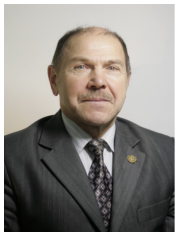
UDC 616-009.5; 004.934

# PRESENTATION AND PROCESSING OF DATA FOR THE DIAGNOTICSS OF NEUROLOGICAL DISEASES



***V.A. Vishnyakov***
*Doctor of Technical Sciences, Professor of the Department of Infocommunication Technologies, BSUIR*
*vish@bsuir.by*

***Yu ChuYue***
*PhD student in the Department of infocommunication technologies at BSUIR.*
*ycy18779415340@gmail.com*

***Xia YiWei***
*PhD student in the Department of infocommunication technologies at BSUIR.*
*903016812@qq.com*

***V.A. Vishnyakov***
*Area of professional interests/research: information management and security in infocommunications, Internet of things, blockchain, IT in education, intelligent control systems.*

***Yu ChuYue***
*PhD student in the Department of Infocommunication technologies of the BSUIR.*

***Xia YiWei***
*PhD student in the Department of Infocommunication technologies of the BSUIR.*

**Annotation.** Big data is triggering a revolution in scientific research mindset. The commonality of data, the overall characteristics of the network are hidden in the data network. In the era of big data, IT systems need to change from having data around the processor to processing power revolves around the data. The processing and presentation of data becomes one of the significant evaluation criteria to measure the effectiveness of the model in performing the task. This article focuses on the data processing process and presentation for diagnosing neurological diseases. Based on the experiments, the data is described, presented and processed.

**Key words:** Machine learning, data processing, data description, data presentation.

## Introduction.

With the rapid development of global information science and technology over the Internet, the term "big data" is increasingly familiar to people [1].

In a general sense, Big Data is a collection of data that cannot be sensed, accessed, managed, processed, and served within a tolerable period of time using traditional IT technologies as well as software or hardware tools. In 2008,《Nature》 published a specific journal "Big Data" [2], which introduced the challenges posed by massive data from various aspects. A special edition of 《Science》 on data processing,

"Dealing with data" [3], was launched in 2011 to discuss the challenges posed by the Data Deluge. The aim is to discuss big data applications in biomedicine, Internet, economics, environmental science, supercomputing, etc.

The high integration of the triadic worlds of human, machine, and material has triggered an explosion in the scale of data and a high degree of complexity in data patterns.

Traditional data-centric disciplines, such as medicine, are generating more and more data, which hide great opportunities and values.

Based on the source of data, big data can be divided into two broad categories in a preliminary way: a category from the physical world and another from human society. The former is mainly scientific experimental data or sensor data, while the latter is related to human activities,

especially associated with the Internet [4].

These two types of data vary tremendously in the way they are processed and their targets, with different ways of presentation of the data.

In a general sense, data used to diagnose neurological diseases usually include various types of clinical data, imaging data, physiological data, and molecular data [5], which can help physicians gain a more comprehensive understanding of patients' conditions, perform diagnosis, therapy, and prognosis assessment of neurological diseases, etc., it can also be used to train machine learning models such as neural networks to enable automated disease diagnosis and prediction.

In this article, we took Alzheimer's disease and Parkinson's disease as examples, specifically, the way voice data used to diagnose neurological diseases are processed and presented in machine learning models.

**Data description.**

Parkinson's disease and Alzheimer's disease both have their own public datasets due to being internationally recognized as incurable neurological diseases. Among the voice datasets available in the field of Parkinson's disease diagnosis are the UCI Machine Learning Library [6], the mPower dataset [7], and the Parkinson's Disease Data Set [8].

In the case of the UCI Machine Learning Library, the data used in this dataset was collected from 188 PD patients (107 males and 81 females), aged between 33 and 87 years old, as shown in Figure 1.

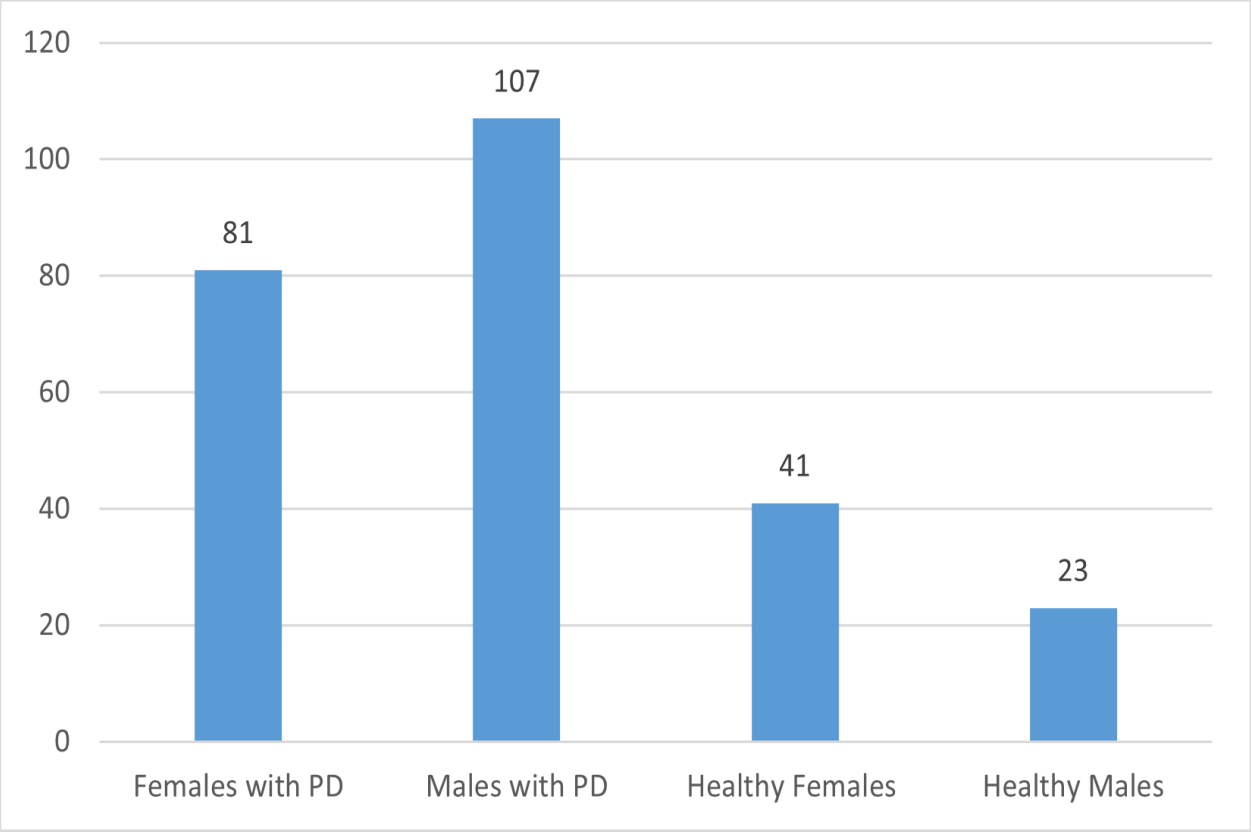



*Figure 1.* Data composition of the UCI Parkinson's dataset

The control group consisted of 64 healthy individuals (23 males and 41 females), aged between 41 and 82 years old. During data collection, the microphone was set to 44.1 KHz with

sustained pronunciation of the vowel /a/ collected from each participant after a physician's examination, repeated three times, and the data was audio format with a .WAC suffix, each sample containing 754 attributes, of which 752 attributes were phonetic features, 1 attribute identified age, and 1 attribute identified gender type.

In the field of Alzheimer's disease diagnosis, there are voice datasets such as DementiaBank [9], Dem@Care Datasets [10], and The ADReSS Challenge [11], a competition dedicated to providing a spontaneous voice benchmark dataset to test the effectiveness of different methodologies for voice recognition in Alzheimer's disease. In the case of The ADReSS Challenge 2020, for example, in order to minimize the risk of bias in the prediction task, the data used in this dataset was collected from 2,122 voice fragments of 78 non-affected individuals and 2,155 voice fragments of 78 patients in audio format with a .WAC suffix, with each voice fragment associated with a corresponding transcript, the participants were in the age range of 50-80 and supported both phonological and semantic analysis of the data.

**Data presentation.**

The presentation of Parkinson's voice data and Alzheimer's voice data in machine learning systems usually requires multiple steps, including data preparation, data training and data evaluation, etc. The visualization of data helps users understand and analyze the data, as illustrated below:

A. Data preparation.

Data preparation includes the steps of data collection, organization, processing and filtering with the aim of ensuring the accuracy and usability of data. The concrete content of data collation is to remove or recover missing values, duplicate data, outliers, erroneous data, etc.

B. Data training

In order for machine learning models to better recognize the data, it is necessary to convert the data into a format suitable for model training, such as converting text into vectors, extracting features from voice, etc. In the voice assessment of Parkinson's, many of which have been shown to have varying degrees of outliers compared to healthy populations, the most commonly used acoustic characteristics include: fundamental frequency (F0) and fundamental frequency variation (vF0), frequency perturbation (Jitter), amplitude perturbation (Shimmer), and harmonic-to-noise ratio (HNR).

In addition, other acoustic metrics such as MFCC (Mel Frequency Cepstral Coefficients), recurrence period density entropy, detrended fluctuation analysis (DFA) ), and VOT (Voice Onset Time) have also gained importance. In the voice assessment of Alzheimer's disease, the variables of diagnostic tests are usually defined according to preclinical symptomatology related to semantic memory and verbal fluency tasks.

In addition, algorithms such as CVT (Cepstral Vowel Transform) algorithm, or acoustic feature sets such as eGeMAPS, ComParE, emobase, etc., or feature extraction tools such as YAAFE (Yet Another Audio Feature Extractor), COVAREP (ComParable Voice Analysis & REpresentation Patterns), OpenSMILE toolkit can be used to help improve the accuracy of the model.

If the features are too redundant or to optimize the network parameters, an algorithm is considered for feature selection, where the most relevant or representative features are selected for the voice data to improve the efficiency and accuracy of the model.

C. Data evaluation.

During the training process, visualization tools can be used to monitor the training effect of the neural network, such as plotting loss function images and making accuracy variation tables, all of which can help to tune the parameters and optimize the network structure to improve the performance of the model.

After the training is completed, in order to evaluate the model, calculating metrics such as accuracy, recall, F1 score helps to compare with the effect of other models or visualize the results using confusion matrix and ROC curve to visually show the accuracy of the network.

**Data processing.**

In machine models, data presentation and processing are usually closely related steps, except that data processing focuses more on processing, extracting, and preprocessing the data to facilitate subsequent analysis, while data presentation focuses more on the mode and content of presenting the data.

The presentation of data can directly affect the subsequent data processing, and the results of data processing can in turn affect the way the data are presented.

Figure 2 shows how the Parkinson's voice data was processed in reference [12].

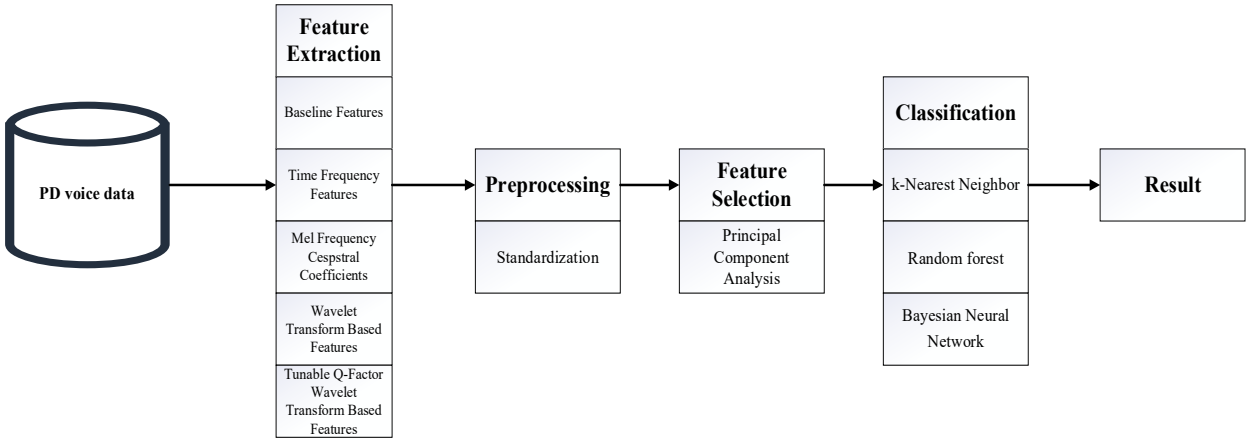



*Figure 2.* Processing of Parkinson's data

**Conclusion.**

Combining the above information, it can be seen that data used to diagnose neurological diseases usually go through several processing steps in machine learning models.

This report describes the presentation and processing of voice data in machine learning models, taking the data of Alzheimer's disease and Parkinson's disease as examples.

Data plays a crucial part in the diagnosis of neurological diseases, safeguarding the reliability, accuracy and integrity of data is a prerequisite for models to make good judgments.

In general, in the future, one needs to overcome the difficulties of data sparsity and redundancy as well as enhance the ability to obtain knowledge from massive and complex data in order to improve the efficiency of research and production so as to move the IT-diagnistics to a new stage of digitization and informatization.

**References**

[1] Zheng, Yili. Research on Big Data Governance in Public Hospitals – A Case Study of Wenzhou E Hospital. Master’s thesis. / Zheng Yili. – Shanghai Normal University. 2021. – 82 c.

[2] Big Data. Nature. – 2008. – 455(7209). –136 c.

[3] Dealing with data. Science. – 2011, – 331(6018). – P. 639–806.

[4] Li Guojie, Cheng. Research Status and Scientific Thinking of Big Data / Li Guojie, Cheng Xueqi. – Journal of the Chinese Academy of Sciences – 2012. – 27(06). – P. 647–657.

[5] Rui, Zhang. Advancing Alzheimer's research: a review of big data promises / Rui Zhang, Gyorgy Simon, Fang Yu. – Int J Med Inform. 2017. – P. 48–56.

[6] Sakar, C.O. A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q-factor wavelet transform. / Sakar, C.O., Serbes, G., Gunduz, A., Tunc, H.C., Nizam, H., Sakar, B.E., Tutuncu, M., Aydin, T., Isenkul, M.E. and Apaydin, H. – Applied Soft Computing, 2019. – P. 255–263.

[7] Brian, M. Bot. The mPower study, Parkinson disease mobile data collected using ResearchKit. / Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, Stephen H. Friend and Andrew D. Trister. . Scientific Data 3, 2016. – 9 p.

[8] Max, A. Little. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease / Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, Lorraine O. Ramig. – Nature Precedings. 2008. – 27 p.

[9] Alyssa, M. Lanzi. DementiaBank: Theoretical rationale, protocol, and illustrative analyses / Alyssa M. Lanzi, Anna K. Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, Matthew L. Cohen. – American Journal of Speech-Language Pathology. 2023. – P. 426– 438.

[10] Karakostas, A. The Dem@Care Experiments and Datasets: a Technical Report. / Karakostas A., Briassouli A., Avgerinakis K., Kompatsiaris I., Tsolaki M. – ArXiv. 2016. – 4 p.

[11] Saturnino, Luz. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. / Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, Brian MacWhinney. - ArXiv, 2020. – 5 c.

[12] Vishniakou, U. A. IT Diagnostics of Parkinson's Disease Based on the Analysis of Voice Markers and Machine Learning. / Vishniakou U. A., Xia YiWei. – Doklady BSUIR. 2023; 3(23).

# ПРЕДСТАВЛЕНИЕ И ОБРАБОТКА ДАННЫХ ДЛЯ ДИАГНОСТИКИ НЕВРОЛОГИЧЕСКИХ ЗАБОЛЕВАНИЙ

***В.А. Вишняков***
*Профессор кафедры инфокоммуникационных технологий, доктор технических наук*

***Ю. Чуйэ***
*Аспирантка кафедры инфокоммуникационных технологий*

***С. Ивей***
*Аспирант кафедры инфокоммуникационных технологий*

*Кафедра инфокоммуникационных технологий*
*Факультет информационной безопасности*
*Белорусский государственный университет информатики и радиоэлектроники,*
*Республика Беларусь*
*E-mail: vish@bsuir.by*

**Аннотация.** Обработка и представление данных становится одним из важных направлений машинного обучения при выполнении задач ИТ-диагностики. Доклад посвящен процессу представления, обработки голосовых данных пациентов и их представлению для диагностики неврологических заболеваний. Наборы данных для обучения нейронных сетей описаны, представлены и обработаны.

**Ключевые слова:** машинное обучение, описание голосовых данных, представление данных, обработка данных.