УДК 123;616.894-053.8-07

ИТ-ДИАГНОСТИКА БОЛЕЗНИ АЛЬЦГЕЙМЕРА НА ОСНОВЕ АНАЛИЗА ГОЛОСОВОЙ ИНФОРМАЦИИ

В.А. ВИШНЯКОВ[1], Ю. ЧУЙЭ[2]

[1]*Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», ул. П. Бровки, 6, Минск, 220600, Беларусь*
ORCID: https://orcid.org/0000-0003-2929-8958

[2]*Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники», ул. П. Бровки, 6, Минск, 220600, Беларусь*
ORCID: https://orcid.org/0009-0009-2342-1856

Поступила в редакцию 15 июня 2023

Рассматривается использование машинного обучения для распознавания болезни Альцгеймера по фонологическим данным на ранней стадии заболевания. Данные, используемые в процессе анализа, взяты из набора данных AdeSS2020 Challenge dataset, который содержит голосовую информацию как пациентов с болезнью Альцгеймера (для обучения нейронной сети), так и пациентов для распознавания. Подход, используемый в этой статье, основан на модели классификации с использованием машинного обучения. Сначала из голосовых данных были извлечены как фонологические, так и семантические признаки, затем выполнено машинное обучение нейронной сети на основе этих признаков с использованием алгоритма случайного леса. Использован также алгоритм GridSearchCV для оптимизации гиперпараметров классификатора случайного леса. В процессе распознавания болезни Альцгеймера точность классификации модели достигла 85 %.

Ключевые слова: машинное обучение, алгоритм случайного леса, нейронная сеть, параметры оптимизации, точность.

Введение. Потеря памяти и нарушение речи являются одними из самых ранних симптомов болезни Альцгеймера. Они стали направлением, в котором ученые работают, чтобы помочь диагностировать заболевание пациентам с болезнью Альцгеймера (БА) с помощью машинного обучения классификатора на базе нейронной сети.

Развитие технологии Интернета вещей (IoT) позволило обеспечить безопасность пациентов с помощью кнопок экстренной помощи, систем глобального позиционирования, интеллектуальных детекторов и т. д. [1]. Достижения в области машинного обучения позволили с помощью ИТ-диагностики обнаруживать изменения в голосе и фонологическом ритме, которые не всегда врачи могут обнаружить на слух. Произношение при болезни Альцгеймера характеризуется вариациями различных временных и акустических фонологических параметров с клиническими симптомами дисфазии, то есть нарушением называния, понимания на слух и письменно, беглой, но невнятной речью и разговором с семантическими ошибками [2].

Известные результаты исследований. Зарубежные исследователи проводили ИТ-анализ вербальных и лингвистических расстройств для пациентов с БА. В статье [3] авторы извлекли переменные из повествовательной речи пациентов с болезнью Альцгеймера, в ходе исследовательского анализа отобрали более 370 доступных признаков для построения модели и достигли 81,92 % точности классификации при отличии пациентов от здоровой контрольной

группы. В работе [4] авторы использовали три алгоритма, основанных на статистических и нейронных методах, для встраивания вербальных аудиосигналов с использованием N-граммы, i-вектора и x-вектора и достигли точности дифференцирования 83,6 % в наборе данных pitt Corpus.

Авторы статьи использовали машинное обучение в качестве методологии для идентификации болезни Альцгеймера, комбинируя фонологические признаки с семантическими для достижения различия между пациентами с болезнью Альцгеймера и здоровыми контрольными группами, используя модель TfidfVectorizer, классификатор случайных лесов и алгоритм GridSearchCV.

Набор данных ADeSS challenge dataset и модель TfidfVectorizer. Исходные данные были получены из набора данных ADeSS challenge dataset [5], основной целью которого является внедрение распознавания БА, при котором можно систематически сравнивать различные методологии. Набор данных содержит 1955 голосовых сегментов от 78 участников, являющихся пациентами, и 2122 голосовых сегмента от 78 участников, не являющихся пациентами с БА, причем записи были акустически улучшены с помощью удаления стационарного шума.

В наборе данных использовалось задание по описанию картинки «кража печенья» из Бостонского диагностического исследования афазии [6], чтобы вызвать повествовательную речь. Протокол предписывал интервьюерам показывать фотографию пациентам и поощрять их описывать картинку как можно подробнее. Доступный набор данных содержит полный улучшенный звук и стандартизированные аудиоблоки. В данной статье было использовано полное дополненное аудио и соответствующие транскрипции. Расшифрованные тексты были предоставлены dataset, все расшифровки были аннотированы с использованием системы кодирования ЧАТА [7]. К каждой расшифровке прилагается аудиофайл, позволяющий проводить параллельный лексический и акустический анализ.

TfidfVectorizer − это реализация с открытым исходным кодом для обработки текста на естественном языке, его центральной концепцией является метод TF-IDF (Term frequency − обратная частота документа), распространенный метод взвешивания для поиска информации и интеллектуального анализа данных. Основная идея метода TF-IDF заключается в том, что если слово или фраза встречается в статье с высокой частотой и редко встречается в других статьях, считается, что они имеют высокую релевантность для документа, что означает, что слово или фраза обладают хорошей способностью дифференцировать категории, которые подходят для классификации.

TfidfVectorizer использует частоту обратной области (IDF) и частоту терминов (TF) слов для вычисления их соответствующих значений частоты обратной области (TF-IDF) [8]. Значение TF-IDF, в свою очередь, используется для присвоения веса или важности слову при выполнении анализа. IDF слова «*w*» в текстовом наборе и TF-IDF слова «*w*» в данном документе *d* вычисляются с использованием приведенных ниже уравнений:

$$\text{idf}(w) = \log(N / \text{df}(w)),$$

$$\text{Tf\_idf}(w,d) = \text{tf}(w,d) \times \text{idf}(w),$$

где N − общее количество документов; df(*w*) − количество документов со словом «*w*»; tf(*w*,*d*) − частота употребления слова «*w*» в документе *d*, которая обозначает количество раз, когда «*w*» появлялось в документе *d*, деленное на общее количество слов в документе.

Классификатор на основе случайного леса. Алгоритм случайного леса объединяет несколько деревьев с помощью идеи коллективного обучения, его базовой единицей является дерево решений. В частности, каждое дерево решений является классификатором, для входной выборки N деревьев будут иметь N результатов классификации. Случайный лес определяет категорию с наибольшим количеством голосов в качестве конечного результата, одна из простейших идей объединения в пакеты. Он может эффективно предотвращать риск переобучения, уменьшать дисперсию в дереве решений, обладает хорошей устойчивостью к шуму и выбросам [9], обладая характеристиками способности обрабатывать многомерные данные с высокой точностью.

Алгоритм обучения классификатора случайного леса включает:

1. Выборки из обучающего набора отбираются случайным образом для формирования нового обучающего набора (метод пакетирования, Bootstraping).

2. Случайным образом выбираются некоторые функции из обучающего набора, чтобы сформировать новый набор функций.

3. Дерево решений обучается на основе нового обучающего набора и нового набора функций. Дерево решений обучается путем непрерывного разделения набора данных на более мелкие подмножества до тех пор, пока количество подмножеств не станет настолько малым, что оно каким-то образом будет предопределено или его нельзя будет разделить дальше.

4. Повторение шагов 1−3, чтобы обучить несколько деревьев принятия решений для формирования случайного леса.

Параметры деревьев решений в случайном лесу необходимо выбирать вручную, если этот гиперпараметр выбран неправильно, это может оказать большое влияние на производительность модели, поэтому GridSearchCV был использован авторами для выполнения задачи поиска. GridSearchCV − это вариант GridSearch, который представляет собой алгоритм, использующий перекрестную проверку для выбора наилучшего гиперпараметра.

В традиционном алгоритме GridSearch можно только выполнить обучение модели и настройку гиперпараметров на обучающем наборе, а затем оценить модель на тестовом наборе. Однако это приводит к переобучению модели, поскольку модель также будет воспринимать шум обучающего набора как достоверную информацию.

Чтобы решить эту проблему был использован метод перекрестной проверки. Перекрестная проверка делит обучающий набор на несколько подмножеств, затем каждое подмножество, в свою очередь, используется в качестве набора для проверки, в то время как остальные подмножества используются в качестве обучающего набора, многократно обучая и проверяя модель, кульминацией чего является вычисление среднего значения показателей производительности по всем наборам для проверки.

По сравнению с традиционным алгоритмом GridSearch, GridSearchCV может не только снизить риск переобучения, но и повысить стабильность и способность модели к обобщению.

Методология распознавания. После получения полного улучшенного аудио и соответствующих расшифрованных текстов для фонологического сигнала были выделены пять пространственно-временных характеристик и две демографические характеристики. Для расшифрованных текстов были извлечены три семантических признака и выполнен ввод их в модель TfidVectorizer, чтобы преобразовать текстовые данные в числовые векторы признаков, которые могут быть использованы алгоритмами машинного обучения. Затем набор функций данных был нормализован и разделен на обучающие и тестовые наборы данных в соотношении 8:2. Тестовый набор данных был использован для проверки обученной модели и получения результатов распознавания БА. На рис. 1 показана методология, используемая авторами для распознавания по изменению голоса болезни Альцгеймера.

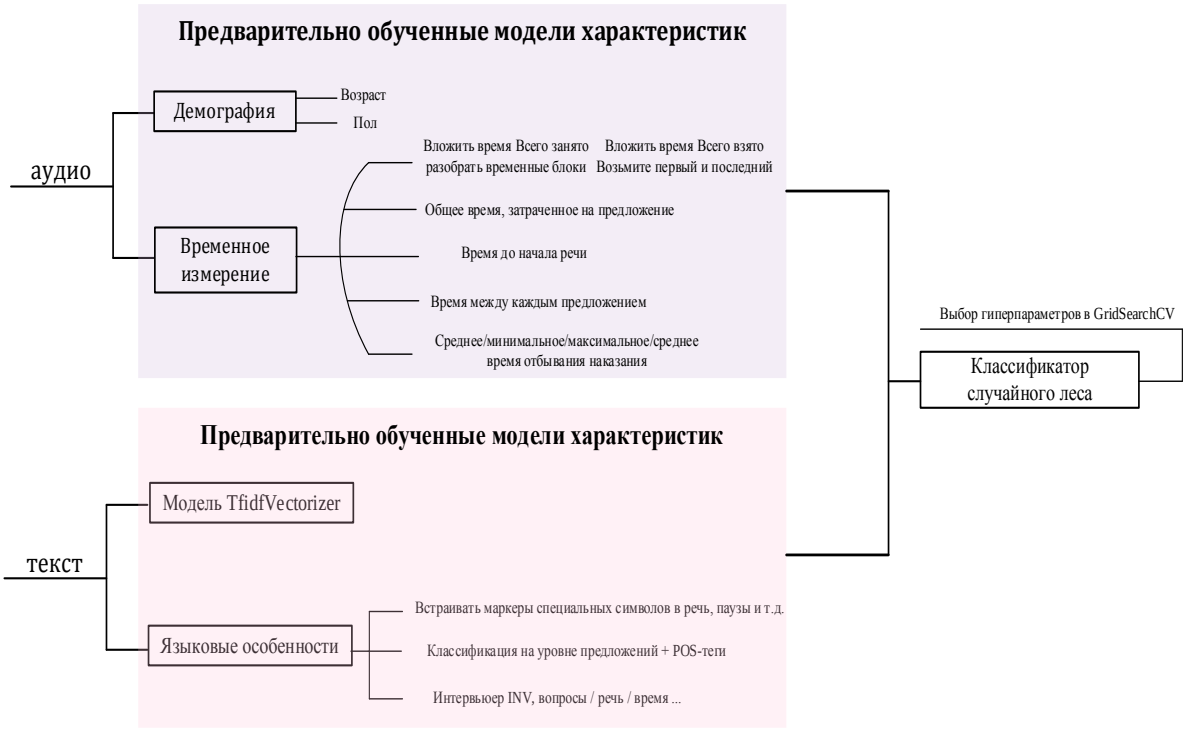

Рис. 1. Методика ИТ-диагностики БА

Результаты распознавания. Язык набора данных для этого эксперимента – английский, язык программирования – Python, один из популярных языков программирования в области машинного обучения. Он не только обладает богатыми библиотеками машинного обучения, такими как scikit-learn, TensorFlow, PyTorch, Keras и др., чтобы помочь практикам быстро создавать модели машинного обучения, но также имеет возможностями визуализации для отображения результатов. Перечень библиотек Python, использовавшихся в экспериментах:

1. Glob – предоставляет функцию сопоставления файлов, которая позволяет сопоставлять имена файлов с помощью подстановочных знаков;
2. Numpy – предоставляет функции манипулирования многомерными массивами и численных вычислений;
3. Pandas – предоставляет возможности анализа для обработки больших объемов структурированных данных;
4. Sklearn – предоставляет алгоритмы машинного обучения, предварительную обработку данных, их оценку;
5. Scipy – предоставляет функции научных расчетов и статистического анализа;
6. Torch – предоставляет платформу глубокого обучения для тензорных вычислений и моделирования нейронных сетей;
7. Transformers: предоставляет модели глубокого обучения и инструменты, связанные с обработкой естественного языка.

В эксперименте на базе обученной модели была использована матрица путаницы для оценки распознавания. Была выполнена оптимизация гиперпараметров случайного леса, используя метод GridSearchCV. После использования классификатора случайного леса матрицы смешения, составленные в соответствии с результатами, приведены на рис. 2 а и б.

.

а) б)

Рис. 2. Результаты:
а – матрица смешения обучающего набора данных; б – матрица смешения тестового набора данных

Экспериментальные результаты распознавания болезни Альцгеймера на тестовом наборе данных приведены в табл. 1.

Таблица 1

Экспериментальные результаты распознавания болезни Альцгеймера

Datasets	Accuracy	Precision	Recall	F1 Score
AD_speech	85 %	85 %	80 %	76 %

Таким образом, точность и прецизионность набора данных составили 85 %. Точность тестов в этом исследовании и существующих исследованиях по одному и тому же набору данных AD была сопоставлена в табл. 2.

Таблица 2

Сравнение с другими исследованиями

Datasets	Researchers	Research methods	Test Accuracy
Ad_speech	Ablimit A [12]	SVM	76,7 %
Ad_speech	Authors	Random forest classifier	85 %

Из табл. 2 видно, что результаты этого эксперимента превзошли классификатор SVM и позволили добиться более точной классификации. Полученная модель распознавания БА не требует клинических исследований и будет использована в сети ИВ для получения системы ИТ-диагностики.

Заключение. 1. В статье авторы представили состояние распознавания речи, актуальное при диагностике болезни Альцгеймера. Использован общедоступный набор данных изменения голоса больных Айцгеймера для машинного обучения нейронной сети. Этот набор данных был применен для обучения нейронной сети с использованием классификатора по алгоритму случайного леса с оптимизированными гиперпараметрами.

2. Применен алгоритм GridSearchCV, который делит пространство гиперпараметров на подпространства и выполняет перекрестную проверку для каждого из них, чтобы выбрать наилучшую комбинацию гиперпараметров. Набор функций данных был нормализован и разделен на обучающие и тестовые наборы данных в соотношении 8:2. Тестовый набор данных был использован для проверки результатов распознавания БА.

3. Экспериментальные результаты распознавания болезни Альцгеймера на тестовом наборе составили 85 %. Результаты эксперимента превзошли классификатор SVM в 76,7 %

и позволили добиться более точной классификации. Полученная модель распознавания БА не требует клинических исследований и будет использована в сети ИВ для получения системы ИТ-диагностики.

IT DIAGNOSTICS OF ALZHEIMER'S DISEASE BASED ON THE ANALYSIS OF VOICE INFORMATION

U.A. VISHNIAKOU, Y. CHUYE

Abstract

The purpose of the article is to use machine learning to recognize Alzheimer's disease from phonological data at an early stage of the disease. The data used in the analysis process is taken from the AdeSS2020 Challenge dataset, which contains voice information of both patients with Alzheimer's disease (for neural network training) so are patients for recognition. The approach used in this article is based on a classification model using machine learning. First, both phonological and semantic features were extracted from the voice data, then machine learning of a neural network based on these features was performed using a random forest algorithm. The GridSearchCV algorithm is also used to optimize the hyperparameters of the random forest classifier. In the process of recognizing Alzheimer's disease, the classification accuracy of the model reached 85 %.

Список литературы

1. Technology-based Tools and Services for People with Dementia and Carers / K. Lorenz [et al.] // Mapping Technology onto the Dementia Care Pathway. Dementia (London). – 2019. – № 18 (2). – P. 725–741.

2. Speaking in Alzheimer's Disease, is that an Early Sign? Importance of Changes in Language Abilities in Alzheimer's Disease / G. Szatloczki [et al.] // Frontiers in aging neuroscience. – 2015. – Vol. 7. – P. 7–195.

3. Fraser, K. S. Linguistic Features Identify Alzheimer’s Disease in Narrative Speech / K. S. Fraser, J. A. Meltzer, F. Rudzizh // J Alzheimer Dis. – 2016. – № 49 (2). – P. 407–422.

4. Zargarbashi, S. S. Y. A Multi-Modal Feature Embedding Approach to Diagnose Alzheimer Disease from Spoken Language / S. S. Y. Zargarbashi, B. Babaali // Computer Science. – 2019. – P. 1–14.

5. Luz, F. Alzheimer’s Dementia Recognition through Spontaneous Speech / F. Luz [et al.] // Front Comput Sci. – 2021. – Vol. 3. – № 3. – A. 780169.

6. Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer's Disease / A. Pompili [et al.] // IEEE Journal of Selected Topics in Signal Processing. – 2020. – № 14 (2). – P. 261–271.

7. MacWhinney, B. The CHILDES project: Tools for analyzing talk. / B. MacWhinney // Child Language Teaching and Therapy. – 2020. – № 8 (2). – 120 p.

8. Kumar, V. A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus / V. Kumar, B. Subba // National Conference on Comunications (NCC), Kharagpur, India, 2020. – P. 1–6.

9. Breiman, L. Random Forests / L. Breiman // Machine Learning. – 2001. – № 45. – P. 5–32.

10. Exploring Dementia Detection from Speech: Cross Corpus Analysis / A. Ablimit [et al.] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2022. – P. 6472–6476.