (LightGBM, Light Gradient Boosting Machine), Naive Bayes classifier (GaussianNB). In the course of the work, various indicators were analyzed on a dataset for the diagnosis of breast cancer in Wisconsin, available in the UCI Repository machine learning system. The experimental results demonstrate high accuracy, ensuring reliable detection of breast cancer. The data obtained confirms the effectiveness of using machine learning algorithms in the field of medical diagnostics.

Keywords: *breast cancer, classification, neural networks, breast cancer diagnosis, machine learning algorithm.*

ВВЕДЕНИЕ

Рак молочной железы – это заболевание, при котором клетки в молочной железе растут бесконтрольно, часто начиная с протоков или долек. Он может быть инвазивным или неинвазивным, с факторами риска, включая генетические, гормональные факторы и образ жизни. Раннее выявление с помощью скрининга значительно улучшает результаты [13]. Jamil R. и др. [6] представили высокоточную модель для выявления микрокальцификации, ключевого показателя ранней стадии рака молочной железы. Их модель Wiener LTI Tophat повышает контрастность изображения, позволяя более точно выявлять раковые участки. Этот метод демонстрирует потенциал передовых методов обработки изображений в улучшении ранней диагностики. Аналогичным образом Batool A., Byun Y.C.[2] исследовали применение алгоритма случайного леса в классификации рака молочной железы, подчеркивая его надежность при обработке сложных наборов данных и превосходную прогностическую эффективность по сравнению с традиционными статистическими методами. Кроме того, Strelcenia E., и Prakoonwit S.[12] провели сравнительное исследование методов разработки признаков и классификации, подчеркнув важность выбора оптимальных признаков для повышения производительности модели. Их исследование показало, что хорошо продуманный выбор признаков значительно влияет на точность классификации, усиливая потребность в усовершенствованных методологиях диагностики рака молочной железы. В связи с расширением применения машинного обучения в медицинской диагностике для дифференциации злокачественных и доброкачественных опухолей используются различные классификационные модели. В этом исследовании оцениваются несколько алгоритмов машинного обучения, сравнивается их эффективность с точки зрения точности, прецизионности, запоминания и показателя F1. Анализируя эти методы, исследователи стремятся определить наиболее эффективную классификационную модель для выявления рака молочной железы, способствующую развитию систем автоматизированной диагностики.

МАТЕРИАЛЫ И МЕТОДЫ

В данном исследовании анализированы различные методы классификации, обычно используемые для диагностики рака молочной железы, такие как многослойный персептрон (MLP, Multi-Layer Perceptron), метод опорных векторов (SVM, Support Vector Machine), логистическая регрессия (Logistic Regression), (математическая регрессия), метод К-ближайших соседей (KNN, K-Nearest Neighbors), экстремальное повышение градиента (XGBoost, Extreme Gradient Boosting), случайный лес (RF, Random Forest), машина для повышения светового градиента (LightGBM, Light Gradient Boosting Machine), Наивный байесовский классификатор (Naive Bayes (GaussianNB)).

Многослойный персептрон (MLP) – это класс искусственных нейронных сетей прямого действия, которые состоят из нескольких слоев узлов, включая входной слой, один или более скрытых слоев и выходной слой. Каждый узел или нейрон в одном слое имеет определенный вес с каждым нейроном в следующем слое. MLP используют методы контролируемого обучения и обратного распространения для обучения, что делает их пригодными для задач классификации [4].

Метод опорных векторов (SVM) – это управляемый алгоритм машинного обучения, обычно используемый для задач классификации. Он работает путем нахождения оптимальной

гиперплоскости, которая разделяет точки данных разных классов в многомерном пространстве, максимизируя разницу между этими классами для повышения точности классификации [8].

Логистическая регрессия (LR) является алгоритмом контролируемого обучения, используемый для задач бинарной классификации. Он моделирует вероятность бинарного результата путем подгонки данных к логистической функции, что делает его пригодным для различения двух классов, таких как доброкачественные и злокачественные опухоли [3].

Метод К-ближайших соседей (KNN) – непараметрический алгоритм обучения на основе примеров, используемый для задач классификации и регрессии. Он работает путем определения "k" ближайших точек данных (соседей) к заданным входным данным и присвоения входным данным наиболее распространенного класса среди этих соседей. Алгоритм использует метрику расстояния, обычно евклидово расстояние, для измерения сходства между точками данных [5].

Экстремальное повышение градиента (XGBoost) – это оптимизированный фреймворк для повышения градиента, разработанный для повышения эффективности и быстродействия. Он последовательно строит совокупность деревьев решений, где каждое дерево пытается исправить ошибки своего предшественника, повышая точность прогнозирования модели. XGBoost использует методы регуляризации для предотвращения переобучения и поддерживает параллельную обработку, что делает его популярным выбором для различных задач машинного обучения [9].

Случайный лес (RF) – это метод коллективного обучения, который в процессе обучения создает несколько деревьев решений и выводит режим их прогнозирования для задач классификации. Этот подход повышает точность прогнозирования и контролирует переобучение путем усреднения результатов различных деревьев, каждое из которых построено на разных подмножествах данных и признаков [10].

Машина для повышения светового градиента (Light Gradient Boosting Machine) основан на методах дерева решений. Он разработан для повышения эффективности и быстродействия, особенно при работе с большими наборами данных. В контексте классификации рака молочной железы LightGBM часто используется для различения доброкачественных и злокачественных опухолей путем анализа различных признаков, извлеченных из медицинских данных, таких как результаты визуализации или истории болезни пациентов [7].

Наивный Байесовский классификатор (GaussianNB) – это вероятностный классификатор, основанный на теореме Байеса, которая предполагает, что признаки, используемые для классификации, условно независимы от обозначения класса. Несмотря на это, упрощающее предположение – NB может эффективно выполнять различные задачи классификации [1].

В исследовании использовались наборы данных по диагностике рака молочной железы в Висконсине, доступные в системе машинного обучения UCI Repository, онлайн-хранилище с открытым исходным кодом (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic). Этот набор данных содержит 569 записей. Эти признаки извлекаются из оцифрованных изображений образцов тканей молочной железы, взятых с помощью тонкоигольного аспирата (FNA), и используются для классификации рака молочной железы.

Таблица 1 - Подробные сведения о наборе данных по диагностике рака молочной железы в штате Висконсин (WDBC).

№	Данные характеристик	Определения
0	id	Уникальный идентификационный номер для каждого образца
1	diagnosis	Классификация опухоли ("M" - злокачественная, "B" - доброкачественная)
2	radius_mean	Среднее значение расстояний от центра до периметра ядра
3	texture_mean	Стандартное отклонение значений серой шкалы на изображении
4	perimeter_mean	Среднее значение длины периметра ядра
5	area_mean	Средняя площадь ядра
6	smoothness_mean	Показатель того, насколько ровной является граница ядра
7	compactness_mean	Отношение периметра в квадрате к площади минус 1,0.
8	concavity_mean	Средняя выраженность вогнутости участков контура ядра
9	concave points_mean	Среднее количество вогнутых участков в контуре ядра
10	symmetry_mean	Показатель симметрии формы ядра

11	fractal_dimension_mean	Мера сложности на границе ядра
12	radius_se	Стандартная ошибка радиуса
13	texture_se	Стандартная ошибка текстуры
14	perimeter_se	Стандартная ошибка периметра
15	area_se	Стандартная ошибка площади
16	smoothness_se	Стандартная ошибка сглаживания
17	compactness_se	Стандартная ошибка компактности
18	concavity_se	Стандартная ошибка вогнутости.
19	concave points_se	Стандартная ошибка вогнутых точек
20	symmetry_se	Стандартная ошибка симметрии
21	fractal_dimension_se	Стандартная ошибка фрактальной размерности
22	radius_worst	Наибольший наблюдаемый радиус
23	texture_worst	Наибольшее наблюдаемое значение текстуры
24	perimeter_worst	Наибольший наблюдаемый периметр
25	area_worst	Наибольшая наблюдаемая площадь
26	smoothness_worst	Наихудший показатель гладкости
27	compactness_worst	Наихудший показатель компактности
28	concavity_worst	Наихудший показатель вогнутости
29	concave points_worst	Наихудший показатель вогнутости точек
30	symmetry_worst	Наихудший показатель симметрии
31	fractal_dimension_worst	Наихудший показатель фрактальной размерности

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В данном разделе представлена оценка эффективности различных классификаторов машинного обучения, используемых для диагностики рака молочной железы. Анализируемые модели включают логистическую регрессию, многослойный персептрон (MLP), метод опорных векторов (SVM), K-ближайших соседей (KNN), XGBoost, случайный лес, LightGBM и наивный Байесовский алгоритм. Их эффективность оценивается на основе ключевых показателей, таких как точность, четкость, запоминаемость, F1 показатель и матрицы путаницы. Результаты позволяют получить представление о классификационных возможностях каждого алгоритма, определяя наиболее подходящую модель для точного и надежного выявления рака молочной железы. Проводится сравнительный анализ, чтобы выявить сильные и слабые стороны каждого подхода. Результаты этой оценки помогают выбрать оптимальные методы машинного обучения для диагностики рака молочной железы. Результаты анализа производительности многослойной нейронной сети (MLP) представлены в таблице 2. Общая точность модели составила 94%, что свидетельствует о высокой прогностической способности при различении доброкачественных и злокачественных опухолей. Четкость модели MLP составила 0,92 для класса 0 (доброкачественные опухоли) и 0,97 для класса 1 (злокачественные опухоли), что указывает на то, что модель эффективно выявляет положительные случаи с высокой степенью достоверности. Что касается запоминаемости, то модель показала 99%-ный результат для класса 0, что означает, что она правильно идентифицировала почти все доброкачественные опухоли, в то время как для класса 1 результат составил 86%, что говорит о том, что некоторые злокачественные опухоли были неправильно классифицированы. Кроме того, показатель F1, который определяет точность и запоминаемость, составил 0,95 для класса 0 и 0,91 для класса 1, что подтверждает высокую классификационную эффективность модели. Усредненные по макросам и средневзвешенные показатели точности, запоминаемости и F1-показателя дополнительно подтверждают общую стабильность модели. Матрица путаницы содержит подробную разбивку результатов классификации. Из 71 случая доброкачественной опухоли модель MLP правильно идентифицировала 70 случаев, при этом один был ошибочно классифицирован как злокачественный. С другой стороны, из 43 случаев злокачественных опухолей модель правильно классифицировала 37 случаев, в то время как шесть были ошибочно классифицированы как доброкачественные. Эти результаты указывают на то, что модель MLP демонстрирует высокую эффективность классификации, особенно при выявлении доброкачественных опухолей, но демонстрирует несколько меньший уровень распознавания злокачественных опухолей.

Таблица 2 - Выходные данные MLP (многослойного персептрона)

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.92	0.99	0.95	71
1	0.97	0.86	0.91	43
Точноть			0.94	114
Макросредние показатели	0.95	0.92	0.93	114
Средневзвешенные показатели	0.94	0.94	0.94	114

Матрица путаницы:
[[70 1]
[6 37]]

Оценка эффективности классификатора с помощью метода опорных векторов (SVM) представлена в таблице 3. Общая точность модели составила 95%, что свидетельствует о высокой эффективности классификации при различении доброкачественных и злокачественных опухолей. Четкость для класса 0 (доброкачественные опухоли) составила 0,92, в то время как для класса 1 (злокачественные опухоли) она достигла 1,00, что означает, что модель идеально идентифицировала все злокачественные опухоли, которые она классифицировала как таковые. Для доброкачественных опухолей отзыв составил 100%, что указывает на то, что все 71 доброкачественный случай был правильно классифицирован, в то время как для злокачественных опухолей отзыв составил 86%, что означает, что некоторые злокачественные случаи были ошибочно классифицированы как доброкачественные. Показатель F1, который обеспечивает баланс точности и запоминания, составил 0,96 для класса 0 и 0,93 для класса 1, демонстрируя эффективность модели при классификации обоих типов опухолей. Усредненные по макросредне- и средневзвешенные баллы также подтверждают высокую эффективность классификатора в целом. Матрица путаницы дает более полное представление о результатах классификации. Модель SVM правильно классифицировала все 71 доброкачественную опухоль без ложноположительных результатов. Однако из 43 злокачественных опухолей 37 были правильно классифицированы, в то время как 6 были ошибочно классифицированы как доброкачественные. Эти результаты показывают, что модель SVM отлично справляется с выявлением доброкачественных опухолей, в то время как ее эффективность при выявлении злокачественных опухолей несколько ниже, что указывает на тенденцию ошибочно классифицировать некоторые злокачественные опухоли как доброкачественные.

Таблица 3 - Выходные данные метода опорных векторов (SVM)

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.92	1.00	0.96	71
1	1.00	0.86	0.93	43
Точноть			0.95	114
Макросредние показатели	0.96	0.93	0.94	114
Средневзвешенные показатели	0.95	0.95	0.95	114

Матрица путаницы:
[[71 0]
[6 37]]

Результаты модели логистической регрессии (LR) представлены в таблице 4. Общая точность модели составила 96%, что свидетельствует о высоком уровне надежности классификации доброкачественных и злокачественных опухолей. Четкость для класса 0 (доброкачественные опухоли) составила 0,95, в то время как для класса 1 (злокачественные опухоли) – 0,97, что свидетельствует о высокой способности правильно выявлять

положительные случаи. Для доброкачественных опухолей показатель запоминаемости составил 99%, что означает, что почти все доброкачественные случаи были правильно классифицированы, в то время как для злокачественных опухолей показатель отзыва составил 91%, что говорит о том, что небольшая часть злокачественных случаев была неправильно классифицирована. Показатель F1, который определяет точность и запоминаемость, составил 0,97 для класса 0 и 0,94 для класса 1, что подтверждает эффективность модели в различении доброкачественных и злокачественных опухолей. Кроме того, усредненные по макросреде и средневзвешенные показатели указывают на стабильно высокие показатели по всем показателям. Матрица путаницы содержит разбивку результатов классификации. Модель логистической регрессии правильно классифицировала 70 из 71 доброкачественных опухолей, при этом один ложноположительный результат был получен, когда доброкачественная опухоль была ошибочно классифицирована как злокачественная. Из 43 злокачественных опухолей модель правильно идентифицировала 39 случаев, в то время как четыре злокачественные опухоли были ошибочно классифицированы как доброкачественные. Эти результаты демонстрируют, что логистическая регрессия обеспечивает высокую степень точности при небольшом количестве ошибочных классификаций. Модель демонстрирует высокую точность для доброкачественных случаев, но несколько сниженную для злокачественных, что может стать областью для дальнейшего совершенствования.

Таблица 4 - Результаты логистической регрессии

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.95	0.99	0.97	71
1	0.97	0.91	0.94	43
Точноть			0.96	114
Макросредние показатели	0.96	0.95	0.95	114
Средневзвешенные показатели	0.96	0.96	0.96	114

Матрица путаницы:
[[70 1]
[4 39]]

Оценка эффективности классификатора K-ближайших соседей (K-NN) обобщена в таблице 5. Общая точность модели составила 96%, что свидетельствует о высокой прогностической эффективности при различении доброкачественных и злокачественных опухолей. Четкость для класса 0 (доброкачественные опухоли) составила 0,93, в то время как для класса 1 (злокачественные опухоли) она составила 1,00, что означает, что все опухоли, классифицированные как злокачественные действительно были злокачественными. Частота выявления доброкачественных опухолей составила 100%, что указывает на то, что все 71 случай доброкачественных опухолей был правильно классифицирован. Однако в случае злокачественных опухолей частота выявления составила 88%, что говорит о том, что небольшое количество случаев злокачественных опухолей было ошибочно классифицировано как доброкачественные. Показатель F1, который отражает баланс между точностью и запоминаемостью, составил 0,97 для класса 0 и 0,94 для класса 1, что подтверждает высокую классификационную способность модели. Усредненные по макросредне- и средневзвешенные баллы еще раз подтверждают эффективность классификатора в обоих классах. Матрица путаницы предоставляет подробную разбивку результатов классификации. Модель K-NN правильно классифицировала все 71 доброкачественную опухоль без ложноположительных результатов. Из 43 злокачественных опухолей 38 были правильно классифицированы, в то время как пять злокачественных опухолей были ошибочно классифицированы как доброкачественные. Эти данные указывают на то, что K-NN превосходно распознает доброкачественные опухоли, в то время как его распознавание злокачественных опухолей несколько ниже, что означает, что он ошибочно классифицирует некоторые злокачественные

опухоли как доброкачественные. Это может иметь значение для клинических применений, где высокая чувствительность к злокачественным опухолям имеет решающее значение.

Таблица 5 - Выходные данные К-ближайших соседей

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.93	1.00	0.97	71
1	1.00	0.88	0.94	43
Точноть			0.96	114
Макросредние показатели	0.97	0.94	0.95	114
Средневзвешенные показатели	0.96	0.96	0.96	114

Матрица путаницы:
[[71 0]
[5 38]]

Характеристики классификатора XGBoost представлены в таблице 6. Общая точность модели составила 96%, что свидетельствует о высоком уровне надежности классификации доброкачественных и злокачественных опухолей. Четкость для класса 0 (доброкачественные опухоли) составила 0,96, в то время как для класса 1 (злокачественные опухоли) она составила 0,95, что указывает на высокую способность правильно идентифицировать оба типа опухолей. Для доброкачественных опухолей отзыв составил 97%, что означает, что 69 из 71 доброкачественного случая были правильно классифицированы, в то время как для злокачественных опухолей отзыв составил 93%, что означает, что 40 из 43 злокачественных случаев были правильно идентифицированы. Показатель F1, который обеспечивает баланс между точностью и запоминанием, составил 0,97 для класса 0 и 0,94 для класса 1, что еще раз подтверждает эффективность модели. Усредненные по макросредам и взвешенные оценки были неизменно высокими, что подтверждает общую высокую классификационную способность модели. Матрица путаницы содержит подробную разбивку результатов классификации. Модель XGBoost правильно классифицировала 69 из 71 доброкачественных опухолей, при этом два ложноположительных результата были ошибочно классифицированы как злокачественные. Из 43 злокачественных опухолей 40 были правильно классифицированы, в то время как три злокачественных случая были ошибочно классифицированы как доброкачественные. Эти результаты указывают на то, что XGBoost обеспечивает высокую точность в обоих классах при небольшом количестве ошибочно классифицированных случаев.

Модель демонстрирует несколько более высокую чувствительность к доброкачественным опухолям, чем к злокачественным, что указывает на незначительную тенденцию к ошибочной классификации некоторых злокачественных опухолей как доброкачественных.

Таблица 6 - Выходные данные XGBoost

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.96	0.97	0.97	71
1	0.95	0.93	0.94	43
Точноть			0.96	114
Макросредние показатели	0.95	0.95	0.95	114
Средневзвешенные показатели	0.96	0.96	0.96	114

Матрица путаницы:
[[69 2]
[3 40]]

Оценка эффективности классификатора «случайный лес» представлена в таблице 7. Общая точность модели составила 96%, что свидетельствует о высокой прогностической способности при различении доброкачественных и злокачественных опухолей. Четкость для класса 0 (доброкачественные опухоли) составила 0,96, в то время как для класса 1 (злокачественные опухоли) она составила 0,98, что свидетельствует о высокой точности правильного выявления злокачественных новообразований. Частота выявления доброкачественных опухолей составила 99%, что означает, что 70 из 71 случая доброкачественных опухолей были правильно классифицированы, в то время как для злокачественных опухолей частота выявления составила 93%, что указывает на то, что 40 из 43 случаев злокачественных опухолей были правильно идентифицированы. Показатель F1, который отражает баланс между точностью и запоминаемостью, составил 0,97 для класса 0 и 0,95 для класса 1, что еще раз подтверждает надежность модели. Усредненные по макросредне- и средневзвешенные по весу баллы были неизменно высокими, что подтверждало эффективность сбалансированной классификации модели. Матрица путаницы позволяет получить более полное представление о результатах классификации модели. Случайный лес правильно идентифицировал 70 из 71 доброкачественной опухоли, при этом один ложноположительный результат был получен, когда доброкачественная опухоль была ошибочно классифицирована как злокачественная. Из 43 злокачественных опухолей 40 были правильно классифицированы, в то время как три злокачественных случая были ошибочно классифицированы как доброкачественные. Эти результаты показывают, что он обеспечивает высокую точность как для доброкачественных, так и для злокачественных опухолей с минимальными ошибками в классификации. Модель демонстрирует несколько более высокую вероятность обнаружения доброкачественных опухолей, чем злокачественных, аналогично другим проанализированным моделям классификации.

Таблица 7 - Выходные данные случайного леса

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.96	0.99	0.97	71
1	0.98	0.93	0.95	43
Точноть			0.96	114
Макросредние показатели	0.97	0.96	0.96	114
Средневзвешенные показатели	0.96	0.96	0.96	114

Матрица путаницы:
[[70 1]
[3 40]]

Характеристики классификатора LightGBM представлены в таблице 8. Общая точность модели составила 96%, что свидетельствует о ее высокой прогностической эффективности при классификации случаев рака молочной железы. Четкость для класса 0 (доброкачественные опухоли) составила 0,96, в то время как для класса 1 (злокачественные опухоли) она составила 0,98, что указывает на высокую специфичность при выявлении злокачественных новообразований. Частота выявления доброкачественных опухолей составила 99%, что означает, что 70 из 71 случая доброкачественных опухолей были правильно классифицированы, в то время как для злокачественных опухолей частота выявления составила 93%, что свидетельствует о том, что 40 из 43 случаев злокачественных опухолей были правильно выявлены. Показатель F1, измеряющий точность и запоминаемость, составил 0,97 для класса 0 и 0,95 для класса 1, что подтверждает надежную классификационную способность модели. Усредненные по макросредним и взвешенные оценки были неизменно высокими, что еще раз подтверждает ее эффективность. Матрица путаницы дает дополнительную информацию о предсказаниях модели. Классификатор LightGBM правильно классифицировал 70 из 71 доброкачественной опухоли, при этом один ложноположительный результат (доброкачественная опухоль была ошибочно классифицирована как злокачественная). Из 43

злокачественных опухолей 40 были идентифицированы правильно, в то время как три злокачественных случая были ошибочно классифицированы как доброкачественные. Эти результаты свидетельствуют о том, что LightGBM обеспечивает высокую точность классификации, аналогичную классификатору Random Forest. Он демонстрирует несколько более высокую вероятность обнаружения доброкачественных опухолей, чем злокачественных, что является тенденцией, наблюдаемой в нескольких проанализированных моделях классификации.

Таблица 8 - Выходные данные LightGBM

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.96	0.99	0.97	71
1	0.98	0.93	0.95	43
Точноть			0.96	114
Макросредние показатели	0.97	0.96	0.96	114
Средневзвешенные показатели	0.96	0.96	0.96	114

Матрица путаницы:
[[70 1]
[3 40]]

Результаты работы наивного байесовского классификатора (GaussianNB) приведены в таблице 9. Общая точность модели составила 97%, что делает ее одним из наиболее эффективных классификаторов для классификации рака молочной железы в этом исследовании. Четкость для класса 0 (доброкачественные опухоли) составила 0,96, в то время как для класса 1 (злокачественные опухоли) она составила 1,00, что указывает на то, что все опухоли, прогнозируемые как злокачественные, были действительно злокачественными. Запоминаемость по доброкачественным опухолям составила 100%, что означает, что все 71 доброкачественный случай был правильно классифицирован, в то время как по злокачественным опухолям повторность составила 93%, что свидетельствует о том, что 40 из 43 злокачественных случаев были правильно идентифицированы. Показатель F1, который обеспечивает баланс точности и запоминания, составил 0,98 для класса 0 и 0,96 для класса 1, что подтверждает высокую прогностическую способность модели. Усредненные по макросредним и взвешенные оценки были неизменно высокими, что подтверждает надежность классификации модели. Матрица путаницы содержит подробную разбивку результатов классификации. Наивный байесовский классификатор правильно идентифицировал все 71 доброкачественную опухоль без ложных срабатываний. Из 43 злокачественных опухолей 40 были идентифицированы правильно, в то время как три злокачественных случая были ошибочно классифицированы как доброкачественные. Эти результаты показывают, что наивный метод Байеса обеспечивает высокую точность и отличное распознавание доброкачественных опухолей без ложных срабатываний. Однако три случая злокачественных новообразований были ошибочно классифицированы как доброкачественные, что может стать критическим фактором при принятии клинических решений. В целом, модель демонстрирует высокую эффективность классификации, что делает ее приемлемым вариантом для выявления рака молочной железы.

Таблица 9 - Выходные данные наивного байесовского классификатора

	Четкость	Запоминаемость	F1-показатель	Всего
0	0.96	1.00	0.98	71
1	1.00	0.93	0.96	43
Точноть			0.97	114
Макросредние показатели	0.98	0.97	0.97	114
Средневзвешенные показатели	0.97	0.97	0.97	114

Матрица путаницы:
[[71 0]
[3 40]]

Чтобы обеспечить всестороннее сравнение классификационных моделей, были проанализированы различные показатели эффективности, включая точность, четкость, запоминаемость и показатель F1. Общее сравнение моделей по этим показателям показано на рисунке 1, что дает четкое визуальное представление об их эффективности. Эта визуализация облегчает детальный анализ сильных и слабых сторон каждой модели, помогая определить наиболее подходящий подход к классификации рака молочной железы.

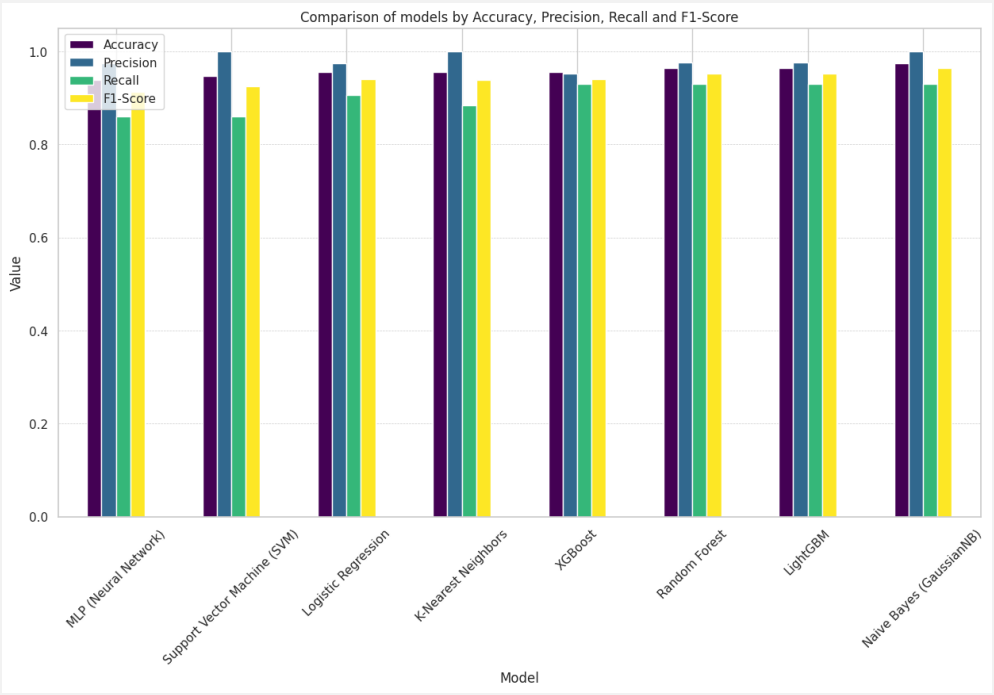

Рисунок 1. Сравнение моделей по точности, запоминаемости и F1-показателю

На рисунке 2 представлены отдельные сравнения моделей, основанные на конкретных показателях эффективности.

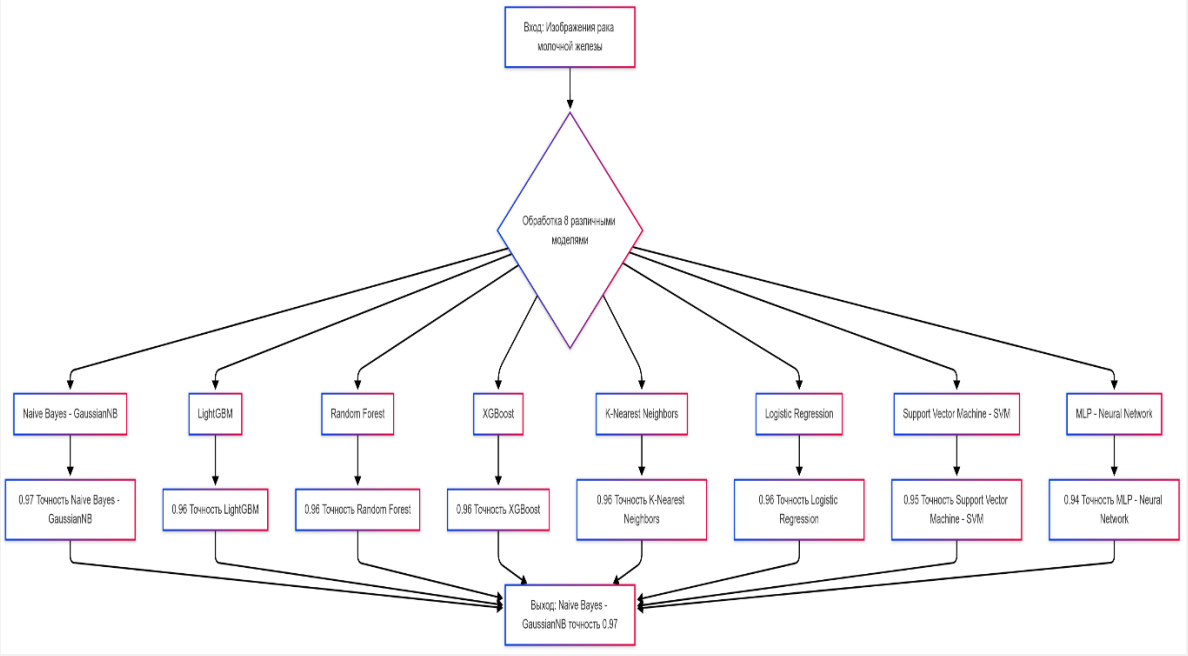

Рисунок 2. Методы классификации для диагностики рака молочной железы

ЗАКЛЮЧЕНИЕ

В данном исследовании оценивалось несколько алгоритмов машинного обучения для классификации рака молочной железы с использованием набора данных о диагностическом раке молочной железы штата Висконсин. Проанализированные модели включали логистическую регрессию, многослойный персептрон (MLP), метод опорных векторов (SVM), K-ближайших соседей (KNN), XGBoost, случайный лес, LightGBM и наивный Байесовский классификатор, эффективность которых оценивалась на основе точности, четкости, запоминаемости и показателя F1. Результаты показали, что наивная байесовская модель достигла наивысшей точности (97%), превзойдя другие модели по общей эффективности классификации. Кроме того, Random Forest, LightGBM и XGBoost продемонстрировали сильные прогностические возможности, особенно в отношении баланса точности и запоминания. Традиционные модели машинного обучения, такие как логистическая регрессия и SVM, также показали хорошие результаты, достигнув точности более 95%, что свидетельствует об их надежности в классификации рака молочной железы. Сравнительный анализ этих моделей дает ценную информацию для выбора подходящего метода классификации в медицинской диагностике. Хотя высокая точность имеет решающее значение, необходимо также учитывать такие факторы, как эффективность вычислений, интерпретируемость и применимость в реальной клинической практике. В будущей работе можно было бы изучить гибридные подходы, архитектуры глубокого обучения или методы отбора признаков для дальнейшего повышения эффективности классификации. В конечном счете, это исследование подчеркивает потенциал моделей машинного обучения в улучшении раннего выявления рака молочной железы, способствуя более точной и своевременной диагностике, что имеет решающее значение для эффективного лечения и результатов лечения пациентов.

ЛИТЕРАТУРА

1 Ara S., Das A., & Dey A. (2021, April). Malignant and benign breast cancer classification using machine learning algorithms. In *2021 International Conference on Artificial Intelligence (ICAI)* (pp. 97-101). IEEE.
2 Batool A., & Byun Y.C. (2023, August). Breast cancer classification using random forest algorithm. In *Journal of Physics: Conference Series* (Vol. 2559, No. 1, p. 012002). IOP Publishing.
3 Chen H., Wang N., Du X., Mei K., Zhou Y., & Cai G. (2023). Classification prediction of breast cancer based on machine learning. *Computational intelligence and neuroscience*, *2023*(1), 6530719.
4 Desai M., & Shah M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, *4*, 1-11.
5 Desiani A., Lestari A. A., Al-Ariq M., Amran A., & Andriani, Y. (2022). Comparison of support vector machine and k-nearest neighbors in breast cancer classification. *Pattimura International Journal of Mathematics (PIJMath)*, *1*(1), 33-42.
6 Jamil R., Dong M., Rashid J., Orken M., Pernebaykyzy Z. S., & Ragytovna M. K. (2024). High Accuracy Microcalcifications Detection of Breast Cancer Using Wiener LTI Tophat Model. *IEEE Access*.
7 Kanber B. M., Al Smadi A., Noaman N. F., Liu B., Gou S., & Alsmadi M. K. (2024). Lightgbm: A leading force in breast cancer diagnosis through machine learning and image processing. *IEEE Access*, *12*, 39811-39832.
8 Lim T. S., Tay K. G., Huong A., & Lim X. Y. (2021). Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. *Int. J. Electr. Comput. Eng.(IJECE)*, *11*(4), 3059.
9 Liew X. Y., Hameed N., & Clos J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, *6*, 100154.
10 Minnoor M., & Baths V. (2023). Diagnosis of breast cancer using random forests. *Procedia Computer Science*, *218*, 429-437.
11 Orazayeva A., Tussupov J., Shangytbayeva G., Galymova A., Zhunissova U., Tergeussizova A., ... & Kenzhebayeva Z. (2024). Effective detection of breast pathology using machine learning methods. *International Journal of Electrical & Computer Engineering (2088-8708)*, *14*(5).
12 Strelcenia E., & Prakoonwit S. (2023). Effective feature engineering and classification of breast cancer diagnosis: a comparative study. *BioMedInformatics*, *3*(3), 616-631.
13 World Health Organization. Data. Health data overview for the Republic of Kazakhstan. https://data.who.int/countries/398

REFERENCES

1 Ara S., Das A., & Dey A. (2021, April). Malignant and benign breast cancer classification using machine learning algorithms. In *2021 International Conference on Artificial Intelligence (ICAI)* (pp. 97-101). IEEE.
2 Batool A., & Byun Y. C. (2023, August). Breast cancer classification using random forest algorithm. In *Journal of Physics: Conference Series* (Vol. 2559, No. 1, p. 012002). IOP Publishing.
3 Chen H., Wang N., Du X., Mei K., Zhou Y., & Cai G. (2023). Classification prediction of breast cancer based on machine learning. *Computational intelligence and neuroscience*, *2023*(1), 6530719.
4 Desai M., & Shah M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth*, *4*, 1-11.
5 Desiani A., Lestari A. A., Al-Ariq M., Amran, A., & Andriani, Y. (2022). Comparison of support vector machine and k-nearest neighbors in breast cancer classification. *Pattimura International Journal of Mathematics (PIJMath)*, *1*(1), 33-42.
6 Jamil R., Dong M., Rashid J., Orken M., Pernebaykyzy Z. S., & Ragytovna M. K. (2024). High Accuracy Microcalcifications Detection of Breast Cancer Using Wiener LTI Tophat Model. *IEEE Access*.
7 Kanber B. M., Al Smadi A., Noaman N. F., Liu B., Gou S., & Alsmadi M. K. (2024). Lightgbm: A leading force in breast cancer diagnosis through machine learning and image processing. *IEEE Access*, *12*, 39811-39832.
8 Lim T. S., Tay K. G., Huong A., & Lim X. Y. (2021). Breast cancer diagnosis system using hybrid support vector machine-artificial neural network. *Int. J. Electr. Comput. Eng.(IJECE)*, *11*(4), 3059.
9 Liew X. Y., Hameed N., & Clos J. (2021). An investigation of XGBoost-based algorithm for breast cancer classification. *Machine Learning with Applications*, *6*, 100154.
10 Minnoor M., & Baths V. (2023). Diagnosis of breast cancer using random forests. *Procedia Computer Science*, *218*, 429-437.
11 Orazayeva A., Tussupov J., Shangytbayeva G., Galymova A., Zhunissova U., Tergeussizova A., ... & Kenzhebayeva Z. (2024). Effective detection of breast pathology using machine learning methods. *International Journal of Electrical & Computer Engineering (2088-8708)*, *14*(5).
12 Strelcenia E., & Prakoonwit S. (2023). Effective feature engineering and classification of breast cancer diagnosis: a comparative study. *BioMedInformatics*, *3*(3), 616-631.
13 World Health Organization. Data. Health data overview for the Republic of Kazakhstan. https://data.who.int/countries/398