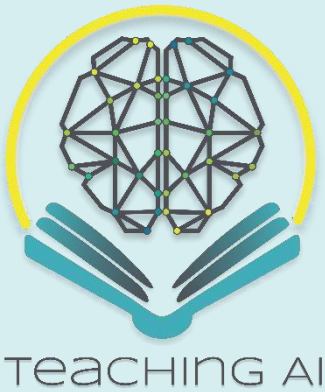


Report WP3-A2:

Analyses of Databases

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency.

Co-funded by the European Union

Result

Analyses of databases

Related to
WP3-A2: Analyses of databases

Statement of originality

This report contains original unpublished work, except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation, or both.

Disclaimer

This report contains material, which is the copyright of TAI Consortium Parties. All TAI Consortium Parties have agreed that the content of the report is licensed under a Creative Commons Attribution Non-Commercial Share Alike 4.0 International License. TAI Consortium Parties does not warrant that the information contained in the Report is capable of use, or that use of the information is free from risk and accept no liability for loss or damage suffered by any person or any entity using the information.

Copyright notice

Note

For anyone interested in having more information about the project, please see the website at: https://tai-erasmus.eu/

Document version history

Version	Date	Prepared by:	Comment
1.0	12.07.2026.	Jelena Šaković Jovanović, Aleksandar Vujović, Janko Jovanović, UOM	Initial version
2.0	16.07.2026.	Jelena Šaković Jovanović, Aleksandar Vujović, Janko Jovanović, UOM	Second version with contributions from and review by the TAI WP3 partners (UNILJ, UNBG, UNSA, and UNIRI)
			Final consolidated version prepared by UoM following contributions, review and verification by the WP3 partners

Table of contents

1. Introduction

The Teaching Artificial Intelligence (TAI) project is implemented within the Erasmus+ Programme 2024, Key Action 2 - Cooperation Partnerships in Higher Education. The overall aim of the project is to strengthen Artificial Intelligence (AI) education through the development of teaching materials, practical learning resources and research-oriented educational activities. Within this framework, Work Package 3 (WP3) focuses on collecting and analysing datasets that can support practical AI and Machine Learning (ML) teaching, learning and student projects.

This report presents the work carried out within WP3-A2: Analysis of Various Databases. The activity builds on WP3-A1, during which 120 datasets were identified across four data modalities. UNILJ - University of Ljubljana was responsible for video datasets, UoM - University of Montenegro for audio/sound datasets, UNBG - University of Belgrade for image datasets and UNSA - University of Sarajevo for numerical datasets. UNIRI - University of Rijeka contributed by reviewing the collected datasets and advising on the structure and formats of the collection. The collected information was recorded in a shared dataset collection document, which is provided in Annex 1 and served as the primary data source for the analysis presented in this report.

The purpose of this report is to analyse the relevance, accessibility, conditions of use, formats, technical characteristics and potential educational applicability of the collected datasets. The analysis also considers their main limitations and the level of preparation required before their use in practical exercises, student projects and research-oriented learning activities.

In addition to the separate analysis of the four data modalities, the report provides a comparative assessment of the complete dataset collection, identifies representative datasets suitable for introductory, intermediate and advanced or research-oriented educational activities, and presents the 12 datasets selected and uploaded to the project repository for subsequent project activities.

The findings of this report provide a basis for the further technical preparation of the selected datasets and the development of Python-based educational materials. The report was coordinated and prepared by the University of Montenegro, with contributions from and review by the TAI WP3 partners.

2. Dataset Analysis Methodology

The analysis conducted within WP3-A2 builds on the dataset collection and standardisation process completed within WP3-A1. The shared dataset collection document prepared during WP3-A1 and provided in Annex 1 was used as the primary data source for the analysis. It contains information on 120 datasets, equally distributed across four data modalities: video, image, audio/sound and numerical datasets.

The analysis was designed to assess the applicability of the collected datasets for AI education and research using the descriptive, legal, structural and technical information recorded by the project partners. The methodology combined a separate analysis of each data modality with a comparative assessment across the four modalities. Particular attention was given to the relevance of the datasets for AI and Machine Learning tasks, their accessibility and conditions of use, their documented quality indicators, data formats and technical characteristics, as well as the preparation required for their use in practical educational activities.

The preparation of the report followed a collaborative approach. Each partner responsible for a particular data modality contributed the corresponding dataset information and domain-specific expertise. UoM harmonised these inputs and organised them within a common analytical structure covering all four modalities.

2.1 Scope and Data Source

The analysis covered all 120 datasets identified within WP3-A1. The information recorded in the shared dataset collection document was used to examine dataset relevance, accessibility, conditions of use, format, technical characteristics and potential educational applicability. The complete shared dataset collection document, containing information on all 120 datasets, is provided in Annex 1.

The analysis was based on both common information recorded for all datasets and technical characteristics specific to each data modality. Most of this information was obtained from descriptions provided by the dataset owners or providers in the relevant repositories or accompanying scientific papers. However, the level of available documentation varied across the datasets. Consequently, some information was incomplete, limited or not publicly available. A detailed description of these fields, as well as the dataset collection and standardisation process, is provided in Report WP3-A1.

2.2 Analysis Dimensions and Procedure

The datasets were analysed according to four main dimensions:

1. relevance to AI and Machine Learning tasks and application areas;
2. accessibility and conditions of use;
3. documented quality and technical suitability;
4. potential educational applicability.

Relevance was assessed on the basis of dataset descriptions and stated application areas.

Accessibility and conditions of use were analysed using the available sources, links and licence information. For datasets supporting several AI tasks, one main application category was assigned according to the primary intended use described in the shared dataset collection document. The classification of individual datasets is provided in Annex 2.

Documented quality and technical suitability were assessed based on dataset scale, structure, format and modality-specific technical characteristics. Where available, information on labels or annotations was also taken into account. The assessment relied on information documented by the dataset providers and recorded in the shared dataset collection document and did not include an independent validation of all dataset files or annotations.

Educational applicability was assessed in relation to the suitability of the datasets for practical exercises, student projects and research-oriented activities, taking into account the need for preprocessing, standardisation or the extraction of smaller subsets.

The analysis was conducted separately for video, image, audio/sound and numerical datasets, using the main common and modality-specific parameters recorded in the shared dataset collection document. The modality-specific findings were reviewed by the partners responsible for the corresponding dataset categories before their inclusion in the consolidated report.

Based on these dimensions, datasets were considered particularly suitable when they combined a clearly defined AI application, sufficiently documented data structures and, where relevant, labels or annotations, technically usable formats, manageable preparation requirements, and conditions of use compatible with the intended educational or research activity. Recommendations were adapted to the expected level of use, distinguishing between introductory exercises, intermediate student projects, and advanced or research-oriented activities.

3. Analysis of Collected Datasets

The collected datasets were analysed separately by data modality in order to identify their main characteristics, application potential, accessibility, technical suitability and possible use in AI education. The analysis was based on the information recorded in the shared dataset collection document and focused on the main patterns and differences within each group of 30 datasets.

For each modality, the analysis considers the represented AI application areas, licensing and access conditions, dataset scale and structure, commonly used formats and modality-specific technical characteristics. Where available, information on labels or annotations was also taken into account. These findings are used to identify the main educational opportunities, practical requirements and potential limitations associated with each group of datasets.

The results are presented for video, audio/sound, numerical, and image datasets in the following sections.

3.1 Video Datasets

3.1.1 Main Applications of Video Datasets

The collected video datasets cover a wide range of applications relevant to Artificial Intelligence and computer vision. As shown in Figure 3.1, the largest group consists of datasets intended for human action and behaviour recognition, accounting for 8 datasets or 26.67% of the collection. Agriculture and animal monitoring represents the second largest category with 6 datasets (20.00%), followed by traffic, surveillance and safety-related applications with 5 datasets (16.67%).

Object detection and tracking, autonomous driving and navigation, and multimodal understanding and affect analysis are each represented by 3 datasets (10.00%). General video classification and recognition accounts for 2 datasets, or 6.67% of the collection. This distribution demonstrates that the video datasets support both widely used computer vision tasks and more specialised educational and research applications.

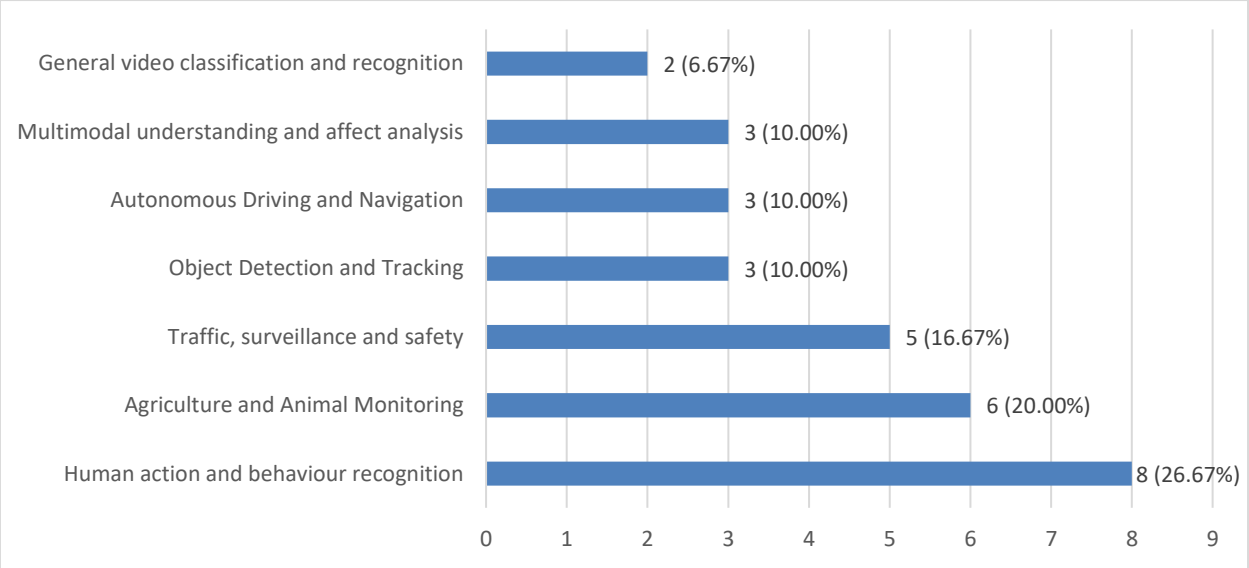


Figure 3.1. Distribution of video datasets by main application

For the purpose of the analysis, each dataset was assigned to one main application category according to its primary intended use. Although some datasets may support more than one AI task, only the dominant application was considered in order to avoid multiple counting. A detailed list of the datasets assigned to each main application category is provided in Annex 2.

3.1.2 Licence and Conditions of Use

The analysis of the licence information shows that the collected video datasets differ considerably in their conditions of use. As presented in Figure 3.2, the largest group consists of datasets distributed under Creative Commons licences, accounting for 12 datasets or 40.0% of the collection. This category includes both attribution-based licences and Creative Commons licences with additional non-commercial, share-alike or no-derivatives conditions.

A further 7 datasets, or 23.3%, are available under separate non-commercial, academic or research-use conditions, while 4 datasets, or 13.3%, are classified as open or public domain. Two datasets, representing 6.7%, are subject to custom or restricted licence arrangements, and licence information was not clearly specified for 5 datasets, or 16.7%.

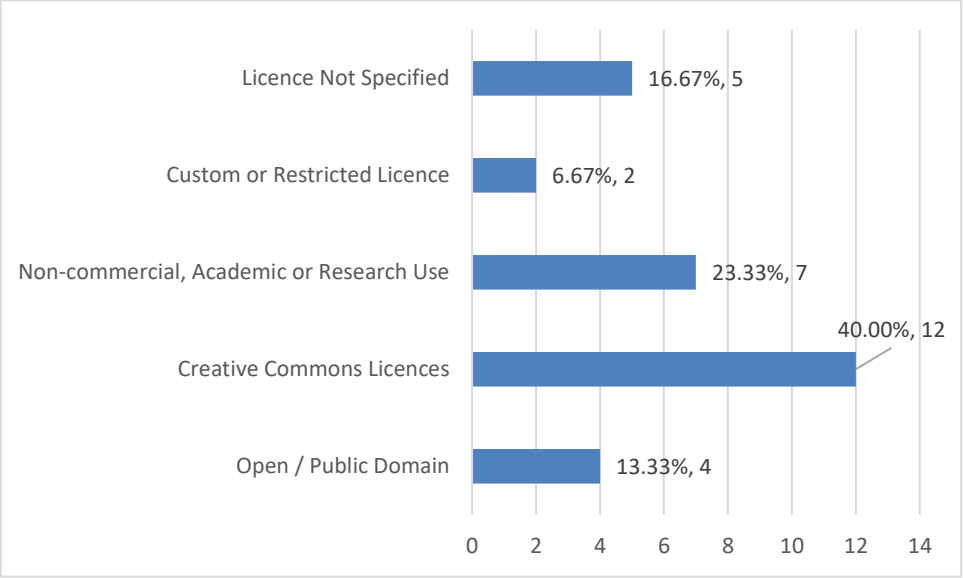


Figure 3.2. Distribution of video datasets by licence and conditions of use

These results indicate that a substantial part of the video dataset collection can be used in higher education and research. However, the precise licence conditions should be checked before datasets or derived materials are incorporated into teaching resources or publicly distributed project outputs.

3.1.3 Dataset Scale: Number of Videos and Total Size

The scale of the collected video datasets was analysed using two complementary parameters: the number of video recordings and the total dataset size. The number of videos indicates the amount of available training and evaluation material, while the total size provides an indication of storage, download and computational requirements.

Number of Videos

As shown in Figure 3.3, the collected datasets differ considerably in the number of video recordings they contain. The classification is based on the number of videos, scenes, clips or video segments reported in the shared dataset collection document. Ten datasets (33.3%) include fewer than 100 videos, while seven datasets (23.3%) contain between 100 and 999 videos. Four datasets (13.3%) include between 1,000 and 9,999 videos, and six datasets (20.0%) contain between 10,000 and 99,999 videos. The remaining three datasets (10.0%) contain 100,000 or more video recordings or segments.

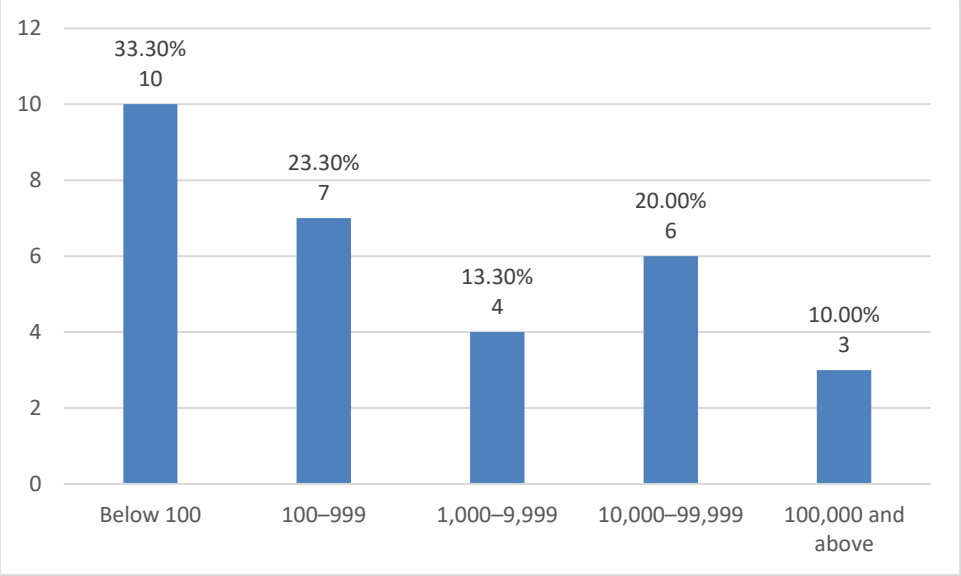


Figure 3.3. Distribution of video datasets by number of videos

The results show that more than half of the datasets contain fewer than 1,000 videos. These collections may be easier to manage in practical classes and student projects. Datasets containing several thousand or hundreds of thousands of recordings provide greater potential for model training and evaluation but generally require more preparation and computational resources.
For multimodal and continuously recorded datasets, scenes, video segments or continuous camera streams were treated as the closest equivalent to individual video units, based on the information available in the dataset collection document.

Total Dataset Size

The total size of the datasets also varies substantially. Ten datasets (33.3%) fall within the range of 10–100 GB, making this the most represented category. Nine datasets (30.0%) have a total size between 1 and 10 GB, while five datasets (16.7%) are smaller than 1 GB. Three datasets (10.0%) range from 100 GB to 1 TB, and one dataset (3.3%) exceeds 1 TB. Size information was not available for two datasets (6.7%).

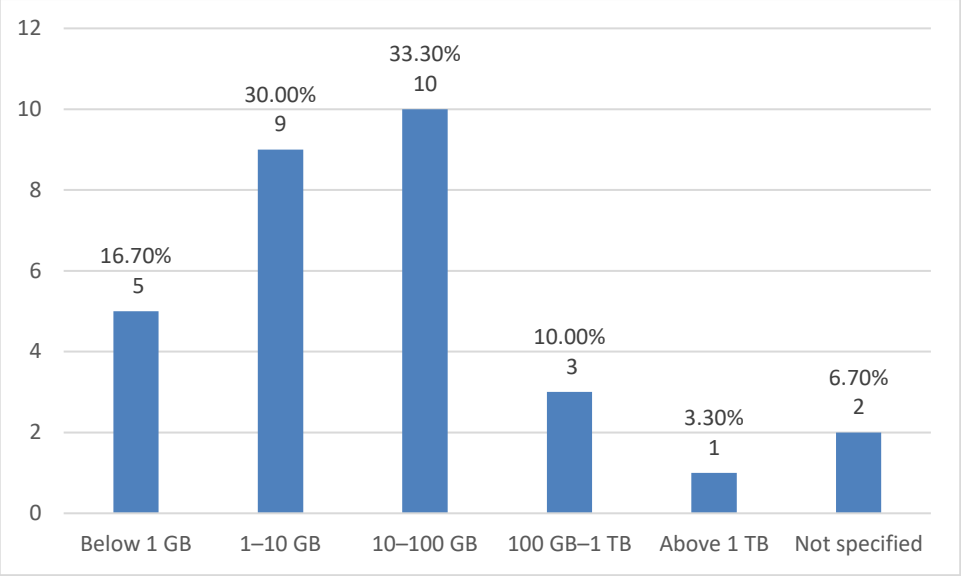


Figure 3.4. Distribution of video datasets by total size

Almost half of the collected datasets are smaller than 10 GB and can therefore be considered relatively manageable for educational use. Datasets between 10 and 100 GB may still be suitable for student projects, although the preparation of smaller subsets may be necessary. Collections above 100 GB are more demanding in terms of download, storage and processing and are therefore generally more appropriate for advanced projects or research-oriented activities.

Taken together, the number of videos and total size show that dataset scale cannot be assessed using only one parameter. A dataset containing many short or low-resolution clips may be smaller than a collection containing fewer but longer or higher-resolution recordings. Both indicators should therefore be considered when selecting video datasets for practical AI teaching.

3.1.4 Video Formats and Technical Characteristics

The technical suitability of the collected video datasets was assessed through their main data formats, video length, resolution and frame rate. These parameters influence compatibility with commonly used AI and video-processing tools, the level of preprocessing required and the computational resources needed for practical educational activities.

Video Formats

As shown in Figure 3.5, the MP4/MPEG-4 family is the dominant format, used in 17 datasets (56.7%). AVI is represented in 4 datasets (13.3%), while 3 datasets (10.0%) use specialised

formats such as ROS bag, TFRecord or CSD/HDF5. Frame- or image-based collections and mixed-format datasets each account for 2 datasets (6.7%). WebM and MKV are each represented by one dataset (3.3%).

The predominance of MP4/MPEG-4 and AVI is favourable for educational use because these formats are widely supported by common video-processing tools and Python libraries. Specialised and multimodal formats may require additional conversion, dedicated software or more advanced technical knowledge before they can be used in practical exercises.

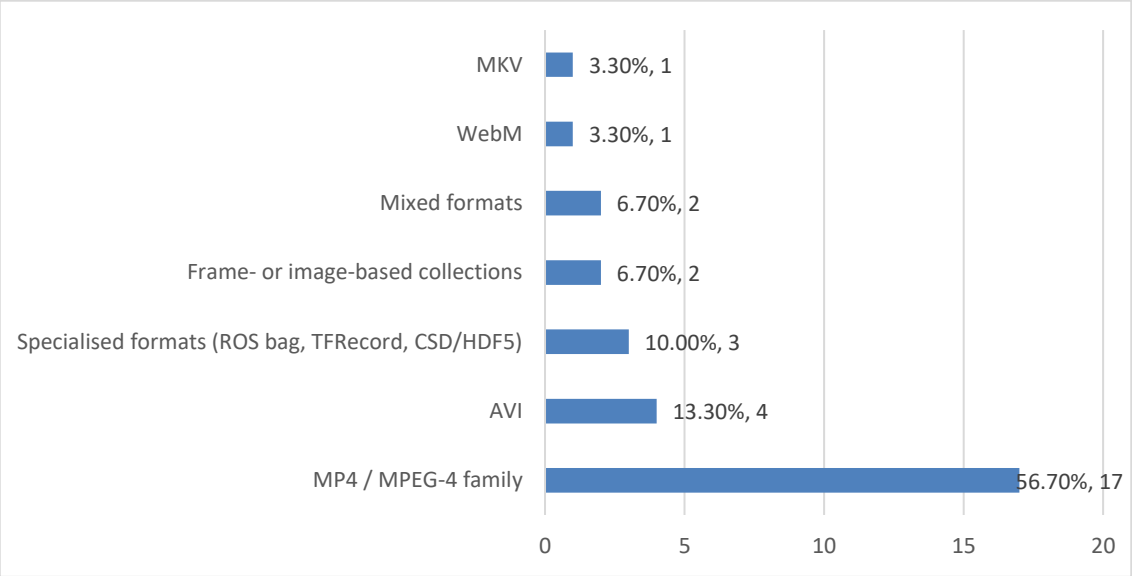


Figure 3.5. Distribution of video datasets by main data format

Video length

The collected video datasets also differ considerably in the duration of individual recordings. Some datasets contain very short clips lasting only a few seconds, while others include longer recordings, continuous video streams or segments with variable duration. This variation reflects the different purposes for which the datasets were developed.

Short video clips are particularly suitable for tasks such as action recognition, gesture classification and event detection because they usually focus on a single activity or clearly defined event. Longer recordings provide richer temporal information and are more appropriate for tracking, behaviour analysis, surveillance, autonomous systems and multimodal video understanding. However, they also require more storage, memory and processing time.

For the purpose of the analysis, video length can be grouped into the following categories:

- below 10 seconds;
- 10–60 seconds;

- 1–10 minutes;
- above 10 minutes;
- variable, continuous or not specified.

Datasets containing long or continuous recordings may require temporal segmentation, frame sampling or the extraction of shorter clips before being used in practical educational activities. In contrast, datasets consisting of short and clearly defined clips are generally easier to integrate into introductory and intermediate AI exercises.

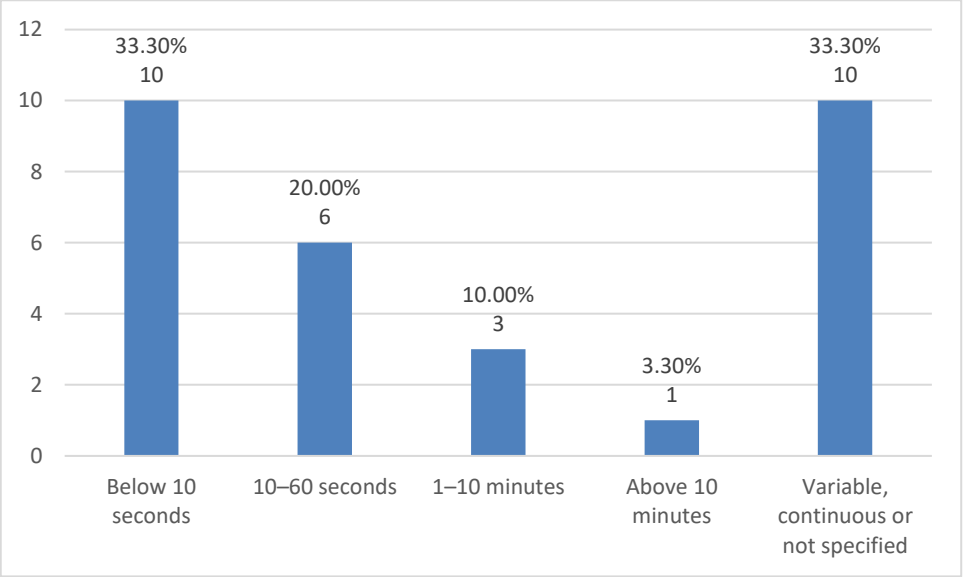


Figure 3.6 Distribution of video datasets by video length

Where average or median duration was available, it was used for classification. Datasets with wide duration ranges, continuous recordings or unavailable duration information were classified as variable, continuous or not specified.

Video Resolution

For the purpose of this analysis, the reported video resolutions were grouped into low resolution, HD or near-HD, Full HD, 4K or higher, multiple or variable resolutions, and not specified. The collected datasets show substantial variation across these categories. Low-resolution videos represent 8 datasets (26.7%), while 6 datasets (20.0%) contain HD or near-HD material. Seven datasets (23.3%) include multiple or variable resolutions, reflecting differences between recordings, cameras or sensors. Full HD is the main resolution in 3 datasets (10.0%), while 4 datasets (13.3%) provide 4K or higher-resolution recordings. Resolution information was not specified for 2 datasets (6.7%).

Lower-resolution datasets generally require fewer computing resources and may therefore be more suitable for introductory exercises. Full HD and 4K datasets provide richer visual

information but require more storage, memory and processing capacity. Datasets containing multiple resolutions may also require resizing or standardisation before model training.

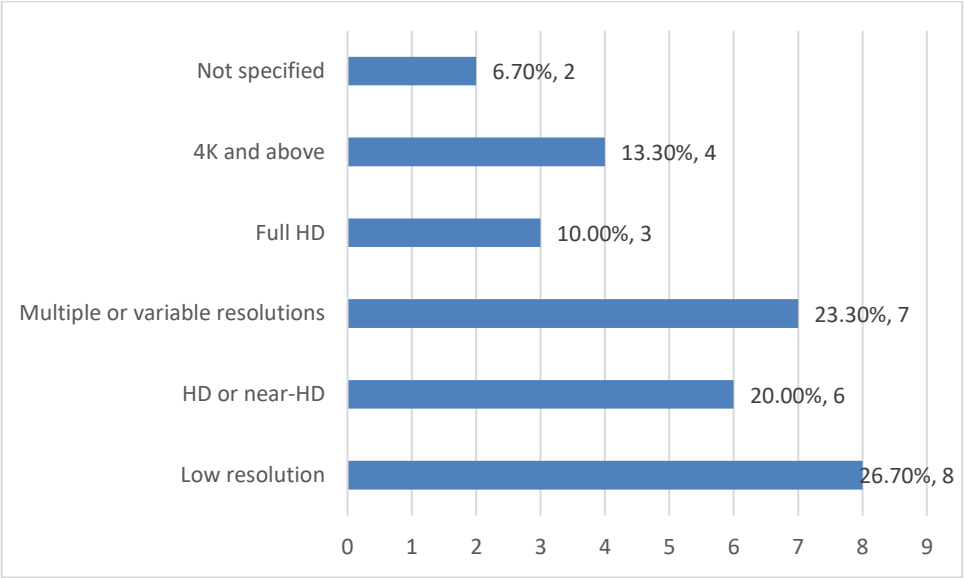


Figure 3.7. Distribution of video datasets by resolution category

Frame Rate

Half of the collected datasets, 15 datasets (50.0%), use frame rates between 20 and 30 fps. Six datasets (20.0%) have frame rates below 20 fps, while one dataset (3.3%) exceeds 30 fps. Five datasets (16.7%) contain variable or multiple frame rates, and frame-rate information was not specified for 3 datasets (10.0%).

Frame rates between 20 and 30 fps are suitable for many common tasks, including action recognition, classification, detection and tracking. Lower frame rates may reduce computational requirements but can provide less temporal detail, while higher or variable frame rates may require frame sampling or temporal standardisation before datasets are used within a common educational workflow.

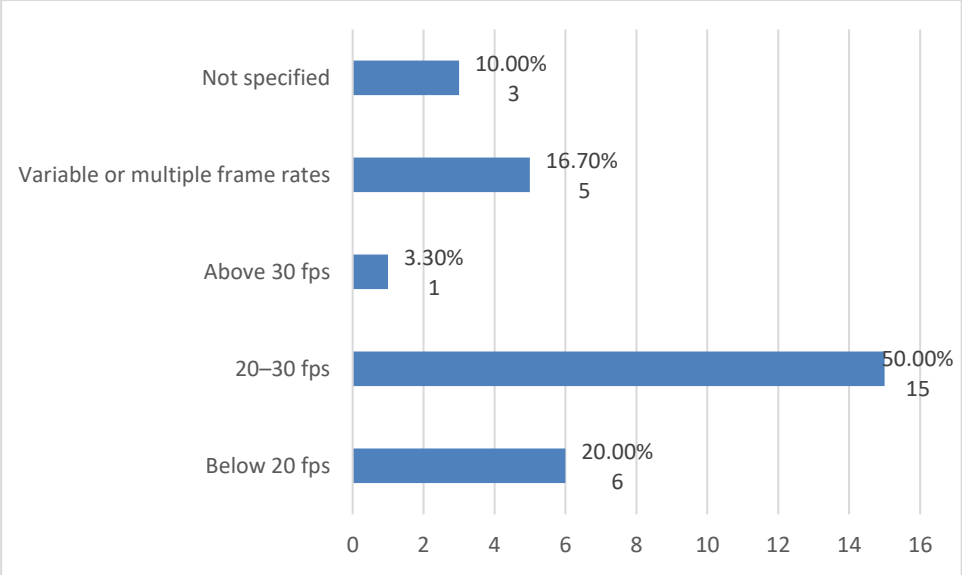


Figure 3.8. Distribution of video datasets by frame rate

The recorded codec information also shows that H.264 and related MPEG-4 codecs are widely represented, although several datasets use VP9, XVID or specialised codecs, and codec information is incomplete for some collections. Overall, most video datasets use formats and technical specifications that can be processed using standard AI and computer-vision tools, while datasets with specialised structures, very high resolutions or variable frame rates require additional preparation.

3.1.5 Educational Applicability of Video Datasets

Based on the recorded characteristics, the collected video datasets offer a wide range of opportunities for AI teaching and research, from basic video classification exercises to advanced multimodal learning and autonomous-system applications. In practice, the most appropriate dataset is not necessarily the largest or the most technically complex one. For introductory activities, clearly defined tasks, understandable labels, manageable dataset size and compatibility with commonly used tools are usually more important. As students' progress, larger collections, more complex annotations and multimodal data can be introduced to support more demanding model-development and research activities.

For introductory practical exercises, smaller datasets containing short videos and clearly defined classes are particularly useful. The Highway Traffic Videos Dataset, for example, contains 254 five-second videos with manually assigned labels describing light, medium and heavy traffic. Its small size of approximately 92 MB, low resolution and clearly defined classification task make it suitable for demonstrating the complete machine-learning workflow, including video loading, frame extraction, preprocessing, model training and performance evaluation. The KTH Action

Recognition Dataset is another suitable example. It contains 600 videos representing six human actions performed by 25 participants in different scenarios. Its limited number of classes, relatively small size and low-resolution recordings allow students to focus on the basic principles of action recognition without excessive computational requirements.

At the intermediate level, students can work with datasets that include more classes, higher-resolution recordings or more demanding preprocessing requirements. The WLASL Dataset, containing 12,000 videos of 2,000 American Sign Language words, can support projects involving gesture and sign-language recognition. Although the large number of classes makes the complete dataset relatively demanding, carefully selected subsets could be used for intermediate projects focusing on a smaller vocabulary. The CBVD-5 Cow Behavior Video Dataset provides another relevant example. It contains 887 videos, accompanying annotations and Full HD recordings intended for cow-behaviour recognition, tracking and livestock monitoring. Its use would allow students to work with realistic video data while learning how to perform frame sampling, resizing and annotation processing.

Large-scale, multimodal and technically specialised datasets are more appropriate for advanced student projects and research-oriented activities. The nuScenes Dataset, which combines camera data with LiDAR and radar recordings, can support advanced work on autonomous driving, object detection, tracking and sensor fusion. However, its total size of approximately 350 GB and heterogeneous data structure require substantial storage, computational resources and data-preparation skills. Similarly, the FieldSAFE Dataset combines stereo, thermal, web and 360-degree cameras with LiDAR, radar and localisation data collected during a synchronised two-hour agricultural event. Its raw sensor streams and ROS bag structure make it particularly relevant for research involving multimodal obstacle detection and agricultural robotics, but also require specialised software and knowledge of sensor-data processing. The EGO4D Dataset, with more than 24,000 first-person videos and approximately 20 TB of data, provides valuable opportunities for advanced research on egocentric perception, although its scale generally makes the use of selected subsets necessary in an educational setting.

Table 3.1. Examples of video datasets suitable for different educational levels

Educational level	Dataset and suitable AI task	Reason for suitability	Required preparation or conditions
Introductory practical exercises	**Highway Traffic Videos Dataset** - traffic-level classification	Small collection of 254 short videos, clearly defined classes and manually assigned labels; CC0 public-domain licence	Basic frame extraction and video preprocessing

Introductory practical exercises	**KTH Action Recognition Dataset** - human action recognition	Six clearly defined action classes, 600 videos, low resolution and manageable size	Possible AVI/codec conversion, depending on the software used; non-commercial licence
Intermediate student projects	**WLASL Dataset** – sign-language and gesture recognition	Standard MP4 format, manageable total size and a clearly defined recognition task	Extraction of a smaller vocabulary subset is advisable; use is subject to the C-UDA
Intermediate student projects	**CBVD-5 Cow Behavior Video Dataset** - behaviour recognition and tracking	Annotated Full HD videos supporting realistic classification and tracking tasks; CC0 public-domain licence	Frame sampling, resizing and annotation processing may be required
Advanced or research-oriented activities	**nuScenes Dataset** - autonomous driving, detection, tracking and sensor fusion	Multimodal data combining camera recordings, LiDAR and radar	Approximately 350 GB; specialised processing and substantial computing resources required; non-commercial use
Advanced or research-oriented activities	**FieldSAFE Dataset** - multimodal obstacle detection and agricultural robotics	Synchronised data from several camera types, LiDAR, radar and localisation sensors; CC BY 4.0 licence	ROS bag and raw sensor processing, temporal synchronisation and specialised software required

The proposed classification is indicative rather than fixed. A large or technically complex dataset can also be used at an introductory or intermediate level if an appropriate subset is prepared in advance and the task is simplified. Conversely, a relatively small dataset can support advanced work when it is used to investigate more complex modelling, evaluation or preprocessing methods. In all cases, dataset selection should be aligned with the intended learning outcomes, available teaching time and computational resources. Access conditions and licence requirements should also be verified before a dataset is incorporated into teaching materials or project outputs. These recommendations are based on the characteristics documented by the dataset providers and recorded in the shared dataset collection document.

3.2 Audio and Sound Datasets

The collected audio and sound datasets cover a broad range of AI applications, including speech processing, environmental sound recognition, music analysis, bioacoustics and biomedical sound analysis. The datasets differ considerably in their application areas, scale, formats, technical characteristics and licensing conditions. The following analysis examines these characteristics and their relevance for AI education and research.

3.2.1 Main Applications of Audio and Sound Datasets

The collected audio and sound datasets cover a wide range of applications relevant to Artificial Intelligence, machine learning and digital signal processing. As shown in Figure 3.9, the largest group consists of datasets intended for music analysis and music information retrieval, accounting for 10 datasets or 33.33% of the collection. This category includes music genre and instrument recognition, automatic music tagging and transcription, audio synthesis, melody extraction, music generation and source separation.

Environmental sound and general audio event recognition represents the second largest category, with 7 datasets (23.33%). These datasets support tasks such as environmental sound classification, acoustic scene recognition, audio tagging and sound event detection. Speech recognition, speaker identification and speech synthesis are represented by 6 datasets (20.00%), covering automatic speech recognition, keyword spotting, speaker verification, text-to-speech synthesis and spoken digit classification.

Bioacoustics and animal sound recognition account for 2 datasets (6.67%), while another 2 datasets (6.67%) are intended for speech emotion recognition. Biomedical sound analysis is also represented by 2 datasets (6.67%), focusing on heartbeat and respiratory sound classification. The remaining dataset (3.33%) supports vehicle speed estimation using audio and video recordings.

This distribution demonstrates that the collection supports both widely used audio-processing tasks and more specialised educational and research applications.

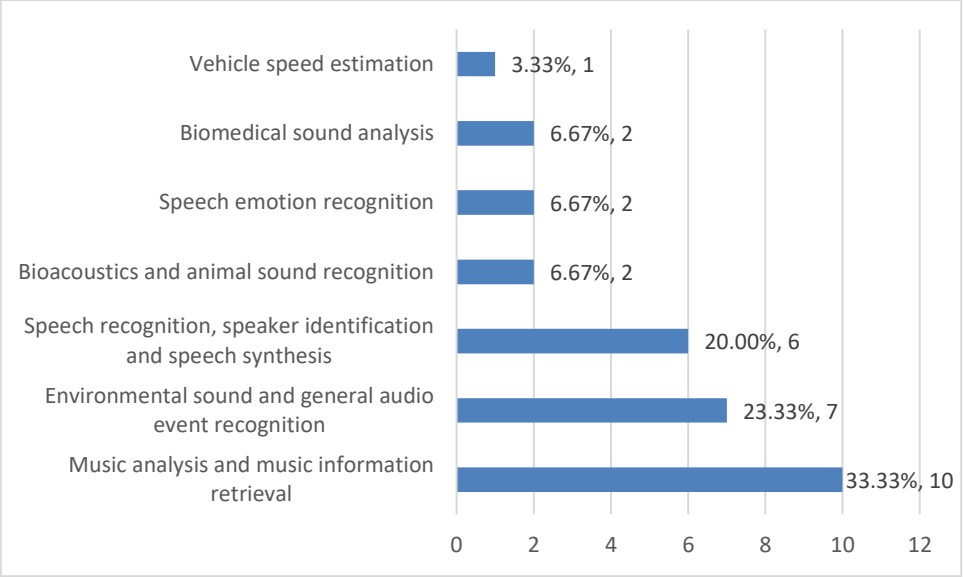


Figure 3.9. Distribution of audio and sound datasets by main application

For the purpose of the analysis, each dataset was assigned to one main application category according to its primary intended use. Although some datasets may support more than one AI task, only the dominant application was considered in order to avoid multiple counting. A detailed list of the datasets assigned to each main application category is provided in Annex 2.

3.2.2 Licence and Conditions of Use

The analysis of the licence information shows that the audio and sound datasets are generally accompanied by clearly stated conditions of use. As presented in Figure 3.10, Creative Commons licences represent the largest category, covering 22 datasets or 73.3% of the collection. This category includes attribution-based licences as well as Creative Commons licences containing non-commercial, share-alike or other specific conditions.

Two datasets, or 6.7%, are available under public-domain or other open-use conditions. A further 5 datasets, representing 16.7%, are subject to separate research-only, non-commercial, custom or otherwise restricted conditions. Licence information was not clearly specified for one dataset, corresponding to 3.3% of the collection.

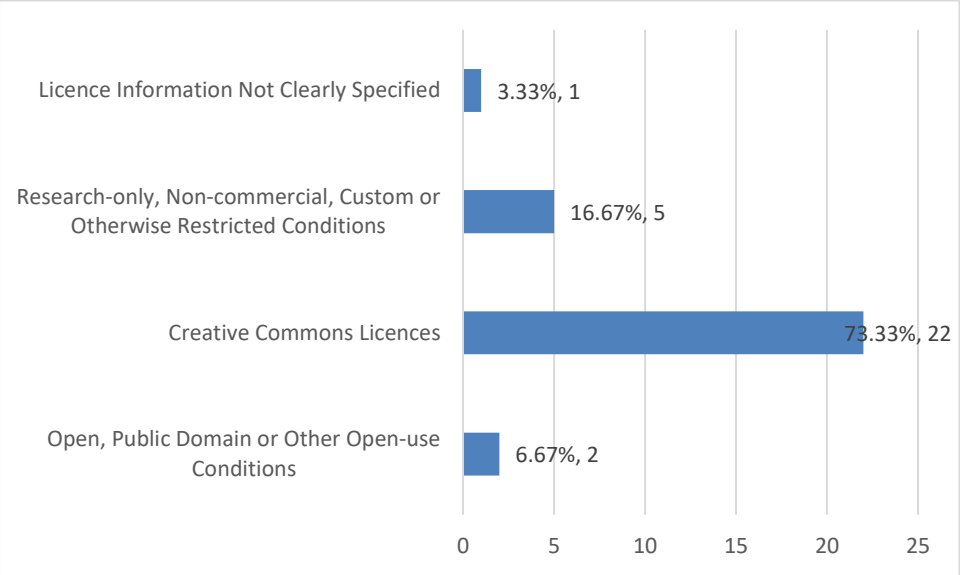


Figure 3.10. Distribution of audio and sound datasets by licence and conditions of use

These results indicate that most audio and sound datasets can support educational and research activities. However, the precise licence conditions should be checked before datasets or derived materials are redistributed or included in publicly available teaching resources.

3.2.3 Dataset Scale: Number of Audio Files and Total Size

The scale of the collected audio and sound datasets was analysed using two complementary parameters: the number of audio files or recordings and the total dataset size. The number of audio files indicates the amount of material available for model training, validation and evaluation, while the total size provides an indication of download, storage and computational requirements.

Number of Audio Files

The scale of the collected audio and sound datasets was assessed according to the reported number of audio files or equivalent units. Since the collections use different organisational structures, tracks, clips, utterances, recordings and performances were treated as the closest equivalents to individual audio files. For the purpose of this analysis, the datasets were grouped into five categories: fewer than 100 files, 100–999 files, 1,000–9,999 files, 10,000–99,999 files, and 100,000 or more files.

As shown in Figure 3.11, one dataset (3.33%) contains fewer than 100 audio files, while 5 datasets (16.67%) include between 100 and 999 files. Nine datasets (30.00%) contain between 1,000 and

9,999 files, and 7 datasets (23.33%) include between 10,000 and 99,999 files. The remaining 8 datasets (26.67%) contain 100,000 or more audio recordings or equivalent audio units.

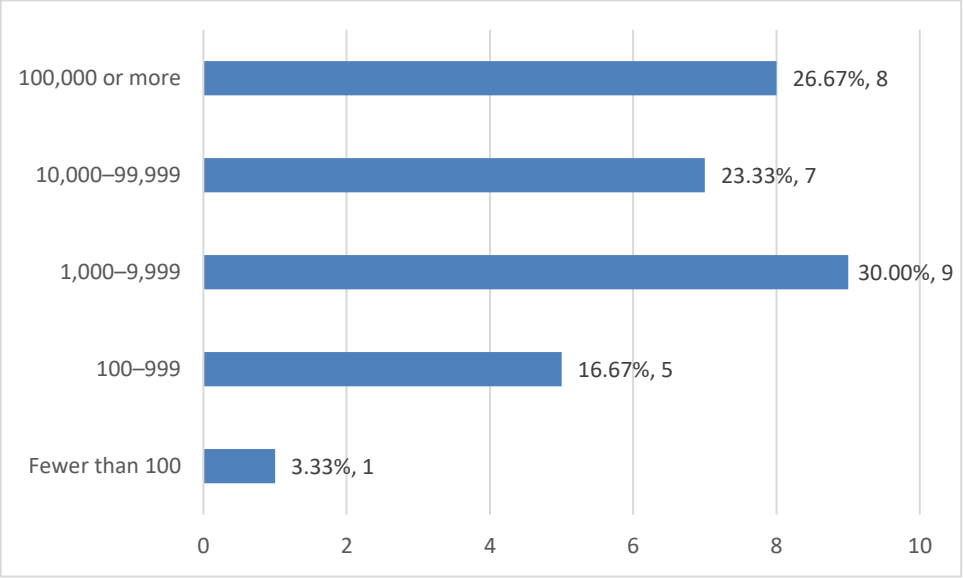


Figure 3.11. Distribution of audio/sound datasets by number of audio files

The results show that 16 datasets contain between 1,000 and 99,999 audio files or equivalent units, while 8 datasets contain at least 100,000 recordings. Smaller collections may be easier to download, inspect and process during practical classes and student projects. In contrast, datasets containing tens or hundreds of thousands of recordings provide greater potential for training and evaluating more complex models, but generally require additional preparation, storage capacity and computational resources.

The number of audio files should be considered together with the duration and total size of the dataset. A collection containing a large number of short audio clips may require less storage than a dataset consisting of fewer but substantially longer recordings. These parameters should therefore be considered jointly when selecting datasets for educational activities.

Total Dataset Size

The total size of the audio and sound datasets also varies substantially. Three datasets (10.00%) are smaller than 1 GB, while 8 datasets (26.67%) range from 1 to less than 10 GB. The largest group consists of 12 datasets (40.00%) with sizes ranging from 10 to less than 100 GB. Four datasets (13.33%) fall within the range of 100 GB to 1 TB, and 2 datasets (6.67%) exceed 1 TB. Size information was not available for one dataset (3.33%).

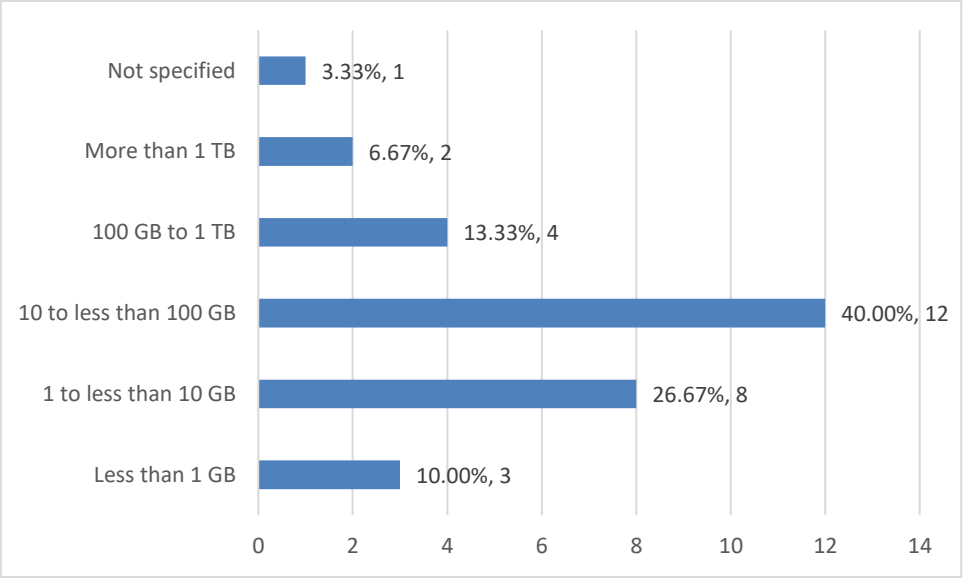


Figure 3.12. Distribution of audio/sound datasets by total size

More than one third of the collected datasets are smaller than 10 GB and can therefore be considered relatively manageable for educational use. Datasets between 10 and 100 GB may also be suitable for student projects, although downloading selected subsets or reducing the number of recordings may be necessary. Collections larger than 100 GB are more demanding in terms of storage, transfer time and processing capacity and are therefore generally more appropriate for advanced projects or research-oriented activities.

3.2.4 Audio Formats and Technical Characteristics

The technical suitability of the collected audio and sound datasets was assessed through their main audio formats, duration of recordings, sampling rates and channel configurations. These parameters influence compatibility with commonly used audio-processing tools and Python libraries, the level of preprocessing required and the computational resources needed for practical educational activities.

Audio Formats

As shown in Figure 3.13, WAV is the dominant audio format, used as the primary format in 19 datasets (63.33%). Five datasets (16.67%) contain mixed audio or multimodal formats, such as WAV combined with MP3, M4A, AIF, MIDI or STEMS files. MP3 is the primary format in 3 datasets (10.00%), while FLAC is represented in one dataset (3.33%). One dataset (3.33%) provides pre-extracted audio features in CSV and TensorFlow Record formats, while another dataset (3.33%) is distributed through compressed source formats such as MP4 and WebM.

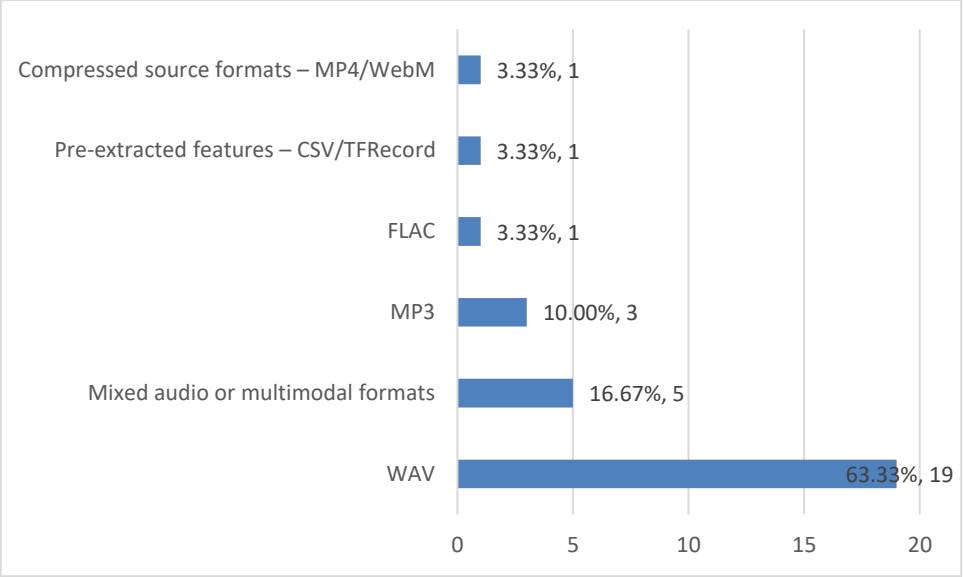


Figure 3.13. Distribution of audio and sound datasets by main data format

The predominance of WAV is favourable for educational use because it is widely supported by audio-processing tools and Python libraries. Mixed, compressed or specialised formats may require conversion or additional preprocessing before practical use.

Audio Duration

The collected datasets differ considerably in the duration of individual audio recordings. As shown in Figure 3.14, one dataset (3.33%) contains recordings shorter than one second. Eight datasets (26.67%) primarily contain recordings lasting from 1 to less than 10 seconds, while 9 datasets (30.00%) contain recordings with a typical duration from 10 to 30 seconds. Three datasets (10.00%) mainly contain recordings longer than 30 seconds, including full music tracks lasting several minutes. The remaining 9 datasets (30.00%) contain recordings with wide or variable duration ranges and were therefore classified separately.

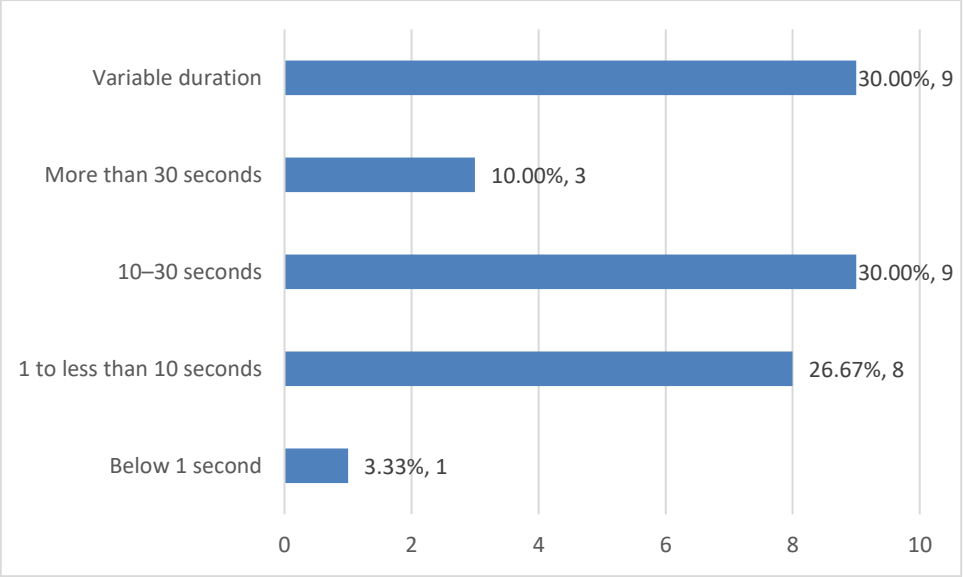


Figure 3.14. Distribution of audio and sound datasets by audio duration

Short and standardised recordings are generally easier to use in practical classes and are suitable for classification, keyword spotting and sound event detection. Longer or variable-duration recordings provide richer temporal information but may require trimming, segmentation or padding before model training.

Sampling Rate

The collected datasets use a range of sampling rates. As presented in Figure 3.15, 44.1 kHz is the most common sampling rate and is used in 11 datasets (36.67%). A sampling rate of 16 kHz is used in 7 datasets (23.33%), while 48 kHz is represented in 3 datasets (10.00%). Two datasets (6.67%) use 22.05 kHz, and one dataset (3.33%) uses 8 kHz. The remaining 6 datasets (20.00%) contain variable or multiple sampling rates.

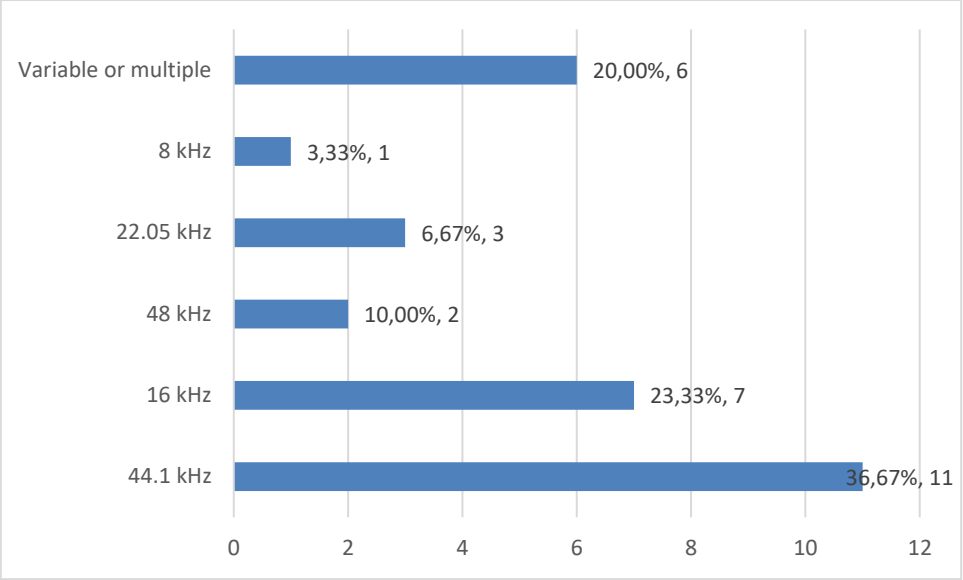


Figure 3.15. Distribution of audio and sound datasets by sampling rate

Sampling rates of 16 kHz are commonly used for speech-related tasks, while 44.1 and 48 kHz are widely used for music and environmental sound analysis. Datasets containing variable sampling rates may require resampling and standardisation before combined processing or model training.

Channel Configuration

Mono recordings are dominant in the collection and are used in 19 datasets (63.33%). Six datasets (20.00%) contain stereo recordings, while 2 datasets (6.67%) include both mono and stereo recordings. The remaining 3 datasets (10.00%) have variable channel configurations or do not provide clearly specified channel information.

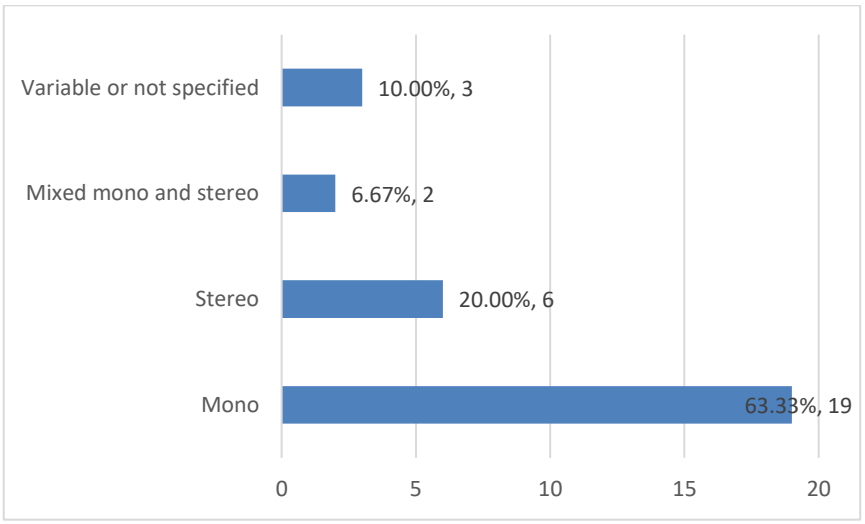


Figure 3.16. Distribution of audio and sound datasets by channel configuration

Mono recordings require less storage and are suitable for many speech- and sound-classification tasks, while stereo recordings preserve additional spatial information relevant to music and acoustic-scene analysis. Mixed channel configurations may require standardisation before model training.

3.2.5 Educational Applicability of Audio and Sound Datasets

The collected audio and sound datasets offer opportunities for teaching a wide range of AI tasks, including sound classification, speech recognition, speaker identification, emotion recognition, music analysis and bioacoustic monitoring. Their educational suitability depends not only on dataset size, but also on the duration and consistency of the recordings, sampling rate, channel configuration, availability of labels or transcripts, data format and the amount of preprocessing required. In an introductory practical class, a small and clearly labelled collection is often more useful than a technically complex benchmark containing hundreds of thousands of recordings. Larger and more heterogeneous datasets become more valuable as students develop the skills and computational resources needed to work with them.

For introductory practical exercises, datasets containing short WAV recordings, clearly defined classes and limited technical variation are particularly suitable. The Free Spoken Digit Dataset (FSDD) is a good example. It contains 3,000 recordings of spoken digits from zero to nine, occupies approximately 10 MB and provides recordings shorter than one second in a consistent mono WAV format. Its manageable size and clear digit labels make it suitable for introducing audio loading, waveform visualisation, feature extraction, spectrogram generation and basic classification. The ESC-50 Dataset provides a natural next step. It contains 2,000 five-second recordings organised into 50 balanced environmental sound classes. Because the files have a consistent sampling rate and channel configuration, students can focus on comparing audio representations and classification methods without first having to resolve extensive data-standardisation issues.

At the intermediate level, students can work with collections that introduce greater technical variation or more complex recognition tasks. UrbanSound8K contains 8,732 labelled urban sound excerpts from ten classes, with start and end times recorded for each excerpt. The dataset can support sound-event classification and the comparison of different feature-extraction approaches. At the same time, its variable sampling rates provide a realistic opportunity to teach resampling and audio standardisation before model training. The RAVDESS Dataset can be used for speech-emotion recognition through its 2,452 short audio-only recordings. Since the complete collection also includes video-only and audio-video files, it can provide a useful transition from single-modality audio analysis to more advanced multimodal emotion-recognition activities. Depending on the selected task, students may need to standardise channel configuration, extract the audio-only subset or align information from different modalities.

Advanced student projects and research-oriented activities can make use of substantially larger or more specialised datasets. LibriSpeech, which contains approximately 1,000 hours of read English speech and around 292,000 utterances with corresponding transcripts, is suitable for

automatic speech recognition and the evaluation of large speech-processing models. Its 60 GB size and large number of files make it more demanding in terms of storage, data organisation and model-training time. AudioSet represents an even more complex case, with more than two million ten-second clips covering a broad hierarchy of acoustic events. Its scale, YouTube-based source structure, variable technical characteristics and clip-level weak labels make it particularly relevant for research on large-scale audio tagging and sound-event recognition. However, working with the collection may require extensive data retrieval, validation, preprocessing and computational resources. Specialised collections such as Maestro, which combines aligned piano audio and MIDI data, can also support advanced work on music transcription, generation and multimodal music representation.

Table 3.2. Examples of audio and sound datasets suitable for different educational levels

Educational level	**Dataset and suitable AI task**	**Reason for suitability**	**Required preparation or conditions**
Introductory practical exercises	**Free Spoken Digit Dataset (FSDD)** – spoken-digit classification	Approximately 10 MB, 3,000 short recordings, consistent mono WAV format and clearly defined digit labels	Basic waveform normalisation and feature extraction; CC BY-SA 4.0 licence
Introductory practical exercises	**ESC-50** - environmental sound classification	2,000 five-second recordings, 50 balanced classes and consistent technical characteristics	Track-level labels are provided, but no temporal annotations; non-commercial use under CC BY-NC 3.0
Intermediate student projects	**UrbanSound8K** – urban sound-event classification	8,732 labelled excerpts from ten classes, accompanied by metadata and excerpt start and end times	Variable sampling rates may require resampling and standardisation; non-commercial use
Intermediate student projects	**RAVDESS** – speech-emotion recognition and multimodal emotion analysis	Short labelled recordings of several emotions, with audio-only and audio-video options	Channel standardisation or extraction of the audio-only subset may be required; CC BY-NC-SA 4.0 licence
Advanced or research-oriented activities	**LibriSpeech** – automatic speech recognition	Approximately 1,000 hours of speech, around 292,000 utterances and corresponding transcripts; CC BY 4.0	Approximately 60 GB; large-scale file handling, FLAC processing and greater training resources may be required

		licence	
Advanced or research-oriented activities	**AudioSet** – audio tagging and sound-event recognition	More than two million human-labelled clips covering a broad range of acoustic events	YouTube-based sources, variable technical characteristics, weak clip-level labels and approximately 2.4 TB of audio; access and usage conditions should be checked

The distinction between educational levels should not be understood as fixed. Many datasets can be adapted by selecting an appropriate subset or simplifying the intended task. The Free Music Archive Dataset, for example, is available as both a smaller collection of approximately 7.2 GB and a full collection of approximately 900 GB. The smaller version could support a student project in music classification, while the complete collection is more appropriate for large-scale research. Similarly, an introductory exercise may use only a limited number of classes from a larger dataset, whereas advanced students may work with the full collection and address issues such as data imbalance, heterogeneous recording conditions or model scalability.

Dataset selection should therefore be aligned with the intended learning outcomes, students' prior knowledge, available teaching time and computational resources. The licence and conditions of use should also be checked before recordings are reused, shared or incorporated into educational materials. The recommendations presented here are based on the characteristics documented by the dataset providers and recorded in the shared dataset collection document.

3.3 Numerical Datasets

The collected numerical datasets cover a broad range of applications relevant to Artificial Intelligence, machine learning, statistical analysis and data-driven decision-making. They include data related to healthcare, finance, energy, environmental monitoring, transportation, economics, engineering and other application areas. The datasets differ considerably in scale, structure, data formats, number of features, availability of time-series information, missing values and licensing conditions. The following analysis examines these characteristics and their relevance for AI education and research.

3.3.1 Main Applications of Numerical Datasets

The collected numerical datasets support a diverse range of AI and data-analysis tasks. As shown in Figure 3.17, the largest group consists of datasets primarily intended for regression and numerical prediction, accounting for 10 datasets (33.33%). These datasets support applications such as house-price prediction, building-energy assessment, materials-property prediction, market analysis and other numerical forecasting tasks.

Classification and decision-support applications represent the second-largest category, with 9 datasets (30.00%). They cover medical diagnosis, customer churn prediction, defect detection, activity recognition, credit-risk assessment and other classification problems. Time-series forecasting and monitoring are represented by 7 datasets (23.33%), including applications related to energy consumption, air quality, climate, epidemiology and transportation demand.

Three datasets (10.00%) primarily support socioeconomic, macroeconomic and panel-data analysis, while one dataset (3.33%) is mainly intended for recommendation and combined predictive analysis.

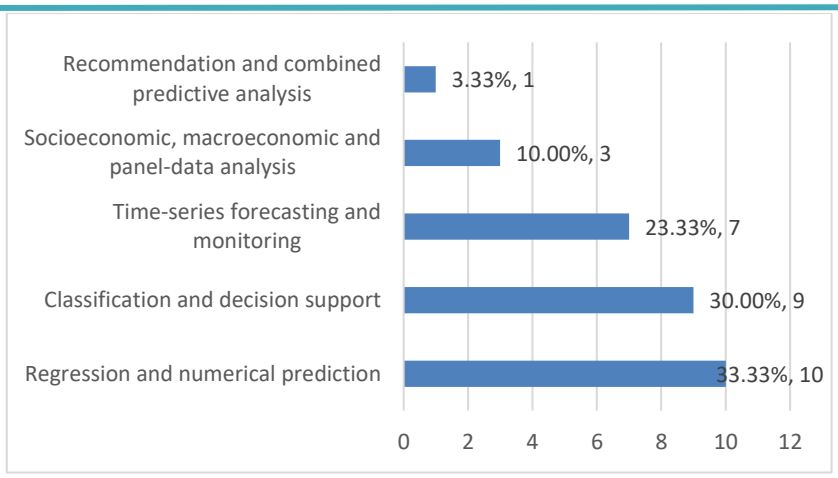


Figure 3.17. Distribution of numerical datasets by main application

This distribution demonstrates that the numerical dataset collection supports both fundamental machine-learning tasks and more specialised educational and research applications.

For the purpose of the analysis, each dataset was assigned to one main application category according to its primary intended use. Although some datasets support several tasks, only the dominant application was considered to avoid multiple counting. A detailed list of the datasets assigned to each main application category is provided in Annex 2.

3.3.2 Licence and Conditions of Use

The analysis of the licence information shows that the numerical datasets are generally available under comparatively open and clearly stated conditions of use. As presented in Figure 3.18, Creative Commons licences represent the largest category, covering 17 datasets or 56.7% of the collection. This category includes both attribution-based licences and Creative Commons licences containing additional non-commercial, share-alike or no-derivatives conditions.

A further 11 datasets, representing 36.7% of the collection, are available under public-domain or other open-use conditions, including CC0, MIT, open-data and similar reuse arrangements. The remaining 2 datasets, or 6.7%, are subject to separate non-commercial, institutional or otherwise restricted conditions. Licence information was recorded for all numerical datasets.

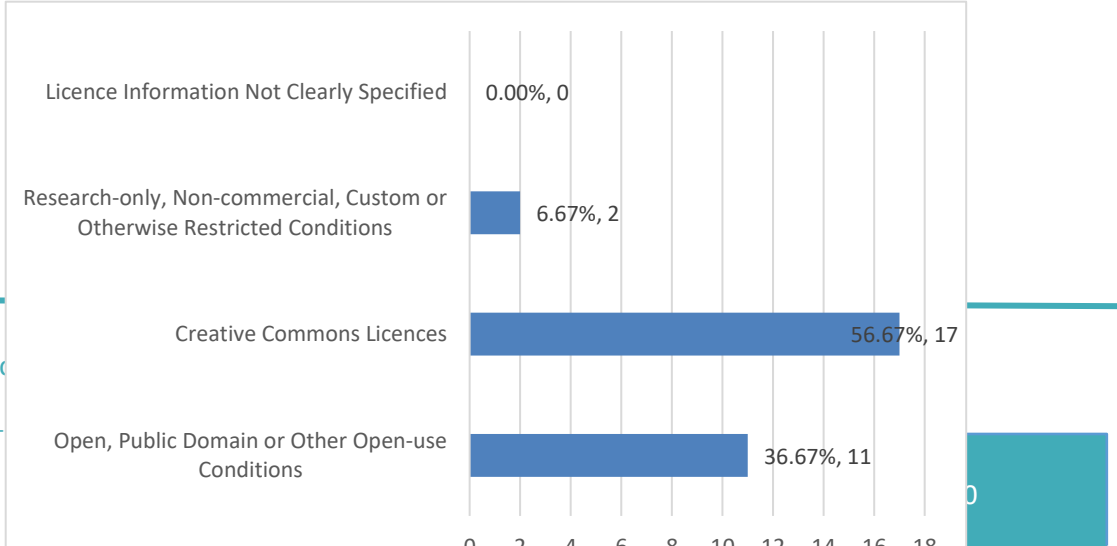


Figure 3.18. Distribution of numerical datasets by licence and conditions of use

Overall, the licensing structure of the numerical dataset collection is favourable for educational and research activities. Nevertheless, the specific conditions of use, including attribution requirements, restrictions on commercial use and institutional limitations, should be checked before datasets or derived materials are redistributed or included in publicly available teaching resources.

3.3.3 Dataset Scale: Number of Records and Total Size

The scale of the collected numerical datasets was analysed using two complementary parameters: the number of records and the total dataset size. The number of records indicates the amount of data available for model training and evaluation, while total size provides an indication of download, storage and processing requirements.

Number of Records

As shown in Figure 3.19, 4 datasets (13.33%) contain fewer than 1,000 records, while 6 datasets (20.00%) contain between 1,000 and 9,999 records. The largest group consists of 12 datasets (40.00%) containing between 10,000 and 99,999 records. Six datasets (20.00%) contain between 100,000 and 999,999 records, while 2 datasets (6.67%) contain one million or more records.

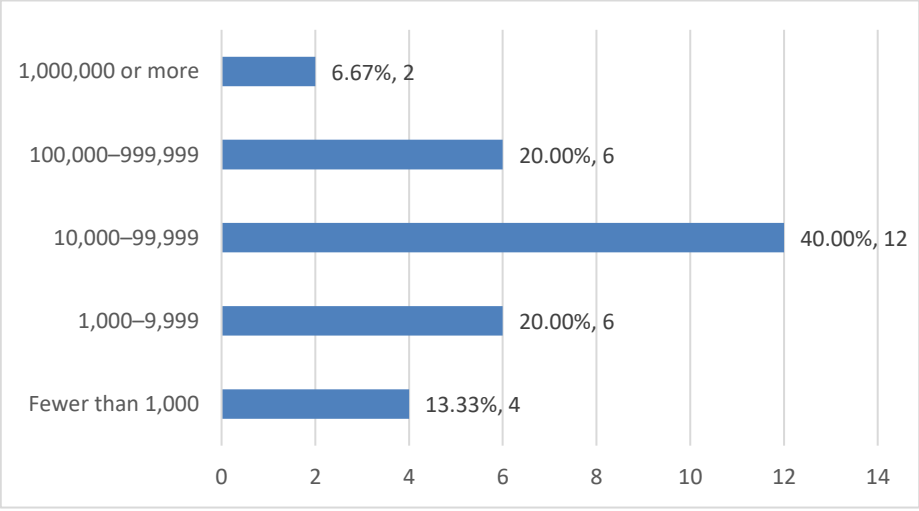


Figure 3.19. Distribution of numerical datasets by number of records

Datasets containing fewer than 10,000 records are generally easier to inspect and process during introductory practical classes. Medium-sized datasets support a broad range of classification, regression and exploratory data-analysis exercises, while datasets containing hundreds of thousands or millions of records are more appropriate for advanced projects and research-oriented activities.

Where a dataset was provided through several files, regions, cities or monthly releases, the reported number of records for the representative or complete collection was used based on the information available in the dataset collection document.

Total Dataset Size

The numerical datasets are generally smaller than the collected video and audio datasets. Nine datasets (30.00%) are smaller than 1 MB, while 8 datasets (26.67%) range from 1 to less than 10 MB. A further 8 datasets (26.67%) range from 10 to less than 100 MB. Four datasets (13.33%) have sizes between 100 MB and 1 GB, and only one dataset (3.33%) exceeds 1 GB.

Figure 3.20. Distribution of numerical datasets by total size

More than half of the numerical datasets are smaller than 10 MB and can therefore be considered highly manageable for educational use. Even most of the larger collections can be processed using standard computers, while the dataset exceeding 1 GB may require greater memory, longer processing time or the extraction of smaller subsets.

3.3.4 Data Formats and Technical Characteristics

The technical suitability of the numerical datasets was assessed through their main data formats,

number of features, feature types, presence of missing values and time-series properties. These characteristics influence compatibility with commonly used data-analysis tools and Python libraries, the level of preprocessing required and the types of educational activities that can be performed.

Data Formats

CSV is the dominant format and is used as the primary format in 21 datasets (70.00%). Five datasets (16.67%) use mixed tabular formats, such as CSV combined with TXT, TSV, XLSX or Parquet. Three datasets (10.00%) are primarily distributed in Excel formats, while one dataset (3.33%) is provided in TXT format.

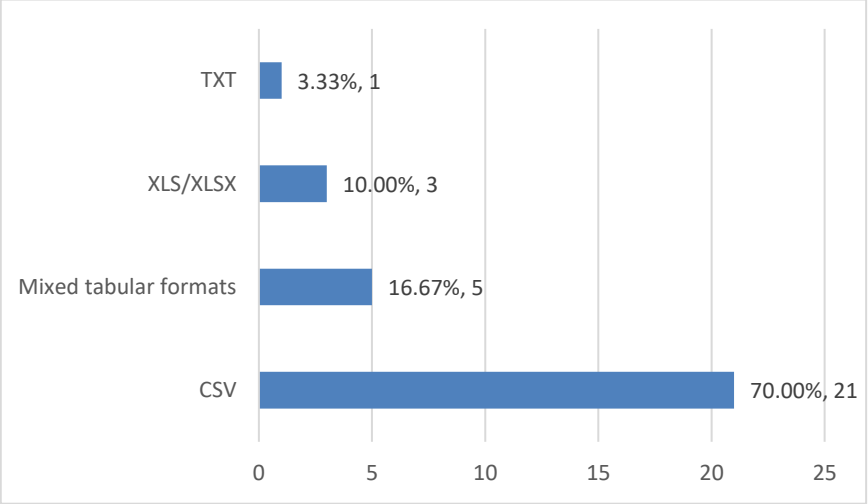


Figure 3.21. Distribution of numerical datasets by main data format

The predominance of CSV is favourable for educational use because this format is widely supported by spreadsheet software, statistical packages and Python libraries. Excel, TXT, TSV and Parquet formats are also compatible with commonly used tools, although mixed-format collections may require additional import or conversion steps.

Number of Features

The numerical datasets also differ substantially in dimensionality. Nine datasets (30.00%) contain 10 or fewer features, while 10 datasets (33.33%) contain between 11 and 20 features. Three datasets (10.00%) contain between 21 and 50 features, and 5 datasets (16.67%) contain between 51 and 100 features. The remaining 3 datasets (10.00%) contain more than 100 features.

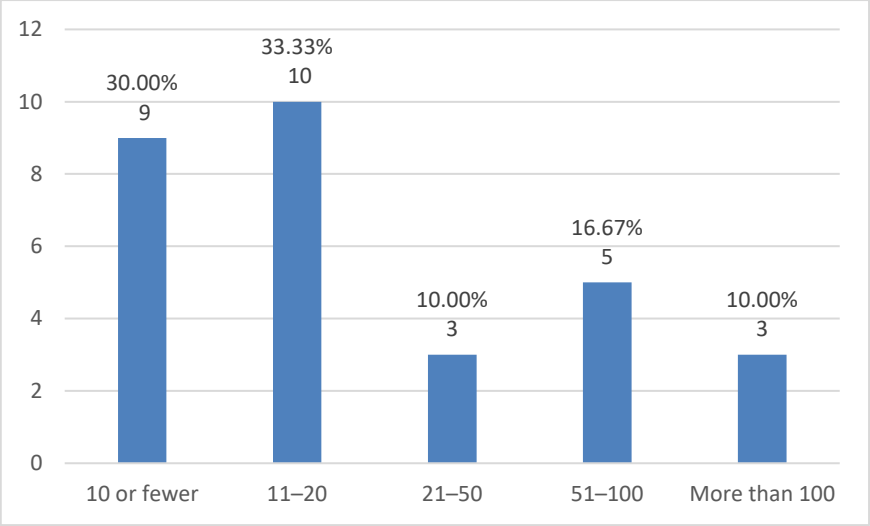


Figure 3.22. Distribution of numerical datasets by number of features

Datasets with a limited number of features are generally easier to understand, visualise and use in introductory exercises. Higher-dimensional datasets provide opportunities for teaching feature selection, dimensionality reduction and more advanced modelling techniques but may require additional preprocessing and interpretation.

Missing Values

Thirteen datasets (43.33%) contain no reported missing values. Four datasets (13.33%) contain only minimal or limited missing values, while 13 datasets (43.33%) include clearly reported missing or sparse data.

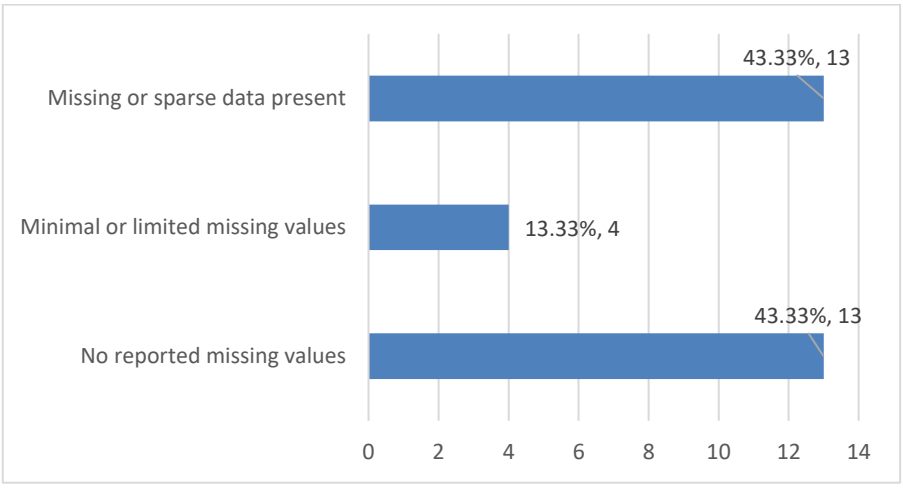


Figure 3.23. Distribution of numerical datasets by availability of missing values

Datasets without missing values are generally easier to integrate into introductory exercises. Datasets containing incomplete or sparse records provide useful opportunities for teaching data-

quality assessment, missing-value visualisation, deletion strategies and imputation methods.

Time-Series Properties

Thirteen datasets (43.33%) contain time-series or temporally ordered data, while 17 datasets (56.67%) are primarily cross-sectional or non-time-series collections.

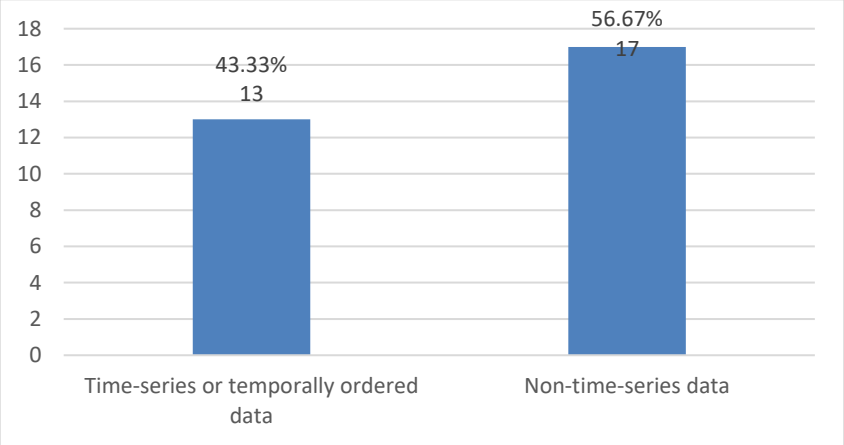


Figure 3.24. Distribution of numerical datasets by time-series properties

The presence of time-series data supports exercises involving trend analysis, seasonality, lag features and forecasting. Non-time-series datasets are suitable for classification, regression, clustering and exploratory data analysis. Datasets with different temporal frequencies may require sorting, aggregation, resampling or time-based train-test splitting.

3.3.5 Educational Applicability of Numerical Datasets

Numerical datasets provide a particularly accessible starting point for AI education because many of them are available in standard tabular formats and can be processed using spreadsheet software, statistical packages and commonly used Python libraries. However, a dataset that is easy to open is not necessarily simple to analyse. Its educational suitability also depends on the number and type of features, the presence of missing values, the clarity of the target variable, the temporal structure of the data and the amount of preparation required before modelling. Smaller and well-structured datasets are generally suitable for introducing the main stages of a machine-learning workflow, while larger, higher-dimensional and time-series collections provide opportunities for more advanced data preparation, modelling and evaluation.

For introductory practical exercises, datasets with a limited number of features, clearly defined target variables and no missing values are especially useful. The Energy Efficiency Dataset, for example, contains 768 records and eight numerical input features describing different building designs, with heating and cooling loads as target variables. Its small size, clearly defined

regression task and complete data make it suitable for introducing exploratory data analysis, correlation analysis, train-test splitting, regression modelling and performance evaluation. The California Housing Dataset provides a slightly larger but still manageable example, with 20,640 records, eight numerical features and no missing values. Since the dataset is intended for house-value prediction and is available in a standard tabular format, students can compare different regression methods and learn how model performance changes when working with a larger number of observations.

At the intermediate level, datasets can introduce missing values, temporal ordering or higher dimensionality. The Air Quality Dataset contains 9,358 hourly records from a gas multisensor device, together with reference pollutant concentrations. Missing values are recorded using the value −200, which provides a practical opportunity to teach data-quality assessment, missing-value identification and cleaning before regression or sensor-calibration tasks. Because the observations are ordered in time, the dataset can also introduce time-aware data splitting and the importance of preserving temporal order during model evaluation. The Human Activity Recognition Using Smartphones Dataset contains 10,299 records and 561 normalised features derived from smartphone accelerometer and gyroscope signals. It supports classification of six human activities and is suitable for teaching feature selection, dimensionality reduction and the comparison of classification models. Its large number of features makes it more demanding than a conventional introductory tabular dataset, even though its overall size remains manageable.

Advanced student projects and research-oriented activities can use large-scale, sparse or continuously updated datasets. The NYC Taxi Trip Record Dataset contains millions of trip records per month and includes pickup and drop-off times, locations, trip distances, fares and payment information. It can support research on demand analysis, travel-time or fare prediction and temporal patterns in urban mobility. Its monthly releases and Parquet or CSV formats make it possible to work with selected periods or progressively larger subsets, depending on the available computing resources. However, the full collection may require efficient file handling, timestamp processing, aggregation and scalable data-analysis methods. The Lending Club Loan Dataset is another advanced example, containing approximately 2.26 million accepted loan records and 151 borrower, loan and repayment attributes. Its size, mixed feature types and many sparse columns provide opportunities for advanced work on credit-risk classification, feature selection and missing-data treatment, but may also require extensive preprocessing and careful interpretation of model results. The World Development Indicators Dataset, with long-term data for 217 economies and approximately 1,400 development indicators, can support research involving panel data, macroeconomic trends and time-series analysis. Its broad scope and sparse structure mean that relevant indicators, countries and periods should normally be selected before analysis.

Table 3.3. Examples of numerical datasets suitable for different educational levels

Educational level	Dataset and suitable AI task	Reason for suitability	Required preparation or conditions
Introductory practical exercises	**Energy Efficiency Dataset** – prediction of building heating and cooling loads	768 records, eight numerical features, clearly defined targets and no missing values; CC BY 4.0 licence	Basic data inspection, scaling and train-test splitting
Introductory practical exercises	**California Housing Dataset** – house-value prediction	20,640 records, eight numerical features, no missing values and a clearly defined regression task; public domain	Basic exploratory analysis, scaling and comparison of regression models
Intermediate student projects	**Air Quality Dataset** – pollutant prediction and sensor calibration	Hourly time-series data with 9,358 records, 15 features and a realistic data-quality problem; CC BY 4.0 licence	Values coded as −200 need to be identified and treated as missing; temporal order should be considered
Intermediate student projects	**Human Activity Recognition Using Smartphones Dataset** – activity classification	10,299 records, 561 normalised features and six activity classes; CC BY 4.0 licence	High dimensionality may require feature selection or dimensionality reduction; TXT files may require structured import
Advanced or research-oriented activities	**NYC Taxi Trip Record Dataset** – demand analysis, fare prediction and mobility modelling	Millions of records per month, time and location information and regular monthly releases; public-domain data	Selection of relevant periods, timestamp processing, aggregation and efficient Parquet or CSV handling may be required
Advanced or research-oriented activities	**Lending Club Loan Dataset** – credit-risk classification and regression	Approximately 2.26 million records and 151 borrower, loan and repayment attributes; CC0 public-domain licence	Approximately 1.5 GB, with mixed feature types and many sparse columns; extensive cleaning and feature selection may be required

The proposed classification is intended as practical guidance rather than a fixed division. The same dataset may support different educational levels depending on how the activity is designed.

A small subset of the NYC Taxi Trip Record Dataset could be used in an intermediate project, while the complete multi-year collection would be more appropriate for advanced analysis. Similarly, a compact dataset can become conceptually demanding when it contains many feature types, missing values or data from a specialised domain. The Cleveland Heart Disease Dataset, for example, contains only 303 records and 13 features, but its medical context requires careful interpretation and makes it more suitable for demonstrating classification methods than for drawing clinical conclusions.

Dataset selection should therefore reflect the intended learning outcomes, students' previous experience, the complexity of the required preprocessing and the available computational resources. For time-series datasets, temporal ordering and appropriate train-test splitting should be considered, while datasets containing categorical or incomplete data may require encoding, cleaning or imputation. Licence conditions should also be verified before datasets are reused or incorporated into educational materials. The recommendations presented here are based on the information documented by dataset providers and recorded in the shared dataset collection document.

3.4 Image Datasets

The collected image datasets cover a broad range of applications relevant to Artificial Intelligence, machine learning and computer vision. They include general and fine-grained image classification, object detection, semantic segmentation, facial analysis, autonomous driving, medical imaging, remote sensing and agricultural applications. The datasets differ considerably in scale, formats, resolution, colour space, annotation level and licensing conditions. The following analysis examines these characteristics and their relevance for AI education and research.

3.4.1 Main Applications of Image Datasets

The collected image datasets support a diverse range of computer-vision tasks. As shown in Figure 3.25, the largest group consists of datasets primarily intended for image classification and fine-grained recognition, accounting for 11 datasets (36.67%). These datasets support tasks such as object, scene, vehicle, clothing and handwritten-digit classification.

Object detection, localisation and pose estimation represent the second-largest category, with 7 datasets (23.33%). Five datasets (16.67%) primarily support semantic segmentation and scene understanding. Face analysis and generative modelling, remote sensing and agricultural analysis, and fashion retrieval and image generation are each represented by 2 datasets (6.67%). One dataset (3.33%) is primarily intended for biomedical image analysis.

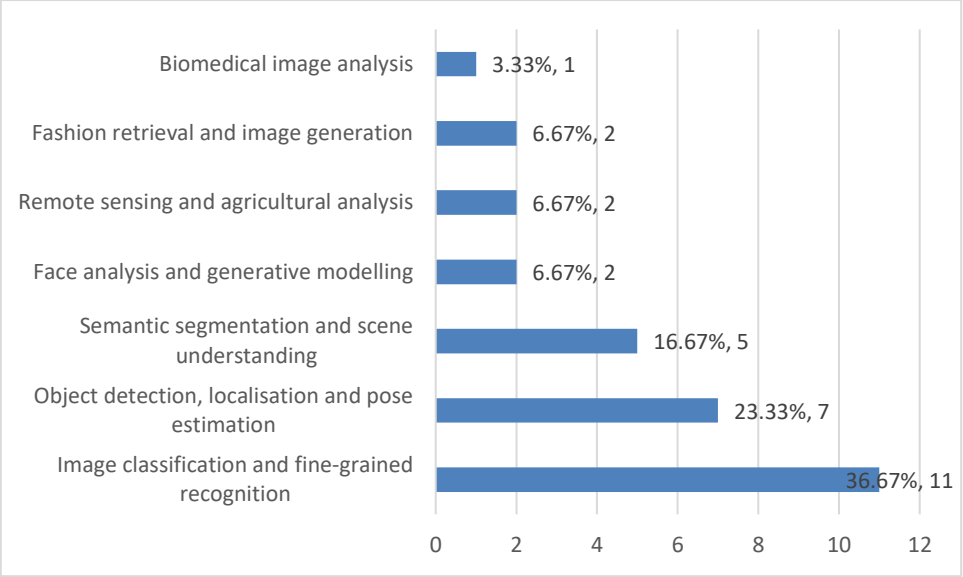


Figure 3.25. Distribution of image datasets by main application

This distribution demonstrates that the image dataset collection supports both widely used computer-vision tasks and more specialised educational and research applications.

For the purpose of the analysis, each dataset was assigned to one main application category according to its primary intended use. Although some datasets support several computer-vision tasks, only the dominant application was considered to avoid multiple counting. A detailed list of the datasets assigned to each main application category is provided in Annex 2.

3.4.2 Licence and Conditions of Use

The analysis of the licence information shows that the collected image datasets differ considerably in their conditions of use. As presented in Figure 3.26, Creative Commons licences represent the largest category, covering 11 datasets or 36.7% of the collection. This category includes attribution-based licences as well as Creative Commons licences containing non-commercial or share-alike conditions.

A further 10 datasets, representing 33.3% of the collection, are subject to separate non-commercial, academic, research-only, custom or otherwise restricted conditions. Nine datasets, or 30.0%, are available under public-domain or other open-use conditions, including CC0, MIT and other open research arrangements. Licence information or stated conditions of use were recorded for all 30 image datasets.

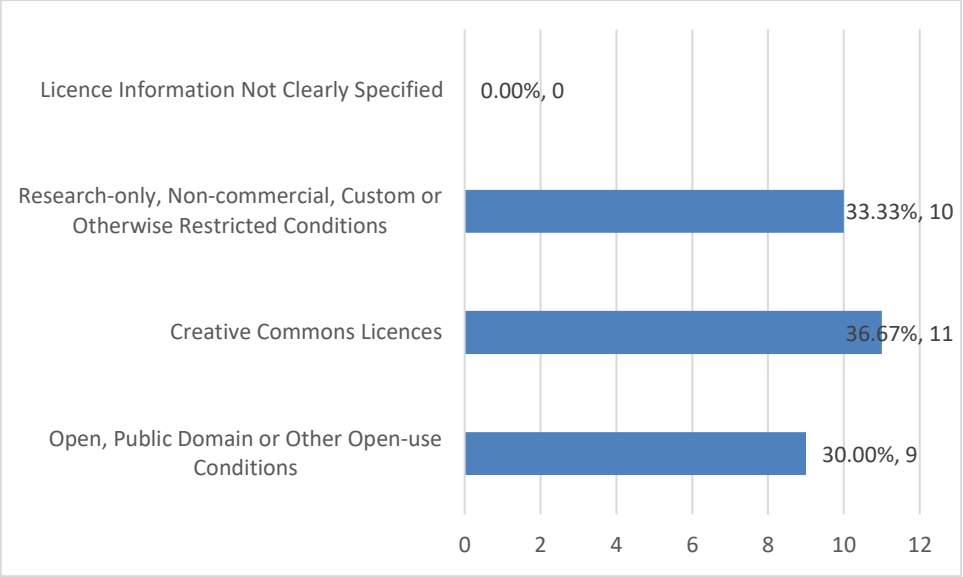


Figure 3.26. Distribution of image datasets by licence and conditions of use

Overall, the image dataset collection provides a broad range of resources for educational and research activities. However, the specific licence conditions should be checked before datasets or derived materials are redistributed or incorporated into publicly available teaching resources.

3.4.3 Dataset Scale: Number of Images and Total Size

The scale of the collected image datasets was analysed using two complementary parameters: the number of images and total dataset size. The number of images indicates the amount of material available for model training and evaluation, while total size provides an indication of download, storage and computational requirements.

Number of Images

As shown in Figure 3.27, 4 datasets (13.33%) contain fewer than 10,000 images, while the largest group, consisting of 14 datasets (46.67%), contains between 10,000 and 99,999 images. Nine datasets (30.00%) contain between 100,000 and 999,999 images, while 3 datasets (10.00%) contain one million or more images.

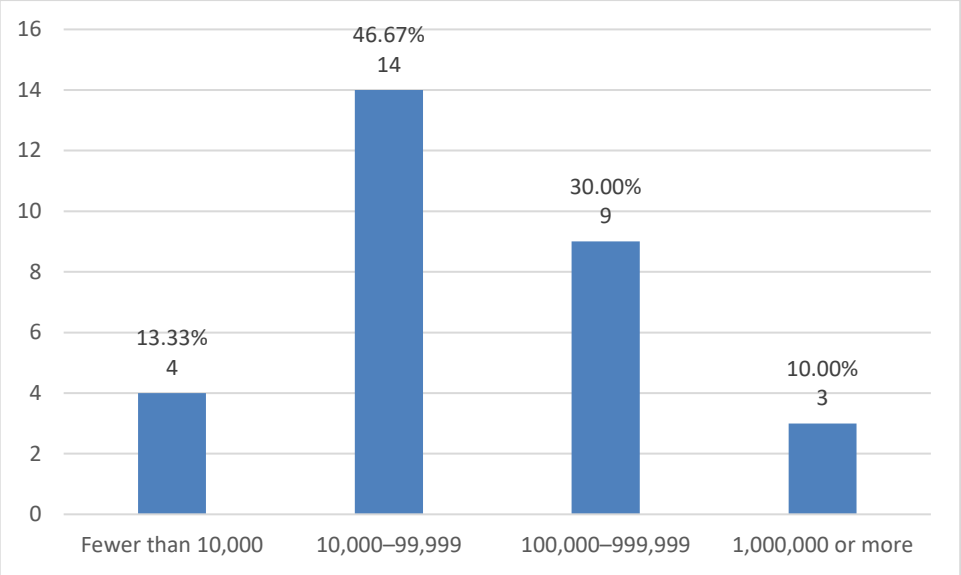


Figure 3.27. Distribution of image datasets by number of images

Smaller image datasets may be easier to manage in practical classes and student projects. Datasets containing tens or hundreds of thousands of images provide greater potential for model training and evaluation but generally require more preparation and computational resources. Collections containing millions of images are more appropriate for advanced projects or research-oriented activities.

Where datasets include frames, image patches or cropped objects rather than conventional image files, these units were treated as the closest equivalents to individual images.

Total Dataset Size

The total size of the collected image datasets also varies substantially. Eight datasets (26.67%) are smaller than 1 GB, while 9 datasets (30.00%) range from 1 to less than 10 GB. A further 9 datasets (30.00%) have sizes ranging from 10 to less than 100 GB, and 4 datasets (13.33%) range from 100 GB to 1 TB.

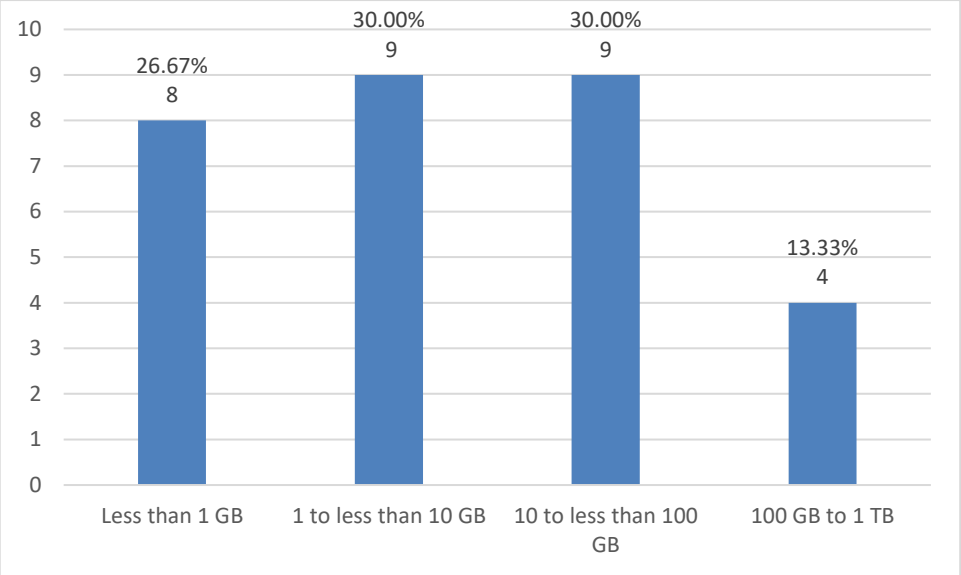


Figure 3.28. Distribution of image datasets by total size

More than half of the image datasets are smaller than 10 GB and can therefore be considered relatively manageable for educational use. Datasets between 10 and 100 GB may require the extraction of smaller subsets, while collections larger than 100 GB are generally more suitable for advanced student projects and research-oriented activities.

3.4.4 Image Formats and Technical Characteristics

The technical suitability of the collected image datasets was assessed through their main image formats, resolution, colour space and annotation level. These characteristics influence compatibility with commonly used computer-vision tools and Python libraries, the level of preprocessing required and the types of educational activities that can be performed.

Image Formats

JPEG is the dominant image format and is used as the primary format in 16 datasets (53.33%). PNG is the main format in 7 datasets (23.33%), while 3 datasets (10.00%) contain mixed standard image formats, such as JPEG combined with PNG or TIFF. Four datasets (13.33%) use proprietary, raw or specialised storage formats.

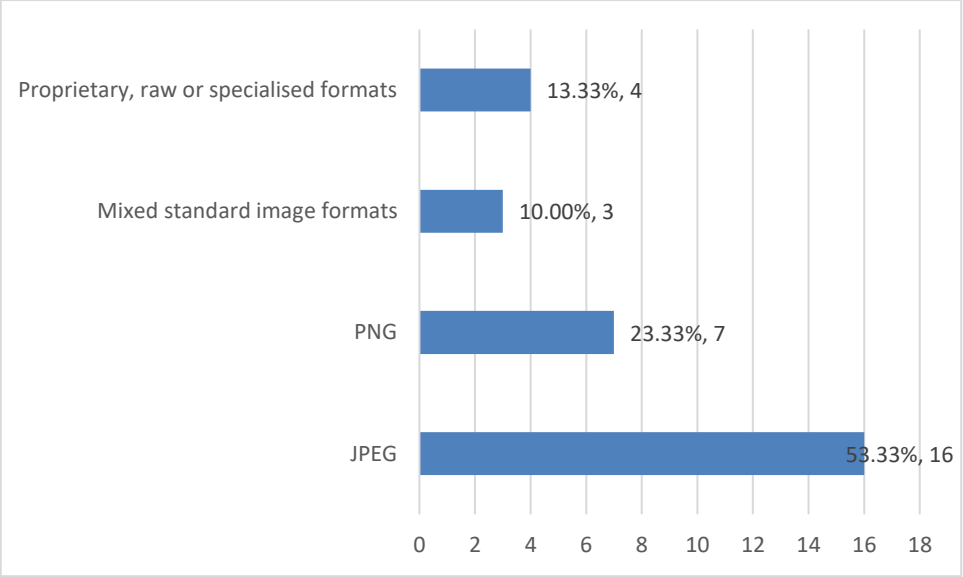


Figure 3.29. Distribution of image datasets by main data format

The predominance of JPEG and PNG is favourable for educational use because these formats are widely supported by computer-vision software and Python libraries. Proprietary, raw and specialised formats may require additional extraction, decoding or conversion before practical use.

Image Resolution

The image datasets contain a wide range of resolutions. Six datasets (20.00%) primarily contain low-resolution images, while 7 datasets (23.33%) mainly contain medium-resolution images. Five datasets (16.67%) contain high-resolution images, and 12 datasets (40.00%) include variable or multiple resolutions.

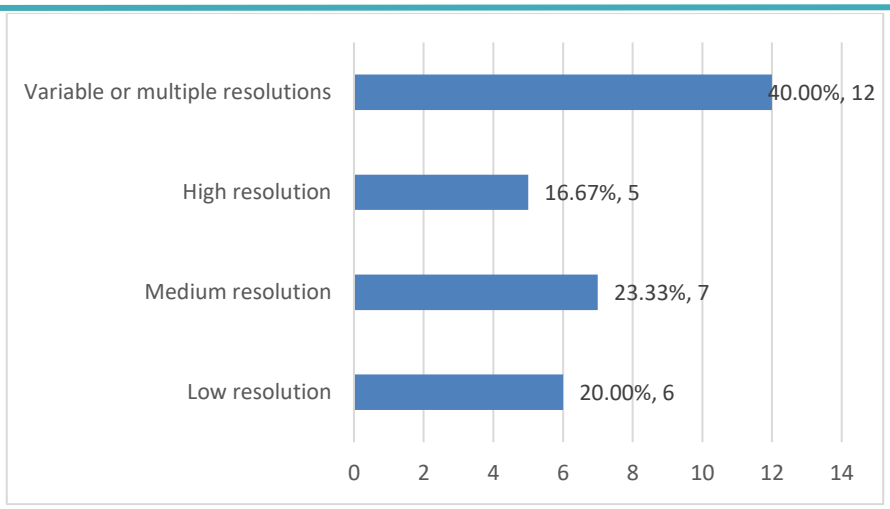


Figure 3.30. Distribution of image datasets by resolution category

Low-resolution datasets generally require fewer computational resources and may be suitable for introductory classification exercises. High-resolution and variable-resolution datasets provide richer visual information but may require resizing, cropping or other standardisation before model training.

Image Colour Space

RGB images dominate the collection and are used in 24 datasets (80.00%). Three datasets (10.00%) contain grayscale images. One dataset (3.33%) includes both RGB and grayscale images, one dataset (3.33%) contains multispectral images, and one dataset (3.33%) combines RGB images with depth information.

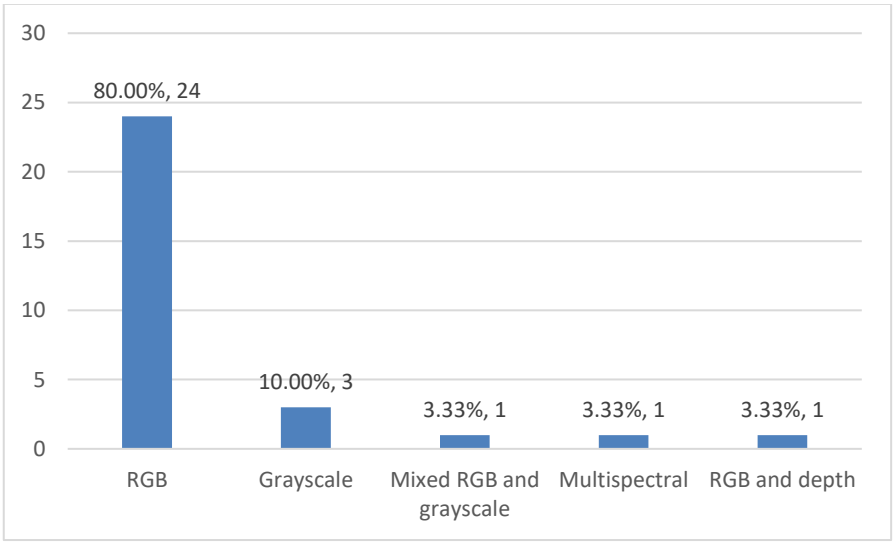


Figure 3.31. Distribution of image datasets by colour space

RGB images are compatible with most standard computer-vision models and tools. Grayscale images require fewer computational resources, although they may be converted to three-channel representations when used with models pre-trained on RGB images. Multispectral and RGB-depth datasets provide additional spectral or spatial information but generally require more specialised preprocessing and modelling approaches.

Annotation Level

The collected datasets also differ in the type and level of available annotations. Twelve datasets (40.00%) primarily provide image-level labels, such as object, scene, identity or class categories. Eleven datasets (36.67%) contain object-level annotations, including bounding boxes, landmarks, keypoints or pose coordinates. Seven datasets (23.33%) include pixel-level annotations for semantic, instance or panoptic segmentation.

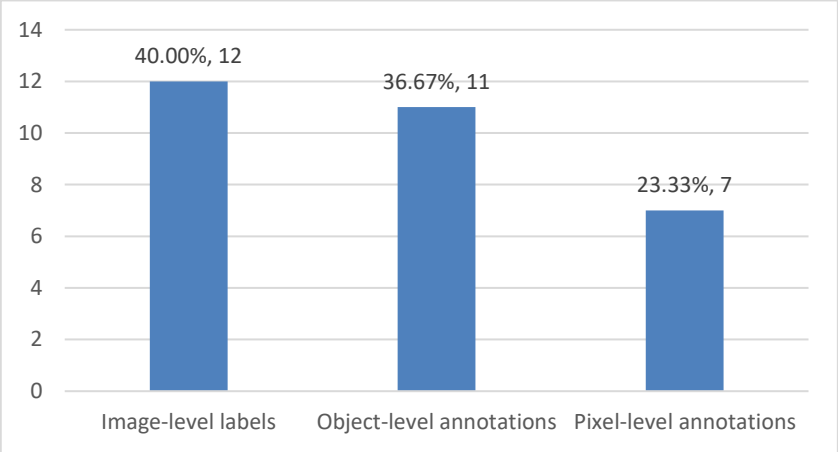


Figure 3.32. Distribution of image datasets by annotation level

Image-level labels are generally suitable for classification exercises, while bounding boxes, landmarks and keypoints support object detection, localisation and pose-estimation tasks. Pixel-level annotations enable semantic and instance segmentation but require more complex models and greater computational resources.

Where datasets provide several annotation types, they were classified according to the most detailed or primary annotation level used for their main intended application.

3.4.5 Educational Applicability of Image Datasets

The collected image datasets support a broad range of educational activities, from basic image classification to object detection, semantic segmentation, pose estimation and three-dimensional scene understanding. Their level of difficulty is not determined by the number of images alone. A collection containing many small, standardised images with image-level labels can be easier to use than a much smaller collection containing variable-resolution images,

bounding boxes or pixel-level segmentation masks. Educational suitability therefore depends on image resolution and format, annotation type, dataset size, technical complexity, required preprocessing and the computational resources available for model training.

For introductory practical exercises, standardised datasets with clearly defined image-level classes are particularly suitable. CIFAR-10, for example, contains 60,000 RGB images of 32 × 32 pixels organised into ten classes, with predefined training and test sets. Its small total size and uniform PNG format allow students to focus on the main stages of an image-classification workflow, including image loading, normalisation, model training and performance evaluation. The PlantVillage Dataset provides a more application-oriented example. It contains 54,306 RGB images with a standardised resolution of 256 × 256 pixels and 38 image-level labels representing crop species and disease categories. Its folder-based structure, JPEG format and public-domain licence make it suitable for introductory classification exercises and initial experiments with transfer learning.

At the intermediate level, students can work with datasets containing more detailed annotations and variable image characteristics. The Oxford-IIIT Pet Dataset includes 7,349 images of 37 cat and dog breeds, together with breed labels, head-region bounding boxes and pixel-level trimap segmentations. This allows the same dataset to support a gradual progression from fine-grained classification to object localisation and semantic segmentation. Because image resolutions vary and annotations are provided in XML and PNG formats, students may need to perform resizing and annotation processing before model training. PASCAL VOC 2012 offers a further step towards realistic computer-vision projects. It contains 11,530 images, 27,450 annotated object instances, bounding-box coordinates and pixel-level segmentation masks. The dataset can support object detection and semantic-segmentation activities, while its XML annotations and variable image resolutions provide useful experience in preparing heterogeneous image data.

Datasets with specialised image channels can also be introduced at the intermediate or advanced level. The EuroSAT Dataset, for example, contains 27,000 satellite-image patches for land-use and land-cover classification. Its RGB JPEG version can support relatively accessible classification exercises, while the 13-band multispectral TIFF version introduces more advanced issues such as handling additional spectral channels, 16-bit image data and remote-sensing-specific preprocessing. This makes EuroSAT a useful example of how the same dataset can support different educational levels depending on the selected data representation.

Advanced student projects and research-oriented activities can make use of larger collections with complex annotation structures or combined two-dimensional and three-dimensional information. The COCO Dataset contains approximately 328,000 images and supports object detection, instance segmentation, keypoint estimation and image captioning. Its variable image resolutions and extensive JSON annotations allow students to work on realistic deep-learning

tasks, but may require subset extraction, annotation parsing and substantial computational resources. The KITTI-360 Dataset is even more technically demanding, with more than 320,000 frames, high-resolution fisheye images and unified two-dimensional and three-dimensional semantic instance annotations. Its structured sequences and PLY-based 3D data make it relevant for autonomous-driving research, 3D scene understanding and novel-view synthesis, although working with the complete 150 GB collection may require dedicated processing tools and significant storage and computing capacity. Similarly, SUN RGB-D, which combines RGB images with depth information and 3D-oriented bounding boxes, can support advanced research on indoor-scene understanding and 3D object detection.

Table 3.4. Examples of image datasets suitable for different educational levels

Educational level	**Dataset and suitable AI task**	**Reason for suitability**	**Required preparation or conditions**
Introductory practical exercises	**CIFAR-10** – image classification	60,000 standardised 32 × 32 RGB images, ten clearly defined classes and predefined training and test sets; CC0/public-domain licence	Basic normalisation, optional data augmentation and model training
Introductory practical exercises	**PlantVillage Dataset** – crop and plant-disease classification	54,306 standardised 256 × 256 JPEG images organised into 38 image-level classes; CC0 1.0 public-domain licence	Preparation of training, validation and test subsets may be required
Intermediate student projects	**Oxford-IIIT Pet Dataset** – fine-grained classification and semantic segmentation	Manageable size with breed labels, bounding boxes and pixel-level trimap annotations; CC BY-SA 4.0 licence	Variable image resolutions, XML annotation processing, resizing and mask preparation
Intermediate student projects	**PASCAL VOC 2012** – object detection and semantic segmentation	11,530 images with bounding boxes and pixel-level segmentation masks; public-domain/open-research conditions	Parsing XML annotations, processing segmentation masks and standardising image dimensions
Advanced or research-oriented activities	**COCO Dataset** – object detection, instance segmentation, keypoint estimation and captioning	Approximately 328,000 images with several detailed annotation types; CC BY 4.0 licence	Approximately 25 GB, variable resolutions and complex JSON annotations; subset extraction and substantial training resources may be required
Advanced or	**KITTI-360** – 3D	More than 320,000 frames	Approximately 150 GB,

research-oriented activities	scene understanding and autonomous-navigation research	with high-resolution images and unified 2D/3D semantic instance annotations	structured sequences and PLY-based 3D data; dedicated processing tools and significant computing resources may be required

The proposed classification is indicative rather than fixed. CIFAR-10 contains substantially more images than the Oxford-IIIT Pet Dataset, but its uniform resolution and simple image-level labels make it easier to use. Conversely, the Oxford-IIIT Pet Dataset is smaller but becomes more demanding when its bounding boxes and segmentation masks are included. Similarly, a selected subset of COCO could support an intermediate object-detection project, while the complete collection is more appropriate for advanced model development and research.

Dataset selection should therefore reflect the intended learning outcomes, the type of annotation required for the selected AI task and the available computational resources. Image-level labels are generally sufficient for classification exercises, while bounding boxes, keypoints and pixel-level masks are needed for detection, pose-estimation and segmentation tasks. High-resolution, multispectral, RGB-depth and 3D datasets may require resizing, channel transformation, specialised file handling or the use of smaller subsets. Licence and access conditions should also be verified before images or annotations are reused or included in educational materials. The recommendations presented here are based on the characteristics documented by the dataset providers and recorded in the shared dataset collection document.

4. Comparative Assessment and Educational Applicability

The analysis shows that the four dataset modalities collectively offer broad potential for AI education and research, although the suitability of individual datasets differs depending on dataset scale, technical complexity, format, licensing conditions and available computational resources.

The comparative assessment therefore focuses on the main similarities and differences between video, audio/sound, numerical and image datasets, as well as the requirements associated with their use in practical exercises, student projects and research-oriented activities. Since the basic data units differ across modalities, the number of videos, audio files, numerical records and images should not be compared directly. Instead, the assessment considers overall manageability, technical requirements, preprocessing needs and conditions of use.

4.1 Comparative Findings

The complete collection comprises 120 datasets, equally distributed across four data modalities: video, audio/sound, numerical and image datasets. Although each modality is represented by 30 datasets, the numbers of individual data items cannot be directly compared because a video recording, an audio file, an image and a numerical record represent fundamentally different units of data.

The consolidated comparison therefore focuses only on characteristics that were recorded and analysed across all four modalities and that can be interpreted consistently. These include licence and access conditions, availability of information on total dataset size, the use of standard or specialised data formats, general preprocessing requirements, relative computational demands and potential educational suitability. Modality-specific parameters, such as video frame rate, audio sampling rate, the number of numerical features and image colour space, are not combined because they are not directly comparable.

Licence and Access Conditions

For the cross-modal comparison, Creative Commons attribution-based licences, including

variants with non-commercial, share-alike or no-derivatives conditions, were grouped into one category. CC0 licences were included under open, public-domain or other open-use conditions. Separate research-only, academic, non-commercial, custom or otherwise restricted conditions were included in a distinct category.

Table 4.1. Consolidated overview of licence conditions across the dataset collection

Licence category	Video	Audio/ Sound	Numerical	Image	Total
Open, Public Domain or Other Open-use Conditions	4	2	11	9	26
Creative Commons Licences	12	22	17	11	62
Research-only, Non-commercial, Custom or Otherwise Restricted Conditions	9	5	2	10	26
Licence Information Not Clearly Specified	5	1	0	0	6
Total	**30**	**30**	**30**	**30**	**120**

Creative Commons licences represent the largest category, covering 62 datasets or 51.7% of the complete collection. A further 26 datasets, or 21.7%, are available under public-domain or other open-use conditions. Together, these two categories account for 88 datasets, corresponding to 73.3% of the collection.

Another 26 datasets, or 21.7%, are subject to separate research-only, non-commercial, custom or otherwise restricted conditions. Licence information was not clearly specified for 6 datasets, representing 5.0% of the collection. These cases occur mainly among the video datasets.

The results show that a substantial part of the collection can support educational and research activities. However, datasets within the same broad category do not necessarily provide identical reuse rights. In particular, Creative Commons licences may include attribution, non-commercial, share-alike or no-derivatives conditions. The specific licence terms should therefore be checked before datasets or derived materials are redistributed or incorporated into publicly available teaching resources.

Dataset Size and Format Characteristics

Information on total dataset size was available for 117 of the 120 datasets. It was recorded for 28 video datasets, 29 audio/sound datasets, and all 30 numerical and 30 image datasets. This provides a generally strong basis for assessing download, storage and processing requirements. Nevertheless, dataset sizes should not be compared using identical thresholds across modalities. A dataset considered large within the numerical group may still require substantially less storage than a medium-sized video, audio or image collection. The modality-specific size categories presented in Chapter 3 are therefore more appropriate for detailed assessment, while the consolidated comparison is limited to the availability of size information and general resource requirements.

Standard and widely supported formats are dominant in all four groups. MP4/MPEG-4 and AVI are commonly represented among video datasets, WAV, MP3 and FLAC among audio and sound datasets, CSV and related tabular formats among numerical datasets, and JPEG and PNG among image datasets. Their prevalence facilitates integration into commonly used data-processing and Machine Learning workflows.

At the same time, the collection includes mixed, specialised and multimodal structures. Examples include ROS bag, TFRecord and CSD/HDF5 in the video group; combinations of different audio formats, MIDI, STEMS and TFRecord files in the audio/sound group; mixed tabular formats and time-series structures among numerical datasets; and raw, proprietary or annotation-specific structures among image datasets. These formats increase the diversity and research value of the collection but may require additional extraction, conversion or specialised processing.

Comparative Assessment of Practical Requirements

Table 4.2 provides a qualitative synthesis of the general patterns identified in Sections 3.1-3.4. The entries were derived from the modality-specific analyses of dataset size, formats, technical characteristics, preprocessing requirements and educational applicability. They do not represent a formal scoring of each individual dataset.

Table 4.2. Comparative overview of common dataset characteristics

Comparative characteristic	Video	Audio/Sound	Numerical	Image
Dataset size information available	28 of 30	29 of 30	30 of 30	30 of 30
Typical dataset size	Highly variable; often medium to high	Highly variable; often medium to high	Generally low	Highly variable; from low to high

Main standard formats represented	MP4/MPEG-4 and AVI	WAV, MP3 and FLAC	CSV, XLSX and TXT	JPEG, PNG and mixed standard image formats
Examples of mixed or specialised formats and structures	ROS bag, TFRecord, CSD/HDF5 and multimodal sensor data	Mixed audio formats, MIDI, STEMS and TFRecord	Mixed tabular formats and time-series structures	Raw, proprietary and annotation-specific structures
Typical preprocessing	Frame extraction and sampling, temporal segmentation, resizing, and frame-rate or format standardisation	Segmentation, resampling, channel standardisation, normalisation and feature extraction	Missing-value treatment, categorical encoding, scaling and time-aware splitting	Resizing, normalisation, colour-space conversion and annotation transformation
Relative computational requirements	Generally high	Generally moderate to high	Generally low to moderate	Generally moderate to high
Typical educational suitability	Mainly intermediate, advanced and research-oriented; subsets may support introductory use	Introductory to advanced, depending on size and technical consistency	Particularly suitable for introductory and intermediate activities	From introductory classification to advanced detection and segmentation
Main practical limitations	Large size, long processing time and multimodal complexity	Variable duration, sampling rates, channels and labelling practices	Missing values, heterogeneous variables and temporal organisation	Large collections, high resolution and complex annotations

Numerical datasets generally provide the most accessible entry point for introductory AI and Machine Learning activities. Their predominantly tabular structure, widespread use of standard formats and comparatively modest storage requirements allow students to focus on data exploration, preprocessing, classification, regression and time-series analysis.

Image and audio/sound datasets support a broad progression from introductory to advanced activities. Smaller and technically consistent collections can be used for basic image or sound

classification, while larger datasets, detailed annotations and heterogeneous technical characteristics support more demanding tasks such as object detection, segmentation, speech recognition, music analysis and acoustic-event detection.

Video datasets generally require substantial preprocessing effort and computational resources because of their temporal structure, frame-level processing and, in some cases, multimodal complexity. Audio/sound datasets may also involve considerable storage requirements, particularly when they contain large numbers of long or high-quality recordings. Large image collections can similarly require substantial storage and processing capacity. Where complete collections are too large or technically demanding, smaller representative subsets may be prepared for classroom use.

Across all modalities, educational suitability is not determined by data type alone. Dataset size, licence conditions, format, technical consistency, documentation, availability of labels or annotations, preprocessing requirements and available computational resources should be considered together. Standard-format and well-documented datasets are generally easier to integrate into introductory and intermediate teaching, while large, multimodal or technically specialised datasets are more appropriate for advanced student projects and research-oriented activities.

4.2 Datasets Suitable for Different Educational Uses

The collected datasets can support different levels and types of educational activities, depending on their scale, technical complexity, format, licence conditions and preprocessing requirements. For the purpose of this assessment, the educational uses were grouped into three broad categories: introductory practical exercises, intermediate student projects and advanced or research-oriented activities.

Smaller datasets in standard formats are generally the most suitable for introductory practical exercises. Numerical datasets in CSV or spreadsheet formats can support activities involving data exploration, visualisation, classification and regression. Smaller image datasets with standardised resolutions and image-level labels are suitable for basic image-classification exercises, while audio datasets containing short recordings in WAV format can be used for introductory sound-classification and speech-recognition tasks. Small video datasets or extracted subsets may also be used, although they generally require more preparation and computational resources.

Intermediate student projects can use datasets that introduce a moderate level of preprocessing, more complex structures, variable technical characteristics or more detailed annotations. Suitable activities include time-series analysis with numerical data, object detection and segmentation with image datasets, sound-event or emotion recognition with audio datasets, and

action recognition or tracking with video datasets. These projects allow students to compare different preprocessing approaches and models and to evaluate performance using more realistic data.

Large-scale, multimodal and technically specialised datasets are most appropriate for advanced student projects and research-oriented activities. They can support deep-learning model development, multimodal learning, high-resolution computer vision, large-scale speech or sound analysis, autonomous-driving applications and other computationally demanding tasks. Their use generally requires greater storage and processing capacity, as well as more extensive preprocessing, standardisation or subset extraction.

Table 4.3. Dataset characteristics suitable for different educational uses

Educational use	Suitable dataset characteristics	Typical examples
Introductory practical exercises	Small or manageable size, standard formats, limited preprocessing and clearly defined labels or target variables	Tabular classification and regression, basic image classification, short-audio classification and small video-classification tasks
Intermediate student projects	Moderate size, more complex features or annotations and a meaningful level of data cleaning or standardisation	Time-series analysis, high-dimensional classification, object detection, semantic segmentation, sound-event recognition, emotion recognition and action recognition
Advanced or research-oriented activities	Large-scale, multimodal or specialised data, extensive preprocessing and higher computational requirements	Deep learning, multimodal analysis, autonomous systems, large-scale speech processing, high-resolution vision and scalable data analysis

To provide more concrete guidance for dataset selection, Table 4.4 presents a representative selection from the complete collection. The recommendations are based on the dataset characteristics analysed in Chapter 3, including scale, format, technical consistency, labels or annotations, preprocessing requirements and licence conditions. The table is not intended to represent an exhaustive ranking of all 120 datasets, and some datasets may be suitable for more than one educational level depending on the selected subset, task and available computational resources.

Table 4.4. Representative datasets recommended for different educational uses

Educational	Dataset	Modality	Suitable AI	Reason for	Important

level			task	recommendation	limitation or preparation requirement
Introductory practical exercises	**Highway Traffic Videos Dataset**	Video	Traffic-level video classification	Small collection of short videos with three clearly defined classes; supports a complete introductory video-classification workflow	Requires frame extraction and basic video preprocessing; low resolution limits more detailed vision tasks
Introductory practical exercises	**Free Spoken Digit Dataset**	Audio/ Sound	Spoken-digit classification	Small size, short mono WAV recordings and clearly defined digit labels; suitable for waveform analysis, spectrograms and basic classification	Basic feature extraction or spectrogram generation is required
Introductory practical exercises	**Energy Efficiency Dataset**	Numerical	Regression	Small, complete and well-structured dataset with clearly defined input features and two target variables	Feature scaling and the interpretation of two related target variables may be required
Introductory practical exercises	**CIFAR-10**	Image	Image classification	Standardised low-resolution images, ten clearly defined classes and predefined training and test sets	Low image resolution limits applicability to more complex detection or segmentation tasks
Intermediate student projects	**CBVD-5 Cow Behavior Video Dataset**	Video	Cow-behaviour recognition, classification and tracking	Contains 887 annotated Full HD videos and supports realistic computer-vision applications in livestock monitoring; CC0 public-domain licence	Frame sampling, resizing and annotation processing may be required; Full HD recordings increase computational requirements
Intermediate	**UrbanSoun**	Audio/Sound	Urban	Clearly labelled	Variable

student projects	**d8K**		sound-event classification	excerpts from ten classes and sufficient scale for comparing audio features and classification models	sampling rates require resampling and technical standardisation; the CC BY-NC 3.0 licence permits non-commercial use only.
Intermediate student projects	**Air Quality Dataset**	Numerical	Regression, sensor calibration and time-series analysis	Introduces missing values, temporal ordering and realistic environmental sensor data	Missing values coded as −200 must be identified and treated; chronological splitting is recommended
Intermediate student projects	**Oxford-IIIT Pet Dataset**	Image	Fine-grained classification, localisation and segmentation	Supports progression from breed classification to bounding-box localisation and pixel-level segmentation	Variable image sizes and XML/PNG annotations require resizing and annotation processing
Advanced or research-oriented activities	**FieldSAFE Dataset**	Video	Autonomous-system perception and sensor fusion	Combines video and other sensor streams and reflects realistic autonomous-system data-processing challenges	ROS bag structure requires specialised software, sensor synchronisation and substantial preprocessing
Advanced or research-oriented activities	**LibriSpeech**	Audio/Sound	Automatic speech recognition	Large-scale speech corpus with transcripts and approximately 1,000 hours of audio; appropriate for deep speech-processing models	High storage and computational requirements; subset selection may be necessary
Advanced or research-oriented activities	**NYC Taxi Trip Record Dataset**	Numerical	Mobility-demand analysis, travel-time	Millions of records and recurring releases support scalable data	Requires efficient file handling, timestamp

			prediction and large-scale temporal modelling	analysis and research on urban mobility	processing, aggregation and usually a selected time period
Advanced or research-oriented activities	**KITTI-360**	Image	Autonomous-driving perception, semantic understanding and 3D scene analysis	Provides complex visual and spatial information suitable for advanced computer-vision and autonomous-driving research	Large scale, multimodal structure and complex annotations require substantial storage, preprocessing and computing resources; the CC BY-NC-SA 4.0 licence includes non-commercial and share-alike conditions.

The introductory recommendations prioritise manageable size, standard formats and clearly defined tasks. They are therefore suitable for teaching the main stages of a Machine Learning workflow without allowing technical preparation to dominate the activity. For example, the Energy Efficiency Dataset and CIFAR-10 allow students to focus directly on regression and classification, while the Free Spoken Digit Dataset and Highway Traffic Videos Dataset introduce audio and video processing through relatively small collections. The individual analyses in Chapter 3 already identify these characteristics as particularly useful for introductory work.

The intermediate recommendations introduce realistic challenges such as variable sampling rates, missing values, temporal organisation, frame or clip processing and more detailed annotations. They remain manageable for student projects but require students to make and justify decisions about preprocessing, model selection and evaluation. CBVD-5, UrbanSound8K, the Air Quality Dataset and the Oxford-IIIT Pet Dataset are especially useful because they introduce realistic requirements related to frame processing, annotation handling, resampling, missing-value treatment and temporal organisation without requiring research-scale infrastructure.

The advanced recommendations were selected because they involve large-scale, multimodal or technically specialised data. FieldSAFE and KITTI-360 support autonomous-system and multimodal perception activities, LibriSpeech supports large-scale speech recognition, and the

NYC Taxi Trip Record Dataset supports scalable numerical and temporal analysis. In each case, the educational value comes partly from addressing realistic challenges in data organisation, preprocessing, resource management and model development.

These recommendations should be treated as guidance rather than fixed assignments. A large dataset can be adapted to an introductory or intermediate activity by preparing a smaller and clearly documented subset, while a comparatively small dataset can support advanced work when the task involves complex modelling, detailed evaluation or methodological comparison. Before selecting a dataset for a specific educational activity, the intended learning outcomes, available time and computing resources, data documentation and applicable licence conditions should be reviewed.

4.3 Datasets Selected for Subsequent Project Activities

4.3.1 Selection Process and Criteria

Following the collection and analysis of 120 datasets, each of the four partners responsible for a particular data modality selected three datasets for subsequent project activities. This resulted in a final selection of 12 datasets, comprising three video, three audio/sound, three numerical and three image datasets. The selected datasets or appropriately prepared subsets were uploaded to the project repository for further technical preparation and the development of Python-based educational materials.

The selection considered the educational relevance of each dataset, the availability of labels, annotations or target values, data format and technical compatibility, dataset size and manageability, licence and conditions of use, and the level of preprocessing required. The partners also considered whether the datasets could support clearly defined AI tasks and be adapted to practical exercises, student projects or more advanced project activities.

4.3.2 Overview of the Selected Datasets

Table 4.5. Overview of datasets selected for subsequent project activities

Partner	Dataset	Modality	Uploaded version	Intended AI task	Labels/targets	Format	Repository size	Licence
UNILJ								
UNILJ								
UNILJ								
UoM								
UoM								

UoM								
UNSA								
UNSA								
UNSA								
UNBG								
UNBG								
UNBG								
UNBG								

Legend:
Partner - partner responsible for selecting and uploading the dataset to the project repository.
Dataset - full name of the selected dataset.
Modality - data type: video, audio/sound, numerical or image.
Uploaded version - indicates whether the full dataset or a prepared subset was uploaded.
Intended AI task - AI task for which the dataset will be used, such as classification, regression, detection, segmentation or time-series analysis.
Labels/targets - available labels, annotations or target values used for model training and evaluation.
Format - main file format of the uploaded dataset version.
Repository size - size of the dataset version uploaded to the project repository.
Licence - licence and the main conditions governing the use of the dataset.

4.3.3 Technical Characteristics and Required Preprocessing

COMMON TEXT - UoM

UNILJ Table 4.6. Technical characteristics of the selected video datasets

No	***Dataset***	*????*	*????*	*????*	**Reason for selection**	**Required preprocessing**
1						
2						
3						

UoM Table 4.7. Technical characteristics of the selected audio/sound datasets

No	***Dataset***	*????*	*????*	*????*	**Reason for selection**	**Required preprocessing**
1						
2						
3						

UNSA Table 4.8. Technical characteristics of the selected numerical datasets

No	***Dataset***	*????*	*????*	*????*	**Reason for selection**	**Required preprocessing**
1						
2						
3						

UNBG Table 4.9. Technical characteristics of the selected image datasets

No	***Dataset***	*????*	*????*	*????*	**Reason for selection**	**Required preprocessing**
1						
2						
3						

4.3.4 Readiness for Subsequent Project Activities

Each partner should write something about the datasets they selected

5. Conclusion – TO BE REWIEVED - UOM

This report presented a structured analysis of 120 datasets collected within WP3-A1, comprising 30 video, 30 audio/sound, 30 numerical and 30 image datasets. The analysis examined their main application areas, conditions of use, scale, data formats, modality-specific technical characteristics and potential applicability in AI education and research.

The results demonstrate that the collected datasets cover a broad range of Artificial Intelligence and Machine Learning tasks, including classification, regression, object detection, segmentation, speech and sound recognition, time-series analysis, action recognition, autonomous driving and multimodal learning. Standard and widely supported formats are well represented across all four modalities, facilitating the use of commonly available data-processing tools, Python libraries and machine-learning frameworks. At the same time, the collection includes large-scale, technically specialised and multimodal datasets that may require additional preprocessing, conversion, standardisation or the extraction of smaller subsets before practical use.

The comparative assessment shows that there is no single dataset modality that is most suitable for all educational purposes. Numerical datasets are generally easier to manage and particularly suitable for introductory exercises involving data exploration, visualisation, classification and regression. Smaller image and audio/sound datasets can support introductory and intermediate activities, while video datasets, high-resolution image collections and large or technically

heterogeneous datasets offer greater potential for advanced student projects and research-oriented applications. The practical suitability of an individual dataset therefore depends on the intended learning outcomes, student level, available time and computational resources, as well as the required level of data preparation.

Licensing and conditions of use represent an additional factor that should be considered when selecting datasets for educational and project activities. Although many datasets are available under open or clearly defined licences, others are restricted to academic, research or non-commercial use or are subject to custom conditions. These conditions should be verified before datasets are incorporated into teaching materials, shared repositories or publicly distributed project outputs.

The preparation of this report was coordinated by UoM and relied on the contributions of the partners responsible for the respective data modalities. UNILJ contributed the information and expertise related to video datasets, UoM to audio and sound datasets, UNBG to image datasets and UNSA to numerical datasets. UNIRI supported the process through the review of the collected datasets and advice concerning the structure and formats of the collection. The consolidated findings were shared with the relevant partners for verification and comments.

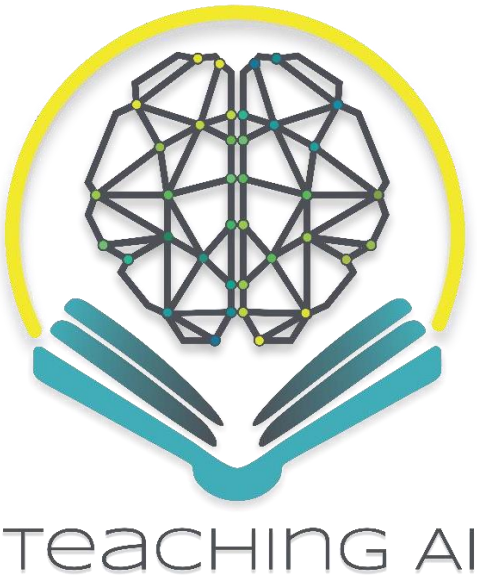
Overall, the analysis confirms that the collected datasets provide a diverse and relevant basis for the further activities of the TAI project. They offer opportunities for developing practical exercises, student projects and research-oriented learning activities across different areas of Artificial Intelligence. The findings of this report supported the selection of appropriate datasets for subsequent technical preparation, the development of Python-based educational materials and their integration into future teaching and learning activities.

6. Annexes

Annex 1: Shared dataset collection document

Annex 1 contains the shared dataset collection document used as the primary data source for the analysis presented in this report. It provides detailed descriptive, legal, structural and technical information on all 120 datasets covering video, audio/sound, numerical and image modalities.

Annex 2: Classification of Datasets by Main Application


TEACHING AI

Lead Partner

UNIVERZA V LJUBLJANI
University of Ljubljana

Partners

UNIVERSITAS STUDIORUM SARAIEVOENSIS
UNIVERZITET U SARAJEVU

УНИВЕРЗИТЕТ У БЕОГРАДУ
UNIVERSITY OF BELGRADE

UCG
Univerzitet Crne Gore

UNIRI