DOI: 10.15514/ISPRAS–2021–33(2)–5

Новая интеллектуальная система для обнаружения сахарного диабета 2-го типа с модифицированной функцией потерь и регуляризацией

[1] *М. Г.Ч. <Mallikagc9@gmail.com>*
[1,2,3,4] *А. Альсадун, ORCID: 0000-0002-2309-3540 <AAlsadoon@studygroup.com>*
[5] *Д.Т.Х. Фам, ORCID: 0000-0002-0813-827X <hangpdt@ued.udn.vn>*
[6] *С.Х. Абдулла, ORCID: 0000-0001-7972-210X <120015@uotechnology.edu.iq>*
[5] *Х.Т. Май, ORCID: 0000-0002-0813-827X <mhthi@ued.udn.vn>*
[1] *П.В.Ч. Прасад, ORCID: 0000-0002-3007-687X <CWithana@studygroup.com>*
[5] *Ч.К.В. Нгуен, ORCID: 0000-0003-2281-0429 <ntquocvinh@ued.udn.vn>*
[1] *Университет Чарльза Стерта,*
Австралия, NSW 2650, Вагга-Вагга, Бурома
[2] *Университет Западного Сиднея,*
Австралия, NSW 2000, Сидней, ул. Элизабет, 255
[3] *Университет Саутерн Кросс,*
Австралия, NSW 2010, Сидней, Сарри Хиллс, ул. Мэри, 84-86
[4] *Азиатско-Тихоокеанский международный колледж,*
Австралия, NSW 2150, Сидней, Парраматта, ул. Фицвильям, 1-3
[5] *Данангский университет – Университет науки и образования,*
Вьетнам, 550000, Дананг, ул. Тон Дык Тханг, 459
[6] *Иракский технический университет,*
Ирак, 10066, Багдад, ул. Аль Синаа

Аннотация. Сахарный диабет 2-го типа (СД2) составляет около 90% случаев диабета, и одним из ключевых аспектов СД2 являются жесткие требования к постоянному мониторингу и выявлению. Это исследование направлено на разработку ансамбля из нескольких моделей машинного и глубокого обучения для раннего обнаружения СД2 с высокой точностью. При большом разнообразии моделей ансамбль обеспечивает больше возможностей, чем отдельные модели. Предлагаемый ансамбль моделей основан на использовании известных моделей логистической регрессии, случайного леса, опорных векторов и глубокой нейронной сети. Выходные данные каждой модели в модифицированном ансамбле используются для определения окончательных выходных данных системы. Датасеты, используемые для этих моделей, включают Practice Fusion HER, Pima Indians diabetic's data, UCI AIM94 Dataset и CA Diabetes Prevalence 2014. По сравнению с ранее разработанными решениями, наше решение на основе ансамблевой модели демонстрирует высокие показатели точности, чувствительности и специфичности. В среднем обеспечиваются точность 87,5% от 83,51%, чувствительность 35,8% от 29,59% и специфичность 98,9% от 96,27%. Время работы предлагаемого решения составляет 96,6 мс, в то время как у наиболее по архитектуре известной системы – 97,5 мс. Предлагаемая усовершенствованная система улучшает возможности прогнозирования СД2 на основе использования ансамбля из нескольких моделей машинного и глубокого обучения. Для получения окончательного точного прогноза с использованием результатов отдельных моделей применяется схема мажоритарного голосования. В работе также изменена функция регуляризации, чтобы учесть регуляризацию всех

моделей в ансамбле, что помогает предотвратить переобучение и поддержать возможность обобщений в предлагаемой системе.

Ключевые слова: прогнозирование диабета 2-го типа; машинное обучение; ансамбль моделей; глубокие нейронные сети; метод опорных векторов; логистическая регрессия; случайный лес

Для цитирования: Г.Ч. М., Альсадун А., Фам Т.Х.Ф., Абдулла С.Х., Май Х.Е., Прасад П.В.Ч., Нгуен Ч.К.В. Новая интеллектуальная система для обнаружения сахарного диабета 2-го типа с модифицированной функцией потерь и регуляризацией. Труды ИСП РАН, том 33, вып. 2, 2021 г., стр. 93-114. DOI: 10.15514/ISPRAS–2021–33(2)–5

Благодарности. Это исследование частично поддержано Университетом Дананга – Университетом науки и образования, Вьетнам, в рамках гранта «T2020-TD-03-BS».

A Novel Intelligent System for Detection of Type 2 Diabetes with Modified Loss Function and Regularization

[1] *M. G.C. <Mallikagc9@gmail.com>*
[1,2,3,4] *A. Alsadoon, ORCID: 0000-0002-2309-3540 <AAlsadoon@studygroup.com>*
[5] *D.T.H. Pham, ORCID: 0000-0002-0813-827X <hangpdt@ued.udn.vn>*
[6] *S. Abdullah, ORCID: 0000-0001-7972-210X <120015@uotechnology.edu.iq>*
[5] *H.T. Mai, ORCID: 0000-0002-0813-827X <mhthi@ued.udn.vn>*
[1] *P.W.C. Prasad, ORCID: 0000-0002-3007-687X <CWithana@studygroup.com>*
[5] *T.Q.V. Nguen, ORCID: 0000-0003-2281-0429 <ntquocvinh@ued.udn.vn>*
[1] *Charles Sturt University,*
Boorooma Street, Wagga Wagga, NSW 2650, Australia
[2] *University of Western Sydney,*
255 Elizabeth St, Sydney, NSW 2000, Australia
[3] *Southern Cross University*
84-86 Mary St., Surry Hills, Sydney, NSW 2010, Australia,
[4] *Asia Pacific International College,*
1-3 Fitzwilliam Street*, Parramatta* NSW *2150*
[5] *The University of Da Nang – University of Science and Education,*
459, Ton Duc Thang St., Danang City, 550000, Vietnam
[6] *University of Technology, Iraq,*
Al Sinaa Street, Baghdad, 10066, Iraq

Abstract. Type 2 Diabetes (T2DM) makes up about 90% of diabetes cases, as well as tough restriction on continuous monitoring and detecting become one of key aspects in T2DM. This research aims to develop an ensemble of several machine learning and deep learning models for early detection of T2DM with high accuracy. With high diversity of models, the ensemble will provide more excessive performance than single models. *Methodology:* The proposed system is modified enhanced ensemble of machine learning models for T2DM prediction. It is composed of Logistic Regression, Random Forest, SVM and Deep Neural Network models to generate a modified ensemble model. *Results:* The output of each model in the modified ensemble is used to figure out the final output of the system. The datasets being used for these models include Practice Fusion HER, Pima Indians diabetic's data, UCI AIM94 Dataset and CA Diabetes Prevalence 2014. In comparison to the previous solutions, the proposed ensemble model solution exposes the effectiveness of accuracy, sensitivity, and specificity. It provides an accuracy of 87.5% from 83.51% in average, sensitivity of 35.8% from 29.59% as well as specificity of 98.9% from 96.27%. The processing time of the proposed model solution with 96.6ms is faster than the state-of-the-art with 97.5ms. *Conclusion:* The proposed modified enhanced system in this work improves the overall prediction capability of T2DM using an ensemble of several machine learning and deep learning models. A majority voting scheme utilizes the output from several models to make the final accurate prediction. Regularization function in this work is modified in order to include the regularization of all the models in ensemble, that helps prevent the overfitting and encourages the generalization capacity of the proposed system.

Keywords: T2DM Prediction, Machine Learning, Ensemble, Deep Neural Networks, SVM, Logistic Regression, Random Forests

For citation: G.C. M., Alsadoon A., Pham D.T.H., Abdullah S., Mai H.T, Prasad P.W.C., Nguen T.Q.V. A Novel Intelligent System for Detection of Type 2 Diabetes with Modified Loss Function and Regularization. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 2, 2021, pp. 93-114 (in Russian). DOI: 10.15514/ISPRAS–2021–33(2)–5.

Acknowledgement. This research is partially supported by The University of Da Nang – University of Science and Education, Vietnam under grant “T2020-TD-03-BS”.

1. Введение

Модели машинного обучения на основе случайного леса (Random Forest) и опорных векторов (Support Vector Machine, SVM) нашли широкое применение в медицинских исследованиях, в то время как нейронные сети также являются очень мощными моделями, способными изучать сильные нелинейные отношения в данных. Эти модели могут быть объединены в ансамбль, в котором применяется каждый из алгоритмов для построения окончательного результата. Предоставление большой размерности и сложной взаимосвязи между признаками позволяет получить лучшие результаты, чем отдельная модель. Эта работа направлена на использование ансамбля моделей с использованием не только одного алгоритма, но и ансамбля алгоритмов для достижения улучшенных характеристик системы [9].

Среди нескольких работ, использующих машинное обучение для прогнозирования сахарного диабета, лучшая эффективность была продемонстрирована при применении подхода с использованием ансамбля различных методов [5]. В этой работе использовался ансамбль глубокой и широкой нейронной сети. Широкая часть сети обрабатывает статические признаки, в глубокая сеть – регулируемые признаки. Демонстрируется, что в ансамбле можно успешно использовать разные модели, способные обрабатывать различные типы признаков. Основным ограничением использования единого алгоритма является то, что модели стремятся к получению очень похожих решений [11]. Когда используются совсем разные модели, их разнообразные решения могут дать лучшие результаты.

В этой статье мы предлагаем систему для прогнозирования СД2, опирающуюся на преимущества разнообразия нескольких моделей машинного обучения, включая случайный лес, логистическую регрессию, опорные вектора и глубокую нейронную сеть. Для всех моделей использовалась регуляризация L2. Для получения окончательного результата прогноза использовался мажоритарный алгоритм голосования Бойера-Мура (Boyer–Moore majority vote algorithm). Эксперименты проводились на 4 датасетах – Practice Fusion EHR Dataset, Pima Indians Diabetics Dataset, UCI AIM94 Dataset и CA Diabetes Prevalence 2014. Средняя точность классификации составила 87,5% (83,21% минимум и 92,98% максимум), а среднее время обработки – 96,6 мс. Эти показатели лучше тех, которые демонстрируют аналогичные известные нам системы.

2. Обзор литературы

В этом разделе представлен обзор различных методов и подходов, применяемых в этой области. Нгуен (Binh P. Nguyen) и др. [5] разработали систему для диагностики СД2, основанную на глубоком и широком обучении с использованием электронной истории болезни людей. Глубокая и широкая модель – это гибридная модель линейной и глубокой нейронной сети, которой свойственны преимущества как линейных моделей, так и нейронных сетей. Авторы достигают современных результатов с точностью 84,28% и AUC 84,13%, что значительно выше, чем у других моделей. Несмотря на то, что модель очень

хорошо работает с ансамблем данных, ограничение этой работы состоит в том, что обычно неясно, какие функции подходят для глубокой и широкой части модели.

В аналогичной работе [12] авторы разработали модель, которая могла бы использовать данные о пациентах с состоянием, похожим на состояние пациента, данные о котором отсутствуют. Этот остроумный метод помогает в практических ситуациях, когда данные не всегда полностью и/или постоянно доступны. Неполные данные и периоды отсутствия данных ежедневного измерения уровня глюкозы в крови приводят к ограничению точности. Авторы предложили метод прогнозирования HbA1c, в котором сверточная нейронная сеть (Convolutional Neural Network, CNN) используется для извлечения признаков из данных временных рядов; затем эти признаки объединяются со статическими признаками пациентов, такими как возраст, пол и т.д., и все это передается в полносвязную нейронную сеть. Авторы сообщают об улучшениях по сравнению с обычными моделями CNN со средней абсолютной ошибкой 4,8. Однако основным ограничением этой работы является то, что ансамбль данных, возможно, очень мал для использования больших глубинных моделей.

Диабетическая ретинопатия – одна из ведущих причин слепоты в мире. Раннее выявление может обеспечить лечение и предотвратить слепоту, но очень сложно диагностировать заболевание на ранних стадиях. Авторы [13] реализовали сиамскую модель CNN, каждая из частей которой обрабатывает данные сканирования левого/правого глаза, чтобы обеспечить предсказания для левого и правого глаза индивидуально. Сиамская сеть обеспечивает признаки для обоих глаз, которые затем передаются в полноценную CNN. Это решение предоставило новый подход, который был побужден тем, как медицинский персонал проводит диагностику ретинопатии у диабетиков с помощью сканирований глаз. Сообщается о получении впечатляющего показателя AUC в 0,95. Однако ограничение предлагаемой модели состоит в том, что она использует очень крупную структуру модели (Inception v3) для каждого глаза, что значительно снижает скорость обучения модели и усложняет обучение.

Контроль уровня глюкозы в крови необходим для лечения диабета. В статье [4] авторы предложили модель, в которой CNN получает многомерные входные данные и преобразует их в данные временных рядов, полезные для прогнозирования на основе рекуррентных нейронных сетей (Recurrent Neural Network, RNN). Это является новым способом совместного использования статических и динамических признаков. Используемые входные данные представляют собой временные ряды уровней глюкозы каждые 30 минут. В статье представлены значения средней квадратичной ошибки (Root Mean Square Error, RMS Error, RMSE) и среднего абсолютного относительного отклонения (Mean Absolute Relative Difference, MARD), которые являются более хорошими, чем у базовых моделей. Однако для модели требуется большой объем данных через равные промежутки времени, что трудно получить в реальном мире.

Авторы работы [2] занимаются прогнозированием послеродового гестационного сахарного диабета (ГСД) с помощью тестов GCT и OGTT. В этой работе использовался алгоритм XGBoost, который представляет собой усиленную модель случайного леса. Авторы сообщают о точности 91% при специфичности и чувствительности 74%. Результаты были выдающимися, потому что данные подходили для этого алгоритма, а сам алгоритм XGBoost прекрасно с ними справлялся и производил результаты на уровне последних мировых достижений.

С увеличением доступности носимых устройств стал значительно более доступным непрерывный мониторинг физиологических и поведенческих особенностей людей. Работа [8] интересна еще и тем, что в модели использовались данные временных рядов из непрерывных измерений глюкозы и активности, а затем они объединялись с независящими от времени демографическими данными аналогично тому, как это делается в передовой системе [5]. Однако в этой работе использовалась долгая краткосрочная память (Long Short-Term Memory; LSTM), обеспечивающая хорошую производительность при работе с данными временных рядов и значительную общую точность. Это решение отличается еще и тем, что в

нем объединяются глубокие и широкие аспекты машинного обучения для использования как временных рядов, так и статических данных.

Авторы работы [3] применили на практике линейный дискриминантный анализ (Linear Discriminant Analysis, LDA), чтобы определить, ел ли пациент с СД1. Для этого используются данные автоматического мониторинга уровня глюкозы. Однако LDA – это относительно простая и линейная модель машинного обучения [10]. Важными особенностями LDA являются линейность метода и потребность в выборе правильного порогового значения классификатора. Для определения правильного порога в [3] использовалось много специфических знаний предметной области.

В работе [14] нечеткие правила выводятся из нечетких деревьев решений, построенных с использованием алгоритма муравейника. Классификатор достиг точности 87,7% и чувствительности 92,2% на Pima Indian Diabetes.

Благодаря использованию ансамбля моделей в [5] были получены многие результаты современного мирового уровня. В своей работе авторы использовали десять различных глубоких и широких моделей и усредняли результаты каждой модели на десяти прогонах для получения окончательного результата. Значительно помогает то, что каждая модель обучается независимо от других моделей, а окончательная ансамблевая модель обобщает знания отдельных моделей.

В работе [15] авторы также предпочли использовать ансамблевый метод. Они занялись проблемой раннего выявления СД2 и гипертонии. Их цель заключалась в развитии модели прогнозирования заболеваний для обеспечения прогнозирования СД2 и гипертонии на основе данных о личных факторах риска. Авторы использовали комбинацию мощных методов и сообщают о впечатляющих результатах. Кроме того, их решение также показало, как можно эффективно обучать ансамбли с помощью проверки K-Cross.

В аналогичной работе [6] авторы продемонстрировали эффективную предварительную обработку данных наряду с использованием ансамбля сильных моделей, который обеспечивает результаты, лучшие, чем у отдельных моделей. Они указали, что использование в моделях прогнозирования СД2 нерегулярно дискретизированных данных приводит к снижению производительности по причинам, связанным с аппроксимацией. В реальном мире для предсказания СД2 очень сложно получить периодические данные для каждого пациента [1].

2.1 Наиболее совершенная система

В работе [5] представлен расширенный ансамбль глубоких и широких нейронных сетей для обнаружения ранних симптомов T2DM. Авторы используют данные электронных медицинских карт (ЭМК) пациентов и применяют современную глубокую и широкую архитектуру для повышения производительности.

На блок-схеме на рис. 1 представлены функции (в черных прямоугольниках) и ограничение (красный прямоугольник) системы прогнозирования СД2 [5], а также разработанная авторами гибридная модель линейной и глубокой нейронной сети (глубокой и широкой нейронной сети). Авторы утверждают, что глубокие и широкие нейронные сети подходят для использования фиксированных и настраиваемых признаков, которые распространены в датасетах о здоровье. Для обучения ансамбля моделей используется перекрестная проверка. Это решение показывает, как модели глубокого обучения могут использоваться для прогнозирования состояния здоровья и как эти модели могут применяться в ансамбле глубоких и широких сетей для получения самых современных результатов для прогнозирования СД2. Представленная модель дает точность 83% в среднем для 10 моделей и 84,28% для ансамбля из 10 моделей. Работа системы состоит из четырех этапов:

предварительная обработка, извлечение признаков, выбор и обучение моделей-членов ансамбля.

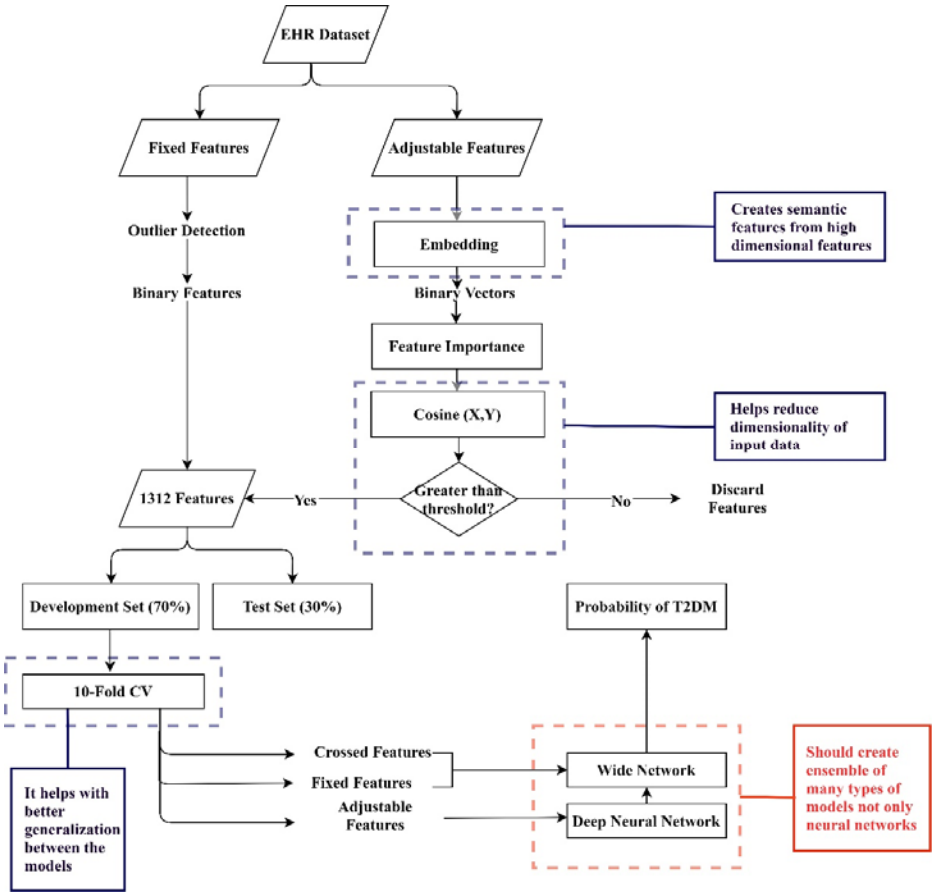

Рис. 1. Блок-схема системы [5]
Fig. 1. Block diagram of the state-of-the-art system [5]

Этап предварительной обработки: предварительная обработка данных, полученных из ЭМК пациентов. Датасет разделяется на фиксированные и настраиваемые признаки. К фиксированным признакам относятся статические характеристики пациентов – индекс массы тела (ИМТ), пол и т. д. Настраиваемые признаки собираются с датчиков и образцов. Фиксированные функции обрабатываются путем обнаружения и удаления выбросов. После этого они, где возможно, преобразуются в двоичные признаки.

Этап извлечения признаков: на этом этапе настраиваемые признаки проходят через слой вложений для построения семантических векторов. Затем их важность оценивается путем ранжирования в порядке косинусного расстояния от меток. Выбираются только наиболее важные признаки, отделяемые пороговыми значениями.

Этап выбора элементов ансамбля: на этом этапе датасет делится на поднаборы данных для разработки и тестирования сети с разделением 70/30. Настраиваемые признаки вводятся в нейронную сеть, состоящую из нескольких полностью связанных слоев с функциями активации ReLU. На заключительном выходном слое объединяется фиксированные и

перекрестные признаки. Объединенные признаки передаются на уровень логистической регрессии для получения окончательного результата в виде распределения вероятностей.

Этап обучения: на этом этапе строится ансамбль моделей. 10-кратная перекрестная проверка используется для создания 10 глубоких и широких моделей. Результаты этих моделей используются в схеме голосования для получения окончательного результата. Результат, который создается большинством моделей, выбирается как окончательный результат ансамбля.

Функция потерь, используемая в данной модели, представляет собой стандартную функцию кросс-энтропии (1):

$$\boldsymbol{L} = \sum_{\boldsymbol{N}} \left(-\boldsymbol{y}\boldsymbol{log}(\boldsymbol{p})\right), \quad (\mathbf{1})$$

где $\boldsymbol{L}$ – функция потерь модели, $\boldsymbol{N}$ – количество образцов, $\boldsymbol{y}$ – истинная вероятность, $\boldsymbol{p}$ – прогнозируемая вероятность

Потеря регуляризации состоит в снижении веса, полученного с помощью параметра регуляризации, который является гиперпараметром модели. Это предотвращает переобучение моделей и способствует возможности обобщения моделей за счет ограничения их весов, как показано в формуле (2).

$$\boldsymbol{L_R} = \boldsymbol{\lambda} \times \sum_{\boldsymbol{n}} |\boldsymbol{W}|^{\boldsymbol{2}}, \quad (\mathbf{2})$$

где $\boldsymbol{L_R}$ – потеря регуляризации, $\boldsymbol{\lambda}$ – параметр регуляризации, $\boldsymbol{n}$ – количество итераций, $\boldsymbol{W}$ – вес модели.

Табл. 1. Ансамбль моделей машинного обучения для прогнозирования начала СД2
Table1: Ensemble of machine learning models to predict onset of T2DM

Алгоритм: ансамбль моделей машинного обучения для прогнозирования начала СД2 Входные данные: 2D-матрица точек данных, где каждый столбец – это признак, а каждая строка – пациент Выходные данные: вероятность обнаружения СД2 в диапазоне от 0,0 до 1,0
BEGIN Шаг 1. Обнаружение выбросов: удалить данные для пациентов выше 95-го процентиля. Шаг 2. Баланс классов: определить вес каждого класса w_i = количество образцов класса / общее количество образцов. Шаг 3. Нормализация: нормализовать каждый признак, чтобы он попал в диапазон (0,1) Шаг 4. Обучить ансамбля моделей, минимизируя функцию потерь (1) и потери регуляризации (2). Шаг 5. Использовать схему мажоритарного голосования для получения окончательного результата на основе результатов отдельных моделей. END

3. Предлагаемая система

В нашем исследовании были проанализированы многие методы прогнозирования СД2. Основные проблемы, которые выделяются в статьях о прогнозировании СД2, – это предварительная обработка данных, выбор признаков и выбор модели.

Среди всех работ лучшим решением можно считать [5]. В ансамблевом методе для прогнозирования используется не одна модель, а коллекция моделей. Для учета всех различных получаемых результатов используется механизм голосования или усреднения. Достижению удовлетворительной производительности в целом помогает то, что разные алгоритмы подходит для выявления разных факторов и закономерностей в данных. Как сообщается в [5], эта система обеспечивают наилучшие точность и AUC ROC при решении задачи прогнозирования СД2.

Использованию ансамблевой модели для выявления сахарного диабета посвящена и впечатляющая работа [8], которую можно считать развитием [5] из-за разнообразия используемых моделей. Однако по своим результатам [8] не опережает [5]. Одна из основных

причин заключается в том, что авторы [8] не регуляризируют свою модель с помощью таких методов, как регуляризация L1/L2 и перекрестная проверка. В нашей работе эти ограничения преодолеваются путем применения соответствующих методов регуляризации.

Предлагаемое нами решение состоит из четырех основных компонентов (рис. 2).

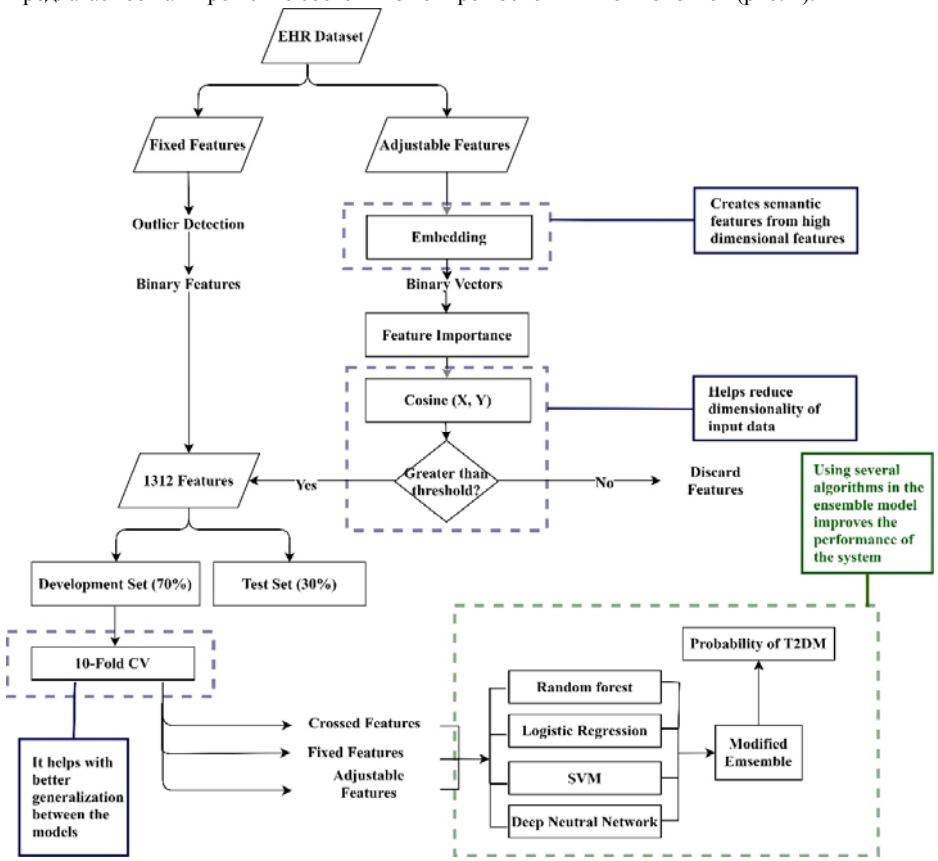

Рис. 2. Блок-схема предлагаемой системы
Зеленая рамка обозначает новые части в предлагаемой нами системе
Fig. 2. Block diagram of the purposed system
The green border refers to the new parts in our purposed system

Этап предварительной обработки: данные получаются из записей ЭМК пациентов, опубликованных компанией Practice Fusion. Набор данных разбит на фиксированные и настраиваемые признаки, как в работе [5]. Фиксированные признаки включают статические характеристики пациента, такие как ИМТ, пол и т.д. Регулируемые признаки – это признаки, собираемые с датчиков и из образцов. Они обрабатываются путем обнаружения и удаления выбросов, а затем преобразуются в двоичные признаки, где это возможно.

Этап извлечения признаков: настраиваемые объекты проходят через слой встраивания для построения сематических векторов. Важность этих призаков оценивается путем ранжирования их в порядке косинусного расстояния от меток аналогично работе [5]. Выбираются только наиболее важные характеристики, отделенные пороговым значением. Однако в предлагаемом решении важность функции рассчитывается с использованием

алгоритма XGBoost, который опирается на использование коэффициента Джини расслоения деревьев решений в модели.

Этап выбора элементов ансамбля: датасет разделяется на наборы для разработки и тестирования с процентным соотношением 70/30, настраиваемые признаки передаются в нейронную сеть, состоящую из нескольких полностью связанных слоев с функциями активации ReLU. В заключительном выходном слое фиксированные признаки объединяются с перекрестными признаками. Затем объединенные признаки загружаются в слой логистической регрессии для получения окончательного результата в виде распределения вероятностей. Однако в предлагаемом решении все функции признаки во все модели ансамбле, который теперь включает не только глубокий и широкий компоненты.

Этап обучения: в работе [5] для построения ансамбля моделей используется 10-кратная перекрестная проверка для создания 10 глубоких и широких моделей. В предлагаемом нами решении с использованием 10-кратной перекрестной проверки создаются *N* моделей, и все они обучаются с использованием всех частей датасета. Затем результаты этих моделей используются в схеме голосования для получения окончательного результата – результата, полученного большинством моделей.

3.1 Предлагаемые формулы

Предлагаемые формулы ориентированы на ансамбль различных моделей машинного обучения, а не только на один тип модели [5]. Это увеличивает разнообразие решений, достигаемых каждой из разных моделей, повышая качество результатов ансамблевой модели.

Формула (3) представляет собой функцию ансамблевых потерь. Она объединяет потерю регуляризации и потерю отдельной модели и является эффективной функцией потерь ансамбля. Это улучшает формулу (1): комбинирование потерь регуляризации отдельных моделей способствует оптимизации ансамбля. Наличие L_M уменьшает обобщения отдельных моделей:

$$L_E = \frac{1}{M}\sum_M (L_R + L_M), \tag{3}$$

где L_E – потеря ансамбля, M – количество моделей, L_R – потеря регуляризации каждой модели, L_M – потеря каждой модели.

Следующая формула (4) представляет собой функцию потери регуляризации всего ансамбля. Она помогает повысить степень обобщения моделей на протяжении всего процесса оптимизации:

$$ML_E = \frac{1}{M}\sum_M L_R, \tag{4}$$

где ML_E – измененная потеря регуляризации ансамбля, M – количество моделей, L_R – потеря регуляризации каждой модели.

Потеря модели ансамбля представляет собой функцию кросс-энтропии, которая эквивалентна сумме энтропии набора данных и расстояния Ку́льбака-Ле́йблера (Kullback-Leibler divergence) между истинными метками и прогнозами модели, как показывает формула (5):

$$ML = ML_E * \sum_N (-y \log(p)), \tag{5}$$

где ML – модифицированная функция потерь модели, ML_E – измененная потеря регуляризации ансамбля, N – количество образцов, y – истинная вероятность, p – прогнозируемая вероятность.

Следующая формула (6) заменяет потерю регуляризации $L1$ на потерю регуляризации $L2$, которая, как было эмпирически показано, обеспечивает большее обобщение:

$$R = \frac{\lambda}{2n}\sum_n |W|, \tag{6}$$

где R – регуляризация, n – количество итераций, W – веса нейронов, λ – параметр регуляризации, $\frac{\lambda}{2n}$ – масштабный коэффициент.

Потеря регуляризации масштабируется по следующей формуле. Параметр регуляризации – это гиперпараметр, который оптимизируется во время обучения, как показано в (7).

$$MR = \frac{\lambda}{2n}, \tag{7}$$

где λ – параметр регуляризации, $\frac{\lambda}{2n}$ – масштабный коэффициент.

В (8) формула (2) была изменено с использованием (6) и (7): потеря регуляризации определяется измененной регуляризацией и количеством итераций, производимых для взвешенной модели, чтобы увеличить разнообразие и обобщение модели.

$$ML_R = MR * \sum_n |W|^2, \tag{8}$$

где ML_R – потеря регуляризации, MR – модифицированная регуляризация, n количество итераций, W – вес модели.

Увеличение потерь в ансамбле за счет объединения модифицированной функции потерь модели и модифицированной потери регуляризации служит для увеличения разнообразия модели. Предлагаемая формула (9) представлена ниже:

$$EEL = ML + ML_R, \tag{9}$$

где $EEL -$ усовершенствованная потеря ансамбля, $ML -$ модифицированная функция потерь модели, $ML_R -$ модифицированная потеря регуляризации.

3.2 Суть предлагаемого подхода

В предлагаемом решении, во-первых, используется ансамбль различных моделей машинного обучения вместо одного типа модели [5]. Это увеличивает разнообразие решений, достигаемых каждой из различных моделей, увеличивая качество за счет объединения результатов в ансамблевой модели.

Во-вторых, для всех моделей ансамбля используется регуляризация $L2$, чтобы обеспечить обобщение и предотвратить переобучение. Гарантируется, что ансамбль во время обучения приобретает разнообразные способности к обобщению.

Наконец, наряду с $L2$-регуляризацией, используемой для моделей, модель нейронной сети в ансамбле также регуляризована путем своевременного прекращения обучения, чтобы избежать переобучения. В результате модели обучаются обобщению за счет дополнительной параметризации без потребности в изучении статистических нюансов набора данных. Улучшается способность к обобщению всего ансамбля.

В нашей системе использовалось несколько моделей вместо одной модели нейронной сети, чтобы снять ограничения на использование системы за счет улучшения результатов. В системе проводится голосование за различные результаты, и выбирается лучший из них по точности и AUC ROC. Предлагаемая система может диагностировать СД2, сводя к минимуму ограничения и обеспечивая дает высокую точность, чувствительность и специфичность.

На этапе обзора литературы мы поняли, что для диагностики СД2 лучше всего подходит ансамблевый метод. Однако в имеющихся решениях для снятия ограничения используется только модель нейронной сети. В своей работе мы пытаемся дополнительно уменьшить ошибку, чтобы повысить точность и получить лучшие результаты. Качественные результаты

можно получать путем обработки результатов многих моделей, таких как Random Forest, логистическая регрессия, SVM, глубокая нейронная сеть, заставляя ансамблевую модель голосовать за лучший результат, который будет выбран в качестве окончательного (табл. 2).

Табл. 2. Набор моделей машинного обучения для прогнозирования начала СД2
Table 2: Ensemble of machine learning models to predict onset of T2DM

Алгоритм: ансамбль моделей машинного обучения для прогнозирования начала СД2. Вход: данные ЭМК пациентов. Выход: Выходные данные классификации: 1 для пациентов с СД2 и 0 для пациентов без СД2.
BEGIN ШАГ 1. Обнаружение выбросов: удалить данные для пациентов выше 95 процентиля. ШАГ 2. Баланс классов: определите вес каждого класса w_t = количество образцов класса / общее количество образцов. ШАГ 3. Нормализация: нормализовать каждый признак: $x = x - \text{mean}(x)/\text{std}(x)$. ШАГ 4. Обучить сумку из N моделей. Модель логистической регрессии с функцией потерь: $L_{LR} = -(y\log(p) + (1-y)\log(1-p))$ Модель SVM с функцией потерь: $L_{SVM} = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta)$ Модель Random forest с функцией потерь: $L_{RF} = \sum p(i)*(1-p(i))$ Нейронная сеть с функцией кросс-энтропии: $L_{NN} = \sum_i (y_i \log(p_i))$ ШАГ 5. Окончательный вывод ансамблевой модели с использованием алгоритма мажоритарного голосования Бойера-Мура. END

4. Результаты и обсуждение

Среда, использованная для проведения экспериментов в данной работе, основана на языке программирования Python. Python обеспечивает экосистему для работы с машинным обучением и анализом данных. В нашем решении использовались четыре разных датасета, как показано в табл. 3.

Табл. 3. Наборы данных
Table 3. Datasets

Название датасета	Общее количество образцов	Количество положительных образцов	Количество отрицательных образцов
Practice Fusion EHR	9948	1904	8044
Pima Indians Diabetics	2500	1500	1000
UCI AIM94	2000	1000	1000
CA Diabetes Prevalence 2014	6000	1000	5000

Табл. 4. Средние значения показателей
Table 4. Average performance

	Чувствительность (%)	Специфичность (%)	Точность (%)	Время обработки (мс)
SOTA	29.59	96.27	83.51	97.5
Предлагаемое решение	35.8	98.9	87.5	96.6

Сначала сравним показатели результатов предлагаемого решения и наиболее совершенной известной нам системы (state-of-the-art, SOTA). Наша модель достигла точности 87,5% по сравнению с 83,51% у системы SOTA, демонстрируя эффективность предлагаемого решения. Среднее время обработки составило 96,6 мс, как показано в табл. 4.

Образцы создавались для разных возрастных групп мужчин и женщин, страдающих диабетом 2 типа. Распределение по возрасту и положительные/отрицательные случаи показаны на рис. 3. Распределение гендерного признак в датасете показано на рис. 4.

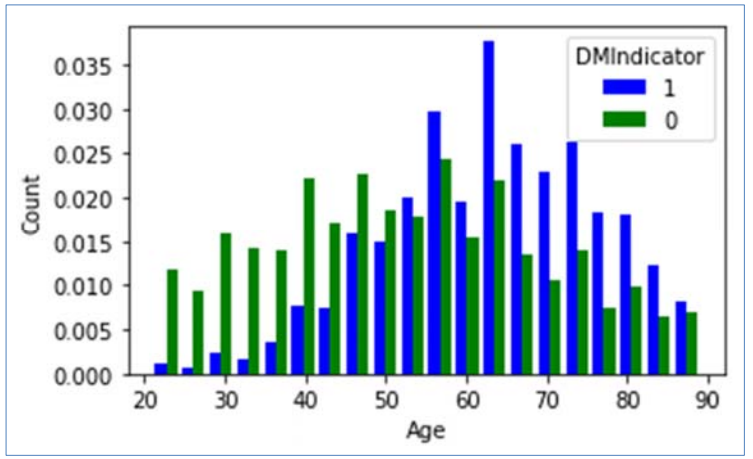

Рис. 3: Распределение по возрасту и положительные/отрицательные случаи
Fig. 3: Distribution of age and the positive/negative cases

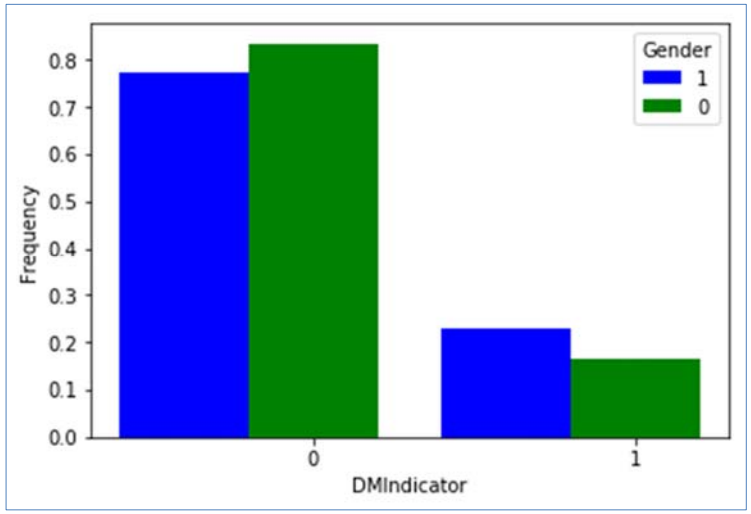

Рис. 4: Гендерный признак в датасете
Fig. 4: Gender feature in the dataset

Образцы данных использовались для обучения с использованием языка программирования Python и соответствующих инструментов анализа данных, таких как Scikit-learn, Pandas и Matplotlib. Сравнивались результаты (точность, чувствительность и специфичность) предложенного решения на основе ансамблевой модели и глубоким и широким решением SOTA. В табл. 5-12 приводятся средние результаты обеих систем, полученные при тестировании для 10 различных групп.

Табл. 5. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета Practice Fusion ЭМК (группа выборки 1: женщины в возрасте от 40 до 70 лет)
Table 5. Accuracy, Sensitivity and Specificity results comparison of SOTA (state-of-the-art) vs Proposed Solution using Practice Fusion EHR Dataset (Sample group 1: women of age 40 to 70)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	82.98	27.22	97.12	97.5
2	83.01	27.23	97.23	97.3
3	83.4	26.56	97.04	97.5
4	82.93	27.23	97.33	97.4
5	83.24	27.04	97.27	97.3
6	83.21	27.52	97.1	97.3
7	83.41	26.8	97.22	97.4
8	82.99	27.1	97.02	97.2
9	83.14	27.23	97.3	97.3
10	83.12	26.82	97.12	97.4
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	84.72	28.03	97.36	96.7
2	85.23	27.64	97.22	96.5
3	84.16	27.14	97.26	96.7
4	84.52	28.09	97.15	96.6
5	84.11	27.15	98.15	96.6
6	85.25	27.83	97.54	96.5
7	84.13	27.94	97.61	96.7
8	84.81	27.42	97.42	96.5
9	84.33	27.25	97.73	96.6
10	84.14	27.73	97.59	96.6

Табл. 6. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета Practice Fusion ЭМК (группа выборки 2: мужчины в возрасте от 30 до 60 лет)
Table 6. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using Practice Fusion EHR Dataset (Sample group 2: men of age 30 to 60)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	82.71	26.71	97.07	97.5
2	82.99	26.61	97.14	97.2
3	82.80	27.22	97.13	97.6
4	83.44	26.77	97.35	97.5
5	82.85	27.08	97.40	97.3
6	83.30	27.05	97.38	97.5
7	82.70	27.04	97.12	97.4
8	83.45	26.61	97.36	97.4
9	83.48	26.84	96.71	97.4
10	83.27	26.82	96.78	97.5

№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	84.91	28.16	97.81	96.7
2	85.07	28.09	98.20	96.5
3	84.66	27.54	97.88	96.7
4	84.90	27.39	98.27	96.6
5	85.14	27.11	98.23	96.6
6	85.11	27.28	97.91	96.5
7	83.95	27.95	98.05	96.7
8	84.62	27.51	97.87	96.5
9	85.09	27.52	98.30	96.6
10	84.13	28.17	97.85	96.6

Табл. 7. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета Prima Indians (группа выборки 1: женщины в возрасте от 40 до 70 лет)
Table 7. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using Prima Indians Dataset (Sample group 1: Women of age 40 to 70)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	82.30	27.11	96.67	43.1
2	83.26	26.72	97.13	43.0
3	82.59	26.77	97.38	43.2
4	83.13	27.16	97.03	43.3
5	83.05	26.81	97.10	43.0
6	82.74	26.72	97.07	43.3
7	81.68	27.11	97.16	43.1
8	82.17	26.72	97.25	43.2
9	81.24	26.71	96.79	43.3
10	82.82	27.27	96.79	43.4
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	83.83	27.37	98.28	42.9
2	83.75	27.78	97.83	42.5
3	84.11	27.29	98.10	42.9
4	84.19	27.14	97.93	42.5
5	84.10	27.37	98.03	42.9
6	84.31	27.17	97.89	42.6
7	84.29	27.71	98.22	42.6
8	83.23	27.33	98.18	42.9
9	84.36	27.17	98.05	42.6
10	83.64	27.48	97.85	43.0

Табл. 8. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета Prima Indians (группа выборки 2: мужчины в возрасте от 30 до 60 лет)
Table 8. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using Prima Indians Dataset (Sample group 2: Men of age 30 to 60)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	82.99	27.26	97.16	43.1
2	82.68	26.87	97.06	43.0
3	82.98	27.15	97.09	43.2
4	81.24	27.05	96.90	43.3
5	82.43	27.25	96.85	43.0
6	81.34	26.80	96.74	43.3
7	81.69	27.08	97.27	43.1
8	82.94	27.12	96.78	43.2
9	81.26	27.10	96.88	43.3
10	83.03	26.97	97.18	43.4
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	83.65	27.34	98.18	42.9
2	84.32	27.56	97.97	42.5
3	83.84	27.82	98.04	42.9
4	83.84	28.15	97.84	42.5
5	83.37	27.77	98.07	42.9
6	84.28	27.12	97.80	42.6
7	84.28	27.42	98.00	42.6
8	83.48	27.52	98.19	42.9
9	83.21	28.17	98.09	42.6
10	83.29	27.67	97.90	43.0

Табл. 9. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета UCI AIM94 (группа выборки 1: женщины в возрасте от 40 до 70 лет)
Table 9. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using UCI AIM94 Dataset (Sample group 1: women of age 40 to 70)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	87.14	26.90	97.31	42.8
2	88.06	26.74	96.77	42.8
3	87.87	27.22	96.69	43.0
4	87.12	27.18	97.04	42.7
5	88.19	26.73	97.29	43.0
6	87.24	26.74	97.11	42.6
7	87.97	27.29	97.22	42.9
8	87.53	26.80	97.34	42.8

9	87.50	27.21	97.08	42.6
10	88.24	27.27	96.96	42.9
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	87.95	28.03	97.96	42.4
2	88.79	27.15	97.93	42.1
3	88.22	28.01	97.87	42.3
4	88.14	27.16	97.82	42.3
5	88.88	28.14	98.19	42.4
6	88.37	28.11	98.10	42.2
7	89.14	27.83	98.24	42.1
8	89.12	27.93	98.30	42.3
9	89.04	27.55	98.13	42.4
10	88.77	28.15	97.99	42.1

Табл. 10. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета UCI AIM94 (группа выборки 2: мужчины в возрасте от 30 до 60 лет)
Table 10. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using UCI AIM94 Dataset (Sample group 2: men of age 30 to 60)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	87.14	27.16	96.87	43.0
2	88.06	27.15	97.28	42.8
3	87.87	27.15	96.66	42.6
4	87.12	27.26	97.09	42.7
5	88.19	26.99	96.60	42.8
6	87.24	27.11	96.62	42.5
7	87.97	27.10	97.19	43.0
8	87.53	26.98	97.31	42.8
9	87.50	26.82	97.36	42.7
10	88.24	26.96	97.11	42.6
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	87.99	28.08	98.06	42.4
2	89.39	27.79	98.12	42.3
3	88.64	28.11	97.86	42.2
4	89.00	27.20	97.91	42.4
5	88.34	28.17	98.12	42.3
6	88.77	27.46	97.88	42.4
7	89.32	27.65	98.27	42.1
8	88.55	27.16	97.84	42.2
9	88.01	27.12	97.87	42.5
10	88.49	27.54	98.05	42.4

Табл. 11. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета CA Diabetes Prevalence 2014 (группа выборки 1: женщины в возрасте от 40 до 70 лет)
Table 11. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using CA Diabetes Prevalence 2014 Dataset (Sample group 1: Women of age 40 to 70)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	90.96	50.71	97.05	69.0
2	91.30	51.13	96.62	69.3
3	90.55	50.58	97.01	69.1
4	90.26	50.67	96.76	69.0
5	90.12	51.32	96.60	69.0
6	90.65	51.88	96.66	69.2
7	90.03	52.76	96.75	69.1
8	91.30	50.94	97.10	69.0
9	90.50	51.09	97.08	69.2
10	90.38	51.13	97.12	69.1
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	91.78	27.47	97.99	69.1
2	92.87	27.47	97.85	68.7
3	92.31	27.72	98.25	68.6
4	91.04	27.49	98.25	68.9
5	92.26	27.97	97.96	68.6
6	91.62	27.25	97.82	68.5
7	92.18	28.00	98.01	68.7
8	92.92	27.94	97.85	68.7
9	92.75	27.37	97.91	68.9
10	91.43	27.92	98.15	68.8

Табл. 12. Сравнение точности, чувствительности и специфичности SOTA и предлагаемого решения с использованием датасета CA Diabetes Prevalence 2014 (группа выборки 2: мужчины в возрасте от 30 до 60 лет)
Table 12. Accuracy, Sensitivity and Specificity results comparison of SOTA vs Proposed Solution using CA Diabetes Prevalence 2014 (Sample group 2: men of age 30 to 60)

№ выборки	**Решение SOTA**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	91.30	52.17	97.14	69.1
2	91.25	52.40	96.91	69.2
3	91.22	51.69	96.69	69.1
4	90.11	52.26	96.87	69.0

5	90.41	52.65	96.64	69.1
6	90.74	52.52	96.65	69.3
7	91.33	52.31	97.05	69.3
8	90.73	51.39	97.33	69.0
9	91.10	51.18	96.70	69.0
10	90.10	51.41	96.95	69.2
№ выборки	**Предлагаемое решение**			
	Точность (%)	Чувствительность (%)	Специфичность (%)	Время обработки (мс)
1	91.64	54.40	98.10	69.1
2	91.82	53.30	98.21	68.7
3	92.98	54.07	97.93	68.6
4	91.64	54.13	97.90	68.9
5	92.55	53.45	97.99	68.6
6	91.67	53.72	97.98	68.5
7	91.39	53.43	98.29	68.7
8	92.59	53.07	98.13	68.7
9	91.47	53.00	97.99	68.9
10	91.05	54.27	98.03	68.8

Используемая в предлагаемом решении ансамблевая модель обеспечивает улучшенные точность, чувствительность и специфичность по сравнению с решением SOTA. Точность в среднем составляет 87,5%, чувствительность – 35,8%, специфичность – 98,9%, время обработки – 96,6 мс.

Эксперименты подтвердили правильность нашего теоретического подхода с использованием нескольких различных моделей. Вместо использования одной модели или набора однотипных моделей, как в решении SOTA, мы используем набор моделей машинного обучения: логистическая регрессия, SVM, Random Forest и нейронные сети. Комбинируя прогнозы этих моделей, мы достигаем более качественных результатов, чем SOTA. Из-за индивидуальных особенностей разные модели изучают разные аспекты обучающих данных. Это приводит к получению более качественного решения, чем при использовании одной модели. Наше решение также снижает проблему переобучения за счет комбинирования регуляризации отдельных моделей. Точность, чувствительность и специфичность в среднем улучшились на 3-4%. Немного сократилось время обработки

Прогнозирование СД2 – активная область науки о данных и медицинских исследований. В решении SOTA используется глубокое и широкое обучение с впечатляющим улучшением точности, чувствительности и специфичности. В нашей исследовательской работе были рассмотрены ограничения модели SOTA. Нам удалось добиться повышения точности до 87,5% по сравнению с 84,28% решения SOTA. Это связано с расширением модели путем формирования ансамбля из четырех разных алгоритмов машинного обучения, что, в свою очередь, улучшает обобщение и общее качество решения. Предлагаемое решение неизменно превосходит современную модель в нескольких экспериментах, описанных в этой работе (см. табл. 13).

Табл. 13. Сравнение решения SOTS и предлагаемого решения
Table 13, Comparison table between state of art and proposed solutions

	Предложенное решение
Суть подхода	Обнаружение T2DM с использованием ансамбля моделей машинного обучения.
Точность	За счет использования ансамблевого подхода точность увеличена до 85,2%.

Чувствительность	Чувствительность увеличена с 29,59% до 32,1%.
Специфичность	Специфичность увеличена с 96,27% до 98,3%.
Предлагаемая формула	Потери ансамбля и регуляризации для увеличения разнообразия моделей ансамбля можно представить как $EEL = ML + ML_R$
Вклад 1	Предлагаемая модель использует преимущества различных моделей машинного обучения для обеспечения прогнозирования, включая глубокое обучение. Различные алгоритмы достигают разнообразных решений, позволяя ансамблю извлечь выгоду из их разнообразия.
Вклад 2	Предлагаемая модель использует алгоритм большинства голосов Бойера-Мура для получения окончательного результата прогноза. Это позволяет различным моделям голосовать за окончательный прогноз. Отдельные прогнозы можно сравнить с прогнозом большинства, чтобы определить, какие модели работают лучше в определенных ситуациях. Это делает результаты немного более интерпретируемыми и надежными.
Вклад 3	Предлагаемое решение использует регуляризацию $L2$ для всех моделей для предотвращения переобучения. Наряду с этим модели нейронных сетей обучаются со своевременным прекращением обучения, чтобы обеспечить правильное обобщение.
	Решение SOTA
Суть подхода	Обнаружение T2DM с использованием ансамбля глубоких и широких сетей.
Точность	Обеспечиваемая точность составляет 83,51%.
Чувствительность	Обеспечиваемая чувствительность составляет 29,59%.
Специфичность	Обеспечиваемая специфичность составляет 96,27%.
Предлагаемая формула	Потери ансамбля и регуляризации для увеличения разнообразия моделей ансамбля могут быть представлены как $EL = L + L_R$
Вклад 1	Решение SOTA использует ансамбль одной и той же модели, что обесценивает ансамбль, поскольку в решениях, достигаемых одной и той же моделью, очень мало или совсем нет разнообразия.
Вклад 2	Решение SOTA усредняет выходную вероятность 10 моделей, чтобы дать окончательную вероятность для двоичной классификации.
Вклад 3	Решение SOTA использует регуляризацию $L1$, что приводит к разреженности. Но более важно избежать переобучения, чем стимулировать разреженность.

6. Заключение и будущие исследования

Обнаружение СД2 является важной проблемой медицинского направления науки о данных. Если на предсказание или идентификацию болезни потребуется много времени, пациент будет позже предупрежден и позже начнет лечиться. Это можно считать основной причиной тысяч смертей ежегодно.

Предлагаемая в этой работе система улучшает общие возможности прогнозирования СД2 с использованием набора из нескольких моделей машинного обучения и глубокого обучения. В настоящее время большинство исследований в этой области сосредоточено на использовании только одной модели для прогнозирования диабета. Однако известно, что ансамбли превосходят отдельные модели. Когда используются разные типы моделей, разнообразие помогает делать более точные прогнозы. Схема мажоритарного голосования использует результаты нескольких моделей для окончательного точного прогноза. Функция регуляризации в нашей работе изменена, чтобы включить регуляризацию всех моделей в ансамбле, что помогает предотвратить переобучение и способствует возможности обобщения предлагаемой системы. Предлагаемая система была реализована на языке Python и при тестировании показала более высокую точность, чувствительность и специфичность.

Имеется много возможностей для совершенствования описанного решения. Для повышения качества ансамблевого решения может быть реализована обучаемая система взвешенного голосования. Это будет способствовать тому, чтобы ансамбль использовал лучшие стороны отдельных моделей. Можно ввести другие типы моделей машинного обучения, а также сложные модели глубокого обучения, чтобы еще больше улучшить показатели системы.

Список литературы / References

[1] Bernardini M., Romeo L., Misericordia P., and Frontoni E. Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, 2020, pp. 235-246.

[2] Houri. O., Gil. Y., Berezowsky A., Wiznitzer A. et al. 339: Future Type-2 diabetes prediction following pregnancy – using a novel machine learning, American Journal of Obstetrics and Gynecology, vol. 222, issue 1, Supplement, 2020, p. S228.

[3] Kölle Konstanze, Biester Torben, Christiansen Sverre et al. Pattern Recognition Reveals Characteristic Postprandial Glucose Changes: Non-Individualized Meal Detection in Diabetes Mellitus Type 1. IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, 2020, pp. 594-602.

[4] Kezhi Li, John Daniels, Chengyuan Liu et al. Convolutional Recurrent Neural Networks for Glucose Prediction. IEEE Journal of Biomedical and Health Informatics vol. 24, no. 2, 2020, pp. 603-613.

[5] Binh P. Nguyen, Hung N Pham, Hop Tran et al. Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records. Computer Methods and Programs in Biomedicine, vol. 182, 2019, article id 105055.

[6] Perveen S., Shahbaz M., Saba T. et al. Handling irregularly sampled longitudinal data and prognostic modeling of diabetes using machine learning technique. IEEE Access, vol. 8, 2000, pp. 21875-21885.

[7] Vivek Rai, Daniel X. Quang, Michael R. Erdos et al. Single-cell ATAC-Seq in human pancreatic islets and deep learning upscaling of rare cells reveals cell-specific type 2 diabetes regulatory signatures. Molecular Metabolism, vol. 32, 2020, pp. 109-121.

[8] Ramazi Ramin, Perndorfer Christine, Soriano C.E. et al. Multi-modal Predictive Models of Diabetes Progression. In Proc. of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2019, pp. 253-258.

[9] Sierra-Sosa D., Garcia-Zapirain B., Castillo C. et al. Scalable Healthcare Assessment for Diabetic Patients Using Deep Learning on Multiple GPUs. IEEE Transactions on Industrial Informatics, vol. 15, no. 10, 2019, pp. 5682 – 5689.

[10] Agata Wesolowska-Andersen, Grace Zhuo Yu, Vibe Nylander et al. Deep learning models predict regulatory variants in pancreatic islets and refine type 2 diabetes association signals. eLife 2020;9:e51503, 2020.

[11] Tomohide Yamada, Kosuke Iwasaki, Shotaro Maedera et al. Myocardial infarction in type 2 diabetes using sodium–glucose co-transporter-2 inhibitors, dipeptidyl peptidase-4 inhibitors or glucagon-like peptide-1 receptor agonists: proportional hazards analysis by deep neural network based machine learning. Current Media Research and Option, vol. 36, no. 3, 2000, pp. 404-410.

[12] Zaitcev A., Eissa R.M., Hui Z. et al. A Deep Neural Network Application for Improved Prediction of HbA1c in Type 1 Diabetes, IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, 2020, pp. 2932-2941.

[13] Xianglong Zeng, Haiquan Chen, Yuan Luo, and Wenbin Ye. Automated Diabetic Retinopathy Detection Based on Binocular Siamese-Like Convolutional Neural Network, IEEE Access, vol. 7, 2019, pp. 30744-30753.

[14] Anuradha, Akansha Singh, and Gaurav Gupta. ANT FDCSM: A novel fuzzy rule miner derived from ant colony meta-heuristic for diagnosis of diabetic patients. Journal of Intelligent & Fuzzy Systems, vol. 36, no. 1, 2019, pp. 747-760.

[15] Muhammad Fazal Ijaz, Ganjar Alfian, Muhammad Syafrudin, and Jongtae Rhee. Hybrid Prediction Model for Type 2 Diabetes and Hypertension Using DBSCAN-Based Outlier Detection, Synthetic Minority Over Sampling Technique (SMOTE), and Random Forest, Applied Sciences, vol. 8, no. 8, 2018, pp. 1-22.

Информация об авторах / Information about authors

Маллика Г.Ч., магистр, разработчик программного обеспечения. Область научных интересов: обработка изображений.

Mallika G.C., Master, Software Developer. Research interests: image processing.

Абир АЛЬСАДУН, кандидат наук, доцент. Область научных интересов: обработка изображений.

Abeer ALSADOON, Ph.D., Associate Professor. Research interests: image processing.

Дуонг Тху Ханг ФАМ, магистр, преподаватель. Область научных интересов: базы данных, интеллектуальный анализ данных, педагогическая методология.

Duong Thu Hang PHAM, Master, Lecturer. Research interests: databases, data mining, pedagogical methodology.

Salma Hameedi ABDULLAH, Ph.D, Lecturer. Research interests: machine learning, computer vision, deep learning.

Сальма Хамиди АБДУЛЛА, Ph.D, Lecturer. Область научных интересов: машинное обучение, компьютерное зрение, глубокое обучение.

Ха Тхи МАЙ, магистр, преподаватель. Область научных интересов: интеллектуальный анализ данных, программная инженерия.

Ha Thi MAI, Master, Lecturer. Research interests: data mining, software engineering.

П.В. Чандана ПРАСАД, доктор философии, доцент. Область научных интересов: обработка изображений.

P.W. Chandana PRASAD, Ph.D., Associate Professor. Research interests: image processing.

Чан Куок Винь НГУЕН, кандидат наук, преподаватель. Область научных интересов: базы данных, интеллектуальный анализ данных, программная инженерия.

Tran Quoc Vinh NGUYEN, PhD, Lecturer. Research interests: databases, data mining, software engineering.