Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.082841

ARTICLE

Quantitative Profiling of Tabular Biomedical Benchmark Datasets: A Meta-Learning Perspective for Algorithm Selection

Yiyan Zhang[1,*], **Yi Xin**[2] **and Qin Li**[2]

[1]Department of Electronic and Information Engineering, Qingdao Huanghai University, Qingdao, China
[2]Department of Biomedical Engineering, School of Medical Technology, Beijing Institute of Technology, Beijing, China
*Corresponding Author: Yiyan Zhang. Email: zhangyy03@qdhhc.edu.cn

Received: 24 March 2026; Accepted: 22 June 2026; Published: 27 July 2026

ABSTRACT: Medical data has specificity compared to other fields of data, and the description of medical data characteristics is still in a qualitative stage. This study included 293 sub-datasets of 138 independent datasets. First, data preprocessing was performed using methods such as incomplete data removal, inconsistent data normalization, and data integration. Then, the characteristics of 293 research datasets were quantified using 26 indicators in three categories: simple indicators, statistical indicators, and informational indicators. Furthermore, statistical analysis was performed on the above-mentioned quantitative characteristics, and stepwise regression and decision tree methods were used for modeling learning. The characteristics of the biological and medical datasets in the study were compared with those of other fields' datasets. By comparing the results of statistical analysis and learning modeling, the study found that the sample size of medical datasets included in the UCI database analyzed in this paper is small, most within 1000. The harmonic mean or geometric mean of continuous variables is significantly higher than the data from other fields. That is to say, the scope of the continuous variable range is large. This study uses quantitative indicators to describe the characteristics of medical datasets to avoid the decrease in credibility caused by subjective analysis, and lays a foundation for further algorithm applicability research.

KEYWORDS: Characteristics of datasets; medical data; quantitative representation; decision tree

1 Introduction

Biomedical research is poised to continuously produce vast quantities of data in diverse formats. Consequently, the imperative to effectively comprehend and extract insights from this data for novel discoveries is growing [1]. The accessibility of research datasets is fundamental to advancements in health and life sciences. However, the inherent heterogeneity and complexity of medical data present a primary challenge for research data management systems: delivering optimal responses to user search queries [2]. With the continuous development and popularization of wearable devices such as wristbands, more and more individual physical data is being collected, and environmental, meteorological data, and other field data related to health conditions have also received increasing attention [3,4]. Each data subtype presents distinct challenges. Geospatial data is inherently characterized by autocorrelation, whereas electronic health record (EHR) data faces significant hurdles regarding standardization and data quality [5]. There are some problems when dealing with massive data, especially how to analyze the data in a reliable way [6,7]. According to the characteristics of the research data, selecting the applicable method for data analysis is an important step. Some scholars have reached a consensus on the characteristics of medical data.

It is generally believed that medical data has the following typical features: heterogeneity, privacy, and incompleteness [8,9]. The sources of medical data are numerous and varied, and it is bound to cause certain medical data to be different from other medical data. Therefore, some scholars have put forward the following novel viewpoints. They believe that some medical data also have particularity in the following aspects: non-canonical form, low mathematical characteristics, high noise, lack of prior knowledge, and high dimensions [10,11]. Project Imaging-X is a survey of 1000+ open-access medical image datasets; the research reveals the key characteristics of the medical imaging data ecosystem: the data volume is several orders of magnitude smaller than that in general fields, the distribution of modalities and tasks is seriously unbalanced, and the degree of fragmentation is high [12].

At present, there is a lack of systematic research on the characteristics of biomedical datasets, and there are still research gaps in the study of the quantitative description of the characteristics of biomedical datasets. Some early meta-learning studies laid the foundation for the quantification of dataset features. In 1990, Rendell and Cho focused on the proportion of positive examples and their concentration in the sample space, and their influence on concept learning [13]. In 1992, Aha used the sample size, the number of target variable categories, the number and type of attributes, the correlation between attributes and target variables, the distribution, and the number and type of noise as the data features to describe the dataset [14]. The number of dataset features considered in the earlier studies was limited, and the subsequent two Esprit projects extended them. One is the "Comparative testing of statistical and logical learning (StatLog)" project [15]. This initiative employed a comprehensive set of 16 dataset characteristics, including statistical measures such as mean and variance, as well as information entropy. The StatLog project has yielded highly valuable metadata, which has served as a foundational resource in the field of meta-learning for decades. The other one is "A meta-learning assistant for providing user support in machine learning and data mining (METAL)" project [16], which follows the 16 dataset feature descriptors in StatLog. After two European Esprit projects, research in this area was very limited. In 2000, Lim et al. added noise to the data set and analyzed the distribution characteristics of the resulting noisy data [17]. Subsequently, in 2006, Ali and Smith incorporated statistical features derived from the MATLAB toolbox and other sources, thereby expanding the repertoire of dataset characteristics to 31 [18]. In recent years, the focus has shifted toward the proposal of various data complexity measures to effectively characterize datasets [19]. Dataset characterization is a challenging issue, and algorithm recommendation performance largely depends on its effectiveness [20]. With the development of algorithm recommendation research, dataset characterization has also further developed [21,22].

Medical data is unique in comparison with other fields of data. How can the specificity of medical data be expressed in an objective and direct manner? This paper mainly studied the different characteristics of medical data from other fields, focusing on the discovery and induction of medical dataset characteristics, and quantifying the characteristics of medical datasets. The research results lay the foundation for the subsequent applicable method research on analyzing these medical data with different characteristics in a reliable way.

2 Materials and Methods

2.1 Data Source

This study aims to investigate and quantitatively characterize the distinctive attributes of universal medical datasets in contrast to general datasets from other domains, and present the characteristics in a quantitative and objective manner. When selecting the base datasets, the study followed the principle of universality, openness, and minimal intervention, and then the University of California, Irvine (UCI) machine learning repository was chosen as the data source. The UCI repository, created as an ftp archive in 1987, it has been widely used as a primary resource for the empirical analysis of machine learning algorithms

within the research community [23]. It comprises a diverse collection of data, encompassing domain-theoretical data and data generated by synthetic generators. Having been cited over 1000 times, the UCI repository is currently ranked among the top 100 databases in the field of computer science and is widely utilized by students, educators, and researchers worldwide. Most of the datasets in the UCI database have been labeled according to their respective fields. To avoid selection bias that may arise from integrating medical data sets from other sources, we did not select data sets other than those from the UCI database.

2.2 Basic Datasets

According to the scope of the datasets studied in this paper, the public datasets included in the UCI database with classification as the target and capable of being transformed into structured formats through straightforward or moderately complex operations were selected. At the beginning of the study, when the base datasets were selected, the UCI database included and shared 335 datasets. After selection, 138 datasets were included in the study.

Information about the sample size, attributes, missing values, and area can be obtained on the sharing link page. Among the 138 datasets included in the study, 52 datasets were identified as "Life" area, and the other 96 datasets belong to "Business", "Computer", "Financial", "Game", or other area. Since the biological datasets and clinical medical datasets have large differences in data formats, such as rows and columns, the area "Life" is further divided into the two fields of medical science and biology, with the labels "Medical" and "Biology" added. The label of the datasets for the non-biomedical field was set to "General". See Appendix A for details. In addition, statistical analysis was performed on the subsets included in each data set.

2.3 Data Preprocessing

The datasets within the UCI repository originate from diverse domains. As a substantial proportion of these datasets are shared in their original, unaltered formats and were acquired and stored using heterogeneous software, significant variations in data formatting are evident. Consequently, to facilitate the characteristic quantification study, preprocessing was conducted on the selected 138 datasets. Given the objective of this study—to compare the characteristics of universal medical datasets with those of other general datasets—the preprocessing operations were restricted to the necessary minimum to preserve the intrinsic characteristics of each dataset. The data preprocessing mainly includes the following aspects:

(1) Incomplete data: For datasets lacking column headers, attribute meanings were clarified by adding appropriate names, and samples containing missing values were removed to ensure data integrity. Meanwhile, the study set 30% as the threshold and removed the attribute for which the proportion of missing values exceeded 30%. Due to the significant individual differences in medical data, methods such as average value calculation, mode calculation, or formula interpolation may lead to considerable deviations when used to fill in the missing values. Therefore, no operation was carried out to fill in the missing values. In the initial statistical analysis, the proportion of missing values was statistically analyzed using quantitative indicators. The preprocessing operation was conducted to meet the input requirements of the subsequent machine learning modeling.

(2) Inconsistent data: This study compares the descriptive attributes of each dataset and standardizes data that can be unambiguously identified. For discrepancies involving uncertainty, the original data were retained and cross-checked. The data with definite correct label values have been corrected manually, while the uncertain and incorrect label data have been eliminated. If no more information was available, then the sample was removed to reduce noise.

(3) Data integration: Data integration operations are required to associate data from different data sources and store them in a unified dataset.

2.4 Quantitative Index

Based on the analysis and comparison of metrics employed in the European Spirit projects (StatLog and METAL), and aligned with the objective of characterizing the datasets, a total of 26 indices were selected for quantification. These 26 quantitative indices are categorized into three groups: simple, statistical, and informational indicators.

2.4.1 Simple Indicators

(1) Number of variables (P): The number of non-target variables represents the dimension of the dataset.
(2) Sample size (N): The number of samples, which is a representation of the amount of data in the usual sense.
(3) Number of categories (N_class): The number of categories included in the target variable. It is used to distinguish whether a classification task is "binary classification" or "multiple classification".
(4) Ratio of largest class (R_largest): The ratio of the number of the largest category samples to the sample size reflects the balance of the dataset.
(5) Ratio of least class (R_least): The ratio of the number of the least category samples to the sample size, together with the R_largest, reflects the balance of the dataset.
(6) Ratio of binary variable (R_binary): The ratio of the number of binary variables to the number of total attribute variables, and reflects the data structure of the dataset.
(7) Ratio of discrete variable (R_discrete): The ratio of the number of discrete variables to the number of total attribute variables also reflects the data structure of the dataset.
(8) Ratio of continuous variable (R_continuous): The ratio of the number of continuous variables to the total number of attribute variables reflects the data structure of the dataset, together with R_binary and R_discrete.
(9) Ratio of missing values (R_missing): The ratio of the number of samples containing missing values to the sample size is a representation of the integrity of the dataset.

2.4.2 Statistical Indicators

(1) Geometric mean (Geomean): It is the nth root of the product of n variable values, and it is less affected by extreme values than the arithmetic mean.
(2) Harmonic mean (Harmean): It is the reciprocal of the arithmetic mean of the reciprocals of each variable, and is also known as the reciprocal average.
(3) Trim mean (Trimean): It involves removing a certain percentage of data from both ends of the dataset, and then calculating the average of the remaining data.
(4) Percentile (Prctile): After sorting a set of data in ascending order, calculate the corresponding cumulative percentiles. The value of the data corresponding to a certain percentile is the Prctile value of that percentile.
(5) Mean absolute deviation (MAD): It is the average of the absolute deviations of each data point from the average value, reflecting the degree of dispersion of the data.
(6) Variance (Var): It is the average of the squares of the differences between the data and the average value, and it can be used to measure the magnitude of the sample's fluctuation.
(7) Standard deviation (Std): It is the arithmetic square root of the variance. Variance and standard deviation are the most important and commonly used indicators for measuring the dispersion trend of data.

(8) Mean of absolute correlation coefficient (MAr): The correlation coefficient is used to study the degree of the linear correlation between variables. The MAr is obtained by taking the absolute value of each correlation coefficient and then calculating the average.
(9) Interquartile range (IQR): It is the difference between the upper quartile and the lower quartile.
(10) Index of dispersion (D): It is defined as the ratio of variance to the mean, and is used to indicate whether the data are clustered or dispersed compared to the standard statistical model.
(11) Skewness: A characterization of the degree of asymmetry of the probability distribution density curve compared to the average value. It reflects the direction and degree of skewness in the data distribution.
(12) Kurtosis: The representation of the height of the peak of the probability density distribution curve at the mean value position.

2.4.3 Informational Indicators

(1) Mean entropy of attribute variables (ME_V): Entropy is a measure of randomness in a variable. First, calculate the information entropy value of each discrete variable, and then the average value of the variable's information entropy can be obtained.
(2) Entropy of class (E_C): Similar to information entropy of non-target variables, the E_C represents the randomness of the distribution of each category.
(3) Mean mutual entropy of class and attribute variables (MME_CV): The mutual entropy reflects the interdependence between two variables.
(4) Equivalent number of variables (ENV): It is the ratio of "E_C" to "MME_CV".
(5) Noise-signal ratio (NSR): It represents the amount of irrelevant information contained in the dataset. If the NSR is larger, it indicates that the dataset contains a considerable amount of noise. In such cases, the data can be compressed without affecting the performance of the model.

The calculation formulas for the above statistical and information indicators can be found in the appendix of Ref. [17].

2.5 Analytical Methods

Some datasets contain several independent sub-datasets. Therefore, a total of 293 independent sub-datasets were selected as the basic datasets in the study. According to the types of variables in the dataset, the 293 datasets included in the study were divided into three categories: mixed variable datasets, discrete variable datasets, and continuous variable datasets. The calculation results of these three quantitative indicators were integrated, and category labels ("medicine", "biology", and "general") were added for each dataset. Thus, a new metadata set for further study was formed.

Firstly, calculate the variance for each group to compare the homogeneity of variances, and then perform one-way analysis of variance for each quantitative indicator to compare the differences between different groups.

The study included 34 datasets with discrete variables. Due to the limited sample size, it is not suitable for modeling to learn the dataset characteristics. Consequently, modeling analysis was restricted to mixed and continuous variable datasets. Using area classification as the target variable and 26 quantitative features as attribute variables, models were constructed via stepwise regression and the C4.5 decision tree algorithm to analyze the metadata. The flowchart of the process for quantifying the characteristics of the medical data set in this study is shown in Fig. 1. For the discrete variable dataset, the results of the one-way analysis of variance indicated that only the quantification characteristic index "MME_CV" showed a difference between groups. The F-value is 5.5685, and the p-value is 0.008982. The statistical significance is extremely significant.

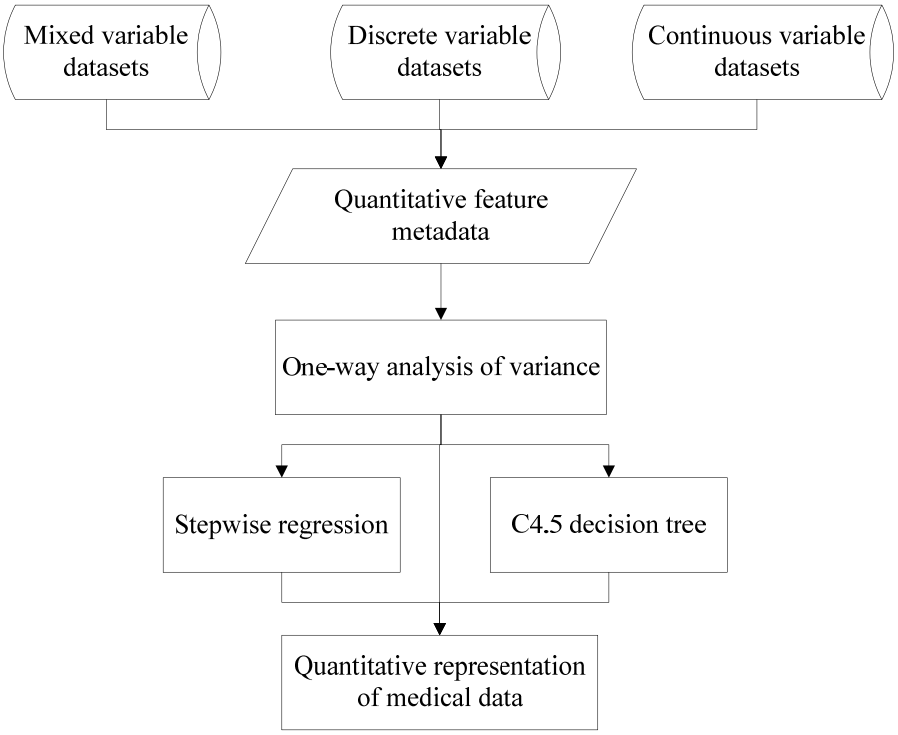

Figure 1: Medical dataset feature quantization flow chart.

3 Results

There are 37 datasets that belong to medical datasets. Table 1 summarizes the information of the dataset with the area label "Medical".

Table 1: Summary of the medical dataset information included in the paper.

Name of Dataset	Sample Size	Number of Attributes	Missing Value	Sub Dataset
Arrhythmia	452	279	Yes	1
Audiology (Standardized)	226	69	Yes	1+1
Breast Cancer Wisconsin (Original)	699	10	Yes	1
Breast Cancer Wisconsin (Prognostic)	198	34	Yes	1
Breast Cancer Wisconsin (Diagnostic)	569	32	No	1
Contraceptive Method Choice	1473	9	No	1
Dermatology	366	33	Yes	1
Echocardiogram	132	12	Yes	1
Haberman's Survival	306	3	No	1
Heart Disease	303	75	Yes	4
Hepatitis	155	19	Yes	1
Horse Colic	368	27	Yes	1+1
Liver Disorders	345	7	No	1
Lung Cancer	32	56	Yes	1
Pima Indians Diabetes	768	8	Yes	1

(Continued)

Table 1 (continued)

Name of Dataset	Sample Size	Number of Attributes	Missing Value	Sub Dataset
SPECT Heart	267	22	No	1+1
SPECTF Heart	267	44	No	1+1
Thyroid Disease	7200	21	N/A	10+7&1
Statlog (Heart)	270	13	No	1
Mammographic Mass	961	6	Yes	1
Arcene	900	10,000	N/A	1+1+1
Parkinsons	197	23	N/A	1
Acute Inflammations	120	6	No	1
Breast Tissue	106	10	N/A	1
Cardiotocography	2126	23	N/A	1
Localization Data for Person Activity	164,860	8	N/A	1
Vertebral Column	310	6	N/A	2
ILPD (Indian Liver Patient Dataset)	583	10	N/A	1
Fertility	100	10	N/A	1
Daphnet Freezing of Gait	237	9	N/A	17
EEG Eye State	14,980	15	N/A	1
Thoracic Surgery Data	470	17	N/A	1
LSVT Voice Rehabilitation	126	309	N/A	1
Diabetes 130-US Hospitals for Years 1999–2008	100,000	55	Yes	1
Parkinson Speech Dataset with Multiple Types of Sound Recordings	1040	26	N/A	1+1
Diabetic Retinopathy Debrecen Data Set	1151	20	N/A	1
Chronic_Kidney_Disease	400	25	Yes	1

*In the last column of Table 1, "1+1" indicates that there is one training set and one test set; "1+1+1" indicates that there is one training set, one test set, and one validation set; "10+7&1" indicates that there are two kinds of subsets with different variable categories, one of which has 10 datasets and 7 of which have test sets, and the other has one dataset.

Table 2 summarizes the information of 17 biological datasets with the area label "Biology".

Table 2: Summary of the biology dataset information included in the paper.

Name of Dataset	Sample Size	Number of Attributes	Missing Value	Sub Dataset
Abalone	4177	8	No	1
Covertype	581,012	54	No	1
Ecoli	336	8	No	1
Iris	150	4	No	1
Molecular Biology (Promoter Gene Sequences)	106	58	No	1
Molecular Biology (Splice-Junction Gene Sequences)	3190	61	No	1
Mushroom	8124	22	Yes	1

(Continued)

Table 2 (continued)

Name of Dataset	Sample Size	Number of Attributes	Missing Value	Sub Dataset
Soybean (Large)	307	35	Yes	4+2
Soybean (Small)	47	35	No	1
Yeast	1484	8	No	1
Zoo	101	17	No	1
PubChem Bioassay Data	N/A	N/A	N/A	21+21
One-Hundred Plant Species Leaves Data Set	1600	64	N/A	3
QSAR Biodegradation	1055	41	N/A	1
Wilt	4889	6	N/A	1+1
Forest Type Mapping	326	27	N/A	1+1
Mice Protein Expression	1080	82	Yes	1

3.1 Dataset Characteristic Quantization Metadata

For different types of datasets, different combinations of quantitative indicators were used to calculate dataset characteristics. The following is a detailed description of the three categories of datasets.

3.1.1 Mixed Variable Dataset

The mixed variable dataset contains both discrete and continuous variables, so all three categories 26 quantified indicators mentioned above were calculated. Take the dataset "Echocardiogram" as an example. First, nine simple quantitative indicators were counted and calculated. The results are shown in Table 3.

Table 3: Nine simple indicators of the dataset "Echocardiogram".

Simple Indicator	Quantitative Value
P	7
N	61
N_class	2
R_largest	0.7213
R_least	0.2787
R_binary	0.1429
R_discrete	0.1429
R_continous	0.8571
R_missing	0.5379

Then, the correlation coefficients of the continuous variables and the other 11 statistical quantitative indicators were calculated. The correlation coefficient matrix is shown in Table 4, and the calculation results of the remaining 11 statistical indicators are shown in Table 5.

After taking the absolute values of all the non-autocorrelation coefficients of the correlation coefficient matrix, the average value is calculated to obtain MAr. The 12 statistical quantitative indicators were summarized as shown in Table 6. Then the five informational quantitative indicators were calculated and summarized in Table 7.

Table 4: Correlation coefficient matrix of seven continuous variables of the dataset "Echocardiogram".

	Age-at-Heart-Attack	Fractional-Shortening	epss	lvdd	Wall-Motion-Score	Wall-Motion-Index
age-at-heart-attack	1	–	–	–	–	–
fractional-shortening	−0.1188	1	–	–	–	–
epss	0.0781	−0.4011	1	–	–	–
lvdd	0.1948	−0.3763	0.5686	1	–	–
wall-motion-score	−0.0564	−0.2741	0.342	0.1848	1	–
wall-motion-index	0.0885	−0.3279	0.4436	0.3426	0.7718	1

Table 5: 11 statistical indicator results of the dataset "Echocardiogram".

	Geomean	Harmean	MAD	Var	Std	IQR	D	Skewness	Kurtosis	Trimean	Percentile
age-at-heart-attack	63.9206	63.3703	7.1780	75.6478	8.6976	12.2500	1.1731	0.4613	2.6439	64.4841	78.0000
fractional-shortening	0.1866	0.1306	0.0816	0.0114	0.1069	0.1225	0.0522	0.7377	4.6606	0.2188	0.3464
epss	0.0000	0.0000	5.3479	54.1099	7.3559	8.1250	4.3796	1.2060	5.4323	12.3550	21.5800
lvdd	4.7694	4.7091	0.6549	0.5999	0.7745	1.1475	0.1242	0.2630	2.3929	4.8302	5.8076
wall-motion-score	14.4912	13.7376	3.7922	29.4151	5.4236	4.7925	1.9227	1.6656	7.7102	15.2990	22.7000
wall-motion-index	1.3460	1.2954	0.3543	0.2017	0.4492	0.6288	0.1435	1.2429	4.2773	1.4062	2.0780

Table 6: Twelve statistical indicators of the dataset "Echocardiogram".

Statistical Indicators	Quantitative Value
Geomean	14.1190
Harmean	13.8738
MAD	2.9015
Var	26.6643
Std	3.8013
MAr	0.0487
IQR	4.5110
D	1.2992
Skewness	0.9294
Kurtosis	4.5195
Trimean	16.4322
Percentile	9.3796

Table 7: Five information indicators of the dataset "Echocardiogram".

Information Indicators	Quantitative Value
ME_V	0.2049
E_C	0.2570
MME_CV	−0.0068
ENV	−37.7941
NSR	−31.1324

Similar to the dataset "Echocardiogram", nine simple, 12 statistical, and five informative quantitative indicators of the remaining 101 mixed variable datasets included in the study were calculated (the datasets refer to separate subsets, training set, test set, and validation set combined into one dataset, the same below). The results of each dataset were summarized to form a new meta-dataset.

3.1.2 Discrete Variable Dataset

The non-target variables of the discrete variable dataset only contain discrete variables, so only the above nine simple indicators and five informational indicators need to be used.

3.1.3 Continuous Variable Dataset

In the continuous variable dataset, except that the target variable is a discrete variable, the rest of the attribute variables are continuous variables. Therefore, nine simple indicators, 12 statistical indicators, and the informational indicator "E_C" were calculated and summarized.

3.2 Statistical Analysis Results

After the preliminary analysis of variance, it can be found that the medical datasets have the following features compared to other datasets. The reflection of these characteristics in the quantitative indicators is shown in Fig. 2.

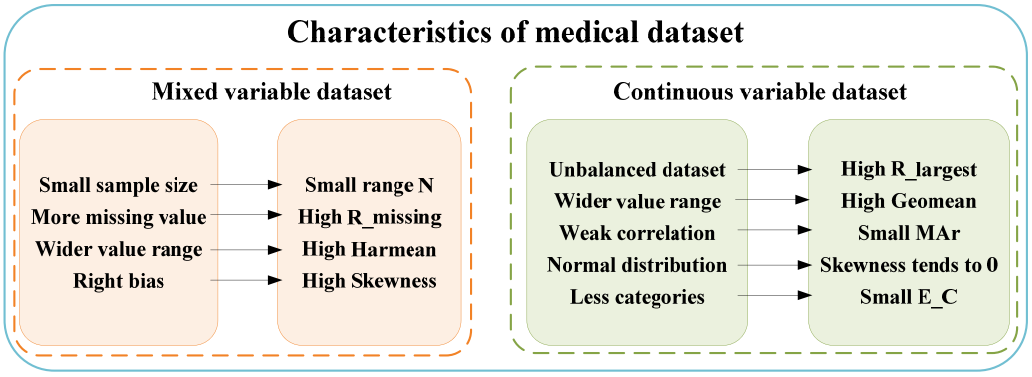

Figure 2: Statistical analysis results of medical dataset characteristics.

As shown in Fig. 2, the mixed-variable medical datasets have certain characteristics, such as a small sample size, more missing values, a wide numerical range, and right-biased data. For the continuous variable datasets, the medical datasets have characteristics such as imbalance, wide numerical range, weak correlation, most of which are subject to normal distribution, and fewer target variable categories.

According to preliminary statistical results, the MME_CV in the discrete variable medical datasets is less than −0.1, while the MME_CV in general datasets is mostly in the range of (−0.25, −0.03). Since only 34 discrete variable datasets were collected in the study and only 5 medical datasets were included, the characteristic recognition of discrete variable medical datasets was insufficient, and some features may be missed.

3.3 Modeling Results

3.3.1 Mixed Variable Dataset

Taking "area" as the target variable, the 26 quantitative characteristics were used as attribute variables to establish a stepwise regression model.

From the stepwise regression model constructed for mixed variable datasets, the characteristics of medical datasets differ from those of other datasets mainly in terms of P, R_binary, Harmean, Std, MAr, IQR, D, Skewness, Prctile, and ME_V.

The C4.5 algorithm was used to construct a decision tree model, as shown in Fig. 3. The accuracy of this model is 96.81%.

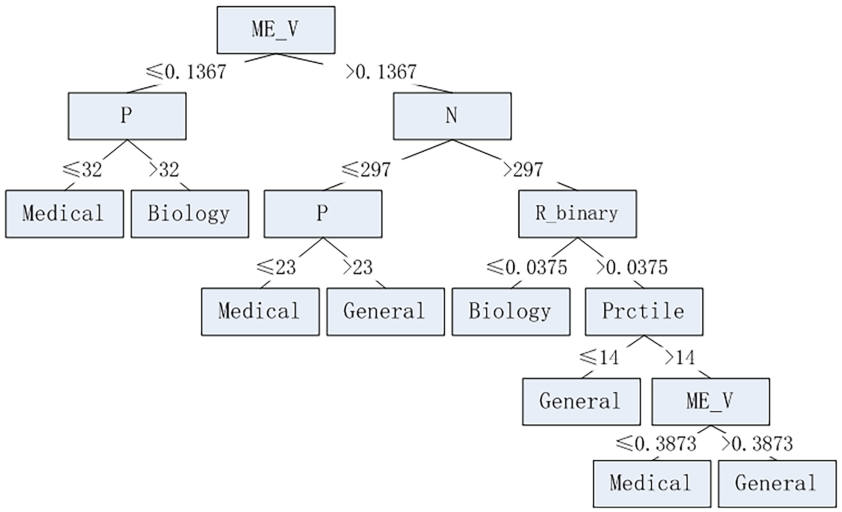

Figure 3: Decision tree for mixed variable datasets.

As shown in Fig. 3, mixed-variable medical datasets have a small sample size, low dimension, and small mutual entropy.

The specific differences can be seen from the preliminary statistics and the decision tree in Fig. 3, and summarized as follows.

(1) P: In most medical datasets, the value of P is all less than 32, while in other data sets, the value of P is higher. Especially in 75% of the biological datasets, the value of P reaches 100 or higher.
(2) N: The values of N in the medical datasets are mostly within 297, while the values of N in the general datasets are larger, most of which reach more than 10,000.
(3) R_binary: The R_binary values of the medical datasets are slightly higher than those of the general datasets. While the R_binary of the 75% biological datasets is significantly higher than that of the medical dataset and the general dataset, the proportion is as high as more than 70%.
(4) Harmean: Medical datasets mostly fall within the range of 0–40, most biological datasets in the range 0–15, and general datasets tend to be closer to 0.
(5) Skewness: The skewness values of medical datasets are mainly distributed in the 0–4 interval, that is, the data distribution of each attribute value is mostly right-biased. While the skewness values of the general datasets are basically around 0, the data in the general datasets are all basically distributed normally.
(6) ME_V: The ME_V in medical datasets mainly concentrates within the range of 0.1–0.35, which is significantly higher than the biological datasets and lower than the general datasets. The ME_V of the

biological datasets is the smallest, most of which are less than 0.1, indicating that discrete variables in the biological datasets usually contain fewer categories.

3.3.2 Continuous Variable Dataset

Taking "area" as the target variable, and selecting 22 quantitative characteristics as attribute variables, a model was constructed using the stepwise regression method.

The stepwise regression model constructed based on the continuous variable dataset reveals that the main differences in the characteristics of the medical datasets differ from the other datasets mainly in the P, N, N_class, R_least, R_missing, Geomean, Harmean, MAD, MAr, IQR, Skewness, Trimean, and E_C.

The C4.5 algorithm was applied to build a decision tree model, and the resulting decision tree is shown in Fig. 4. The accuracy of the model is 97.99%.

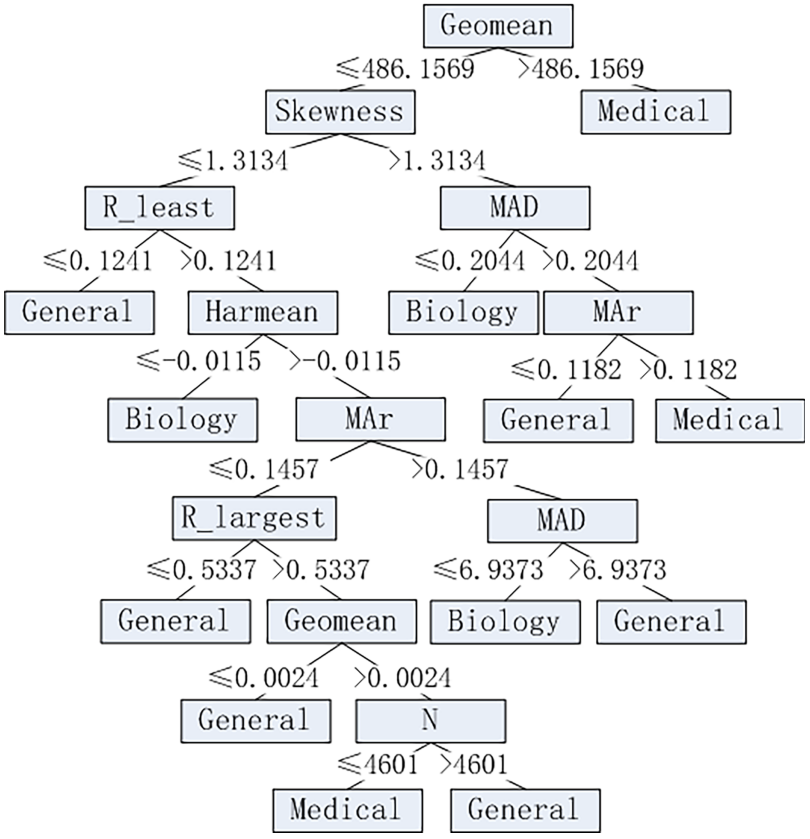

Figure 4: Decision tree for continuous variable datasets.

According to Fig. 4, continuous-variable medical datasets have relatively high Geomean and relatively small sample size.

4 Discussion

The specific differences drawn by the preliminary statistics and the decision tree in Fig. 4 are as follows.

(1) N: The value of N in the medical datasets is smaller than that in the general datasets, and nearly half of the medical datasets analyzed in this paper have a sample size of less than 1000.

(2) N_class: Most medical datasets are binary-class tasks. The number of classes in general datasets is slightly more than that in medical datasets, but no more than 15 as usual; the number of classes in biological datasets spans a wide range, with the maximum reaching up to 100.
(3) Geomean: Nearly 60% of the medical datasets have a geomean for each continuous variable greater than 486.1569, with a very large range, while the geomean of the biological datasets and general datasets is lower than 486.1569.
(4) MAr: The medical datasets have the smallest MAr, most of which are below 0.05; the biological datasets have the largest span, mainly within the 0–0.2 interval; the general datasets have the value between the two categories, mainly in the 0–0.1 range. Compared with the other two area datasets, the correlation between the research variables in the medical dataset is relatively weak.
(5) Skewness: The skewness of the medical datasets is basically around 0, symmetrically distributed; the skewness of the general datasets is slightly larger, but mostly within 1. It can be seen that most of the variables in the two area datasets obey a normal distribution; the skewness of the biological datasets is significantly higher than that of the above two types of datasets, indicating that the distribution of the attribute variables is mostly right-biased.
(6) E_C: The E_C value of the medical datasets is small, and the span is small, mostly within the range of 0.2–0.5. The general datasets have relatively larger values and a wider range; the upper and lower quartile range is 0.3–0.7. The E_C span of the biological dataset is the largest, and more than half of the datasets are greater than 0.7, which is significantly higher than the aforementioned two types of datasets.

Based on the above statistical analysis and modeling results, the medical datasets included in the UCI database analyzed in this paper typically have a small sample size, mostly within 1000. Compared with non-medical datasets, other data mostly come from industrial fields. With the advancement of industrial information technology, industrial data is becoming increasingly standardized and easy to obtain in batches, so the sample size is larger than that of medical data. The medical data are mostly collected by medical research institutions. Although the sample size is relatively small, certain collection scheme rules have been formulated before collection, and the medical field is mostly used for experimental control studies, so the small sample size has little impact on the reliability of such studies. In addition, medical research itself is greatly influenced by individual differences, and there are differences between different hospitals or medical paths, so the generalization of some research results is limited. At the same time, the influence of small sample size is being improved by the development trend of federated learning and multi-center research.

Based on the above statistical analysis and modeling results, the range of continuous variables in the medical datasets is significantly larger than that of datasets in other fields. Because most of the continuous variables in the medical field are laboratory test results, such as biochemical test results, and the normal value ranges of these tests vary from single digits to hundreds of thousands. With the standardization in the industrial field, data deviation is usually very small, so the range of variables is relatively small, showing a centralized trend. These differences are correctly expressed by quantitative indicators and can be explained.

In summary, whether it is a mixed variable dataset or a continuous variable dataset, medical datasets exhibit common characteristics that are different from those of general datasets in other fields, namely, a small sample size and large and high average values of the variable values. Habchi et al. reviewed and discussed the efficiency of deep learning and large language model technologies in cancer diagnosis, analyzed key issues such as data imbalance, and proposed possible solutions. The research indicates that limited dataset sizes and imbalanced data are particularly significant challenges in the field of medical cancer research [24], which are consistent with some of the results presented in this paper.

Massive data is like a treasure house, containing rich and valuable information. However, how to mine out potential knowledge requires us to formulate a more reasonable research plan according to the actual situation and relevant background knowledge, select suitable mining methods for analysis and exploration, and then discuss the mining results from a professional perspective. Finally, get the correct law or conclusion that conforms to reality. The review article by Habchi et al. summarizes various studies on the application of artificial intelligence methods (especially those using the Transformer model) in the diagnosis of thyroid cancer. A new classification system for these methods is introduced, based on the AI algorithms, framework objectives, and the computing environments employed. Additionally, it reviews and compares the characteristics of available thyroid cancer datasets. The research indicates that variables such as data discrepancies, lack of complete data, and evolving clinical conditions can substantially influence outcomes [25]. The review by Ma et al. delved deeply into the issue of data discrepancies, stating that the differences in equipment and collection protocols among various medical institutions result in significant heterogeneity in the data, and the distribution of patient disease stages is also uneven [26]. Xu et al. proposed a "Disease Embedding" method, which encodes sparse and heterogeneous medical records into high-dimensional vectors [27]. This method can quantify the deep similarities and correlations among diseases, providing a new computational framework for uncovering the patterns of hidden diseases. Zang et al. demonstrated how to integrate multimodal data such as clinical features, CT images, and complete pathological section images. The study proved that compared to a single modality, the multimodal AI model can perform risk stratification more accurately and effectively overcome the "risk misjudgment" problem caused by the single data dimensionality of traditional clinical tools [28]. Based on the quantitative characteristics studied in the paper, we aim to identify a data mining algorithm suitable for the current dataset, and then obtain the law of medical significance and apply it to the clinic. We have successfully applied the 26 indicators to the specific research issues in the prostate tumor database and diabetes database of a large tertiary hospital, two large-scale real-world datasets. The validation results confirmed that the proposed decision tree model maintains robust performance on the external data. According to the dataset characteristics, the corresponding algorithm is developed. The study can provide data characteristic analysis support for the algorithm development and parameter selection and optimization research [29,30]. In addition, the study result can also be used as a part of the algorithm selection process for the doctor's scientific research platform, providing theoretical support for the algorithm selection process.

5 Conclusions

Medical data covers a variety of categories, and the data characteristics of each category are more or less different. Currently, there are relatively few systematic studies on the characteristics of biomedical datasets. This paper studies the characteristics of medical datasets and quantifies them, which is conducive to a more objective understanding and representation of medical datasets. Quantitative features of 293 UCI sub-datasets were calculated using combinations of 26 quantitative indicators and analyzed by statistical and modeling methods to achieve a quantitative representation of the biomedical dataset characteristics. The datasets used in this study are limited to publicly available datasets that can be transformed into structured data through simple or slightly complex operations, and are specifically designed for classification tasks. Image data and unstructured text data were not included in the study for comparative analysis. In subsequent research, the applicability of data mining methods on datasets with various characteristics and the improvement of algorithms will be explored based on the quantitative characteristics.

Acknowledgement: None.

Funding Statement: This research was funded by the Qingdao Huanghai University Doctoral Research Foundation Project, grant number 2023boshi02, and Qingdao Huanghai University scientific research project, grant number KYH2025001.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Qin Li and Yi Xin; methodology, Yiyan Zhang; software, Yiyan Zhang; validation, Yiyan Zhang; formal analysis, Yiyan Zhang; writing—original draft preparation, Yiyan Zhang; writing—review and editing, Yi Xin and Qin Li; visualization, Yiyan Zhang; funding acquisition, Yiyan Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available in Appendix A.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1: Detaileds information of the medical dataset included in the paper.

Dataset Name	Access Link	Citation	View
Arrhythmia	https://archive.ics.uci.edu/dataset/5/arrhythmia	19	38,316
Audiology (Standardized)	https://archive.ics.uci.edu/dataset/8/audiology+standardized	0	4665
Breast Cancer Wisconsin (Original)	https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original	6	154,455
Breast Cancer Wisconsin (Prognostic)	https://archive.ics.uci.edu/dataset/16/breast+cancer+wisconsin+prognostic	0	26,700
Breast Cancer Wisconsin (Diagnostic)	https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic	37	681,396
Contraceptive Method Choice	https://archive.ics.uci.edu/dataset/30/contraceptive+method+choice	3	18,582
Dermatology	https://archive.ics.uci.edu/dataset/33/dermatology	14	33,740
Echocardiogram	https://archive.ics.uci.edu/dataset/38/echocardiogram	0	16,743
Haberman's Survival	https://archive.ics.uci.edu/dataset/43/haberman+s+survival	0	17,983
Heart Disease	https://archive.ics.uci.edu/dataset/45/heart+disease	64	1,122,696
Hepatitis	https://archive.ics.uci.edu/dataset/46/hepatitis	12	50,681
Horse Colic	https://archive.ics.uci.edu/dataset/47/horse+colic	0	22,012
Liver Disorders	https://archive.ics.uci.edu/dataset/60/liver+disorders	2	52,387
Lung Cancer	https://archive.ics.uci.edu/dataset/62/lung+cancer	4	64,606

(Continued)

Table A1 (continued)

Dataset Name	Access Link	Citation	View
Pima Indians Diabetes	Not available now	–	–
SPECT Heart	https://archive.ics.uci.edu/dataset/95/spect+heart	8	10,699
SPECTF Heart	https://archive.ics.uci.edu/dataset/96/spectf+heart	0	6700
Thyroid Disease	https://archive.ics.uci.edu/dataset/102/thyroid+disease	18	97,375
Statlog (Heart)	https://archive.ics.uci.edu/dataset/145/statlog+heart	16	38,350
Mammographic Mass	https://archive.ics.uci.edu/dataset/161/mammographic+mass	4	22,004
Arcene	https://archive.ics.uci.edu/dataset/167/arcene	0	9082
Parkinsons	https://archive.ics.uci.edu/dataset/174/parkinsons	15	101,791
Acute Inflammations	https://archive.ics.uci.edu/dataset/184/acute+inflammations	2	12,236
Breast Tissue	https://archive.ics.uci.edu/dataset/192/breast+tissue	0	7403
Cardiotocography	https://archive.ics.uci.edu/dataset/193/cardiotocography	5	34,623
Localization Data for Person Activity	https://archive.ics.uci.edu/dataset/196/localization+data+for+person+activity	0	7023
Vertebral Column	https://archive.ics.uci.edu/dataset/212/vertebral+column	1	25,992
ILPD (Indian Liver Patient Dataset)	https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset	1	45,640
Fertility	https://archive.ics.uci.edu/dataset/244/fertility	1	33,878
Daphnet Freezing of Gait	https://archive.ics.uci.edu/dataset/245/daphnet+freezing+of+gait	0	12,572
EEG Eye State	https://archive.ics.uci.edu/dataset/264/eeg+eye+state	0	27,302
Thoracic Surgery Data	https://archive.ics.uci.edu/dataset/277/thoracic+surgery+data	2	14,637
LSVT Voice Rehabilitation	https://archive.ics.uci.edu/dataset/282/lsvt+voice+rehabilitation	0	3130
Diabetes 130-US Hospitals for Years 1999–2008	https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008	1	159,389
Parkinson Speech Dataset with Multiple Types of Sound Recordings	https://archive.ics.uci.edu/dataset/301/parkinson+speech+dataset+with+multiple+types+of+sound+recordings	0	17,227

(Continued)

Table A1 (continued)

Dataset Name	Access Link	Citation	View
Diabetic Retinopathy Debrecen Data Set	https://archive.ics.uci.edu/dataset/329/diabetic+retinopathy+debrecen	1	18,576
Chronic_Kidney_Disease	https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease	0	122,283

*The citations and views come from access link statistics as of 25 May 2026.

Table A2: Detailed information of the biology dataset included in the paper.

Dataset Name	Access Link	Citation	View
Abalone	https://archive.ics.uci.edu/dataset/1/abalone	57	215,968
Covertype	https://archive.ics.uci.edu/dataset/31/covertype	47	77,542
Ecoli	https://archive.ics.uci.edu/dataset/39/ecoli	6	24,908
Iris	https://archive.ics.uci.edu/dataset/53/iris	352	1,395,737
Molecular Biology (Promoter Gene Sequences)	https://archive.ics.uci.edu/dataset/67/molecular+biology+promoter+gene+sequences	0	10,125
Molecular Biology (Splice-Junction Gene Sequences)	https://archive.ics.uci.edu/dataset/69/molecular+biology+splice+junction+gene+sequences	0	8672
Mushroom	https://archive.ics.uci.edu/dataset/73/mushroom	77	28,251
Soybean (Large)	https://archive.ics.uci.edu/dataset/90/soybean+large	0	22,385
Soybean (Small)	https://archive.ics.uci.edu/dataset/91/soybean+small	0	8114
Yeast	https://archive.ics.uci.edu/dataset/110/yeast	19	34,687
Zoo	https://archive.ics.uci.edu/dataset/111/zoo	29	73,817
PubChem Bioassay Data	https://archive.ics.uci.edu/dataset/209/pubchem+bioassay+data	0	2059
One-Hundred Plant Species Leaves Data Set	https://archive.ics.uci.edu/dataset/241/one+hundred+plant+species+leaves+data+set	0	7169
QSAR Biodegradation	https://archive.ics.uci.edu/dataset/254/qsar+biodegradation	0	8408
Wilt	Not available now	–	–
Forest Type Mapping	https://archive.ics.uci.edu/dataset/333/forest+type+mapping	0	6142
Mice Protein Expression	https://archive.ics.uci.edu/dataset/342/mice+protein+expression	1	19,612

*The citations and views come from access link statistics as of 25 May 2026.

References

1. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's big data to knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014;21(6):957–8. doi:10.1136/amiajnl-2014-002974.
2. Teodoro D, Mottin L, Gobeill J, Gaudinat A, Vachon T, Ruch P. Improving average ranking precision in user searches for biomedical research datasets. Database. 2017;2017:bax083. doi:10.1093/database/bax083.
3. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big data for health. IEEE J Biomed Heal Inform. 2015;19(4):1193–208. doi:10.1109/JBHI.2015.2450362.
4. Fang R, Pouyanfar S, Yang Y, Chen SC, Iyengar SS. Computational health informatics in the big data age: a survey. ACM Comput Surv. 2017;49(1):1–36. doi:10.1145/2932707.
5. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Heal. 2018;39(1):95–112. doi:10.1146/annurev-publhealth-040617-014208.
6. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1(1):2. doi:10.1186/2196-1115-1-2.
7. Fathima AS, Basha SM, Ahmed ST, Mathivanan SK, Rajendran S, Mallik S, et al. Federated learning based futuristic biomedical bigdata analysis and standardization. PLoS One. 2023;18(10):e0291631. doi:10.1371/journal.pone.0291631.
8. Azuaje F. Review of "mining imperfect data" by Ronald K. Pearson. BioMedical Eng OnLine. 2005;4(1):43. doi:10.1186/1475-925X-4-43.
9. Vadovský MPJ. Use and characteristics of medical data mining methods. In: Proceedings of the Scientific Conference of Young Researchers; 2016 Apr 29–30; Baku, Azerbaijan. p. 1–5.
10. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med. 2002;26(1–2):1–24. doi:10.1016/s0933-3657(02)00049-0.
11. Lee CH, Yoon HJ. Medical big data: promise and challenges. Kidney Res Clin Pract. 2017;36(1):3–11. doi:10.23876/j.krcp.2017.36.1.3.
12. Project Imaging-X Contributors. Project imaging-X: a survey of 1000+ open-access medical imaging datasets for foundation model development. [cited 2025 Jan 1]. Available from: https://github.com/uni-medical/Project-Imaging-X/blob/main/project-imaging-x_dataset-survey.pdf.
13. Rendell L, Cho H. Empirical learning as a function of concept character. Mach Learn. 1990;5(3):267–98. doi:10.1023/A:1022651406695.
14. Aha DW. Generalizing from case studies: a case study. In: Machine learning proceedings 1992. Amsterdam, the Netherlands: Elsevier; 1992. p. 1–10.
15. King RD, Feng C, Sutherland A. StatLog: comparison of classification algorithms on large real-world problems. Appl Artif Intell. 1995;9(3):289–333. doi:10.1080/08839519508945477.
16. Smith-Miles KA. Cross-disciplinary perspectives on meta-learning for algorithm selection. ACM Comput Surv. 2009;41(1):1–25. doi:10.1145/1456650.1456656.
17. Lim TS, Loh WY, Shih YS. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn. 2000;40(3):203–28. doi:10.1023/A:1007608224229.
18. Ali S, Smith KA. On learning algorithm selection for classification. Appl Soft Comput. 2006;6(2):119–38. doi:10.1016/j.asoc.2004.12.002.
19. Nojima Y, Nishikawa S, Ishibuchi H. A meta-fuzzy classifier for specifying appropriate fuzzy partitions by genetic fuzzy rule selection with data complexity measures. In: Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011); 2011 Jun 27–30; Taipei, Taiwan. p. 264–71.
20. Wang G, Song Q, Zhu X. An improved data characterization method and its application in classification algorithm recommendation. Appl Intell. 2015;43(4):892–912. doi:10.1007/s10489-015-0689-3.
21. Song Q, Wang G, Wang C. Automatic recommendation of classification algorithms based on data set characteristics. Pattern Recognit. 2012;45(7):2672–89. doi:10.1016/j.patcog.2011.12.025.
22. Roy A, Cruz RMO, Sabourin R, Cavalcanti GDC. Meta-learning recommendation of default size of classifier pool for META-DES. Neurocomputing. 2016;216:351–62. doi:10.1016/j.neucom.2016.08.013.

23. Markelle K, Rachel L, Kolby N. The UCI machine learning repository. [cited 2025 Jan 1]. Available from: https://archive.ics.uci.edu/.
24. Habchi Y, Kheddar H, Himeur Y, Belouchrani A, Serpedin E, Khelifi F, et al. Advanced deep learning and large language models: comprehensive insights for cancer detection. Image Vis Comput. 2025;157(6):105495. doi:10.1016/j.imavis.2025.105495.
25. Habchi Y, Kheddar H, Himeur Y, Ghanem MC. Machine learning and transformers for thyroid carcinoma diagnosis. J Visual Commun Image Represent. 2026;115(4):104668. doi:10.1016/j.jvcir.2025.104668.
26. Ma ZB, Mi Y, Zhang B, Zhang Z, Wu JY, Huang HW, et al. Review on deep learning algorithms for heterogeneous medical image processing. J Softw. 2023;34(10):4870–915. (In Chinese).
27. Xu T, Li Y, Gao X, Rzhetsky A, Jia G. An effective encoding of human medical conditions in disease space provides a versatile framework for deciphering disease associations. Quant Biol. 2025;13(3):e93. doi:10.1002/qub2.93.
28. Zang X, Xia Y, Xiao H, Luo H, Si M, Hou N, et al. A multimodal AI model for precision prognosis in clear cell renal cell carcinoma: a multicenter study. npj Digit Med. 2025;8(1):668. doi:10.1038/s41746-025-02034-x.
29. Zhang Y, Li Q, Xin Y. Research on eight machine learning algorithms applicability on different characteristics data sets in medical classification tasks. Front Comput Neurosci. 2024;18:1345575. doi:10.3389/fncom.2024.1345575.
30. Zhang Y, Xin Y, Li Q. Research on parameter selection and optimization of C4.5 algorithm based on algorithm applicability knowledge base. Sci Rep. 2025;15(1):29418. doi:10.1038/s41598-025-11901-2.