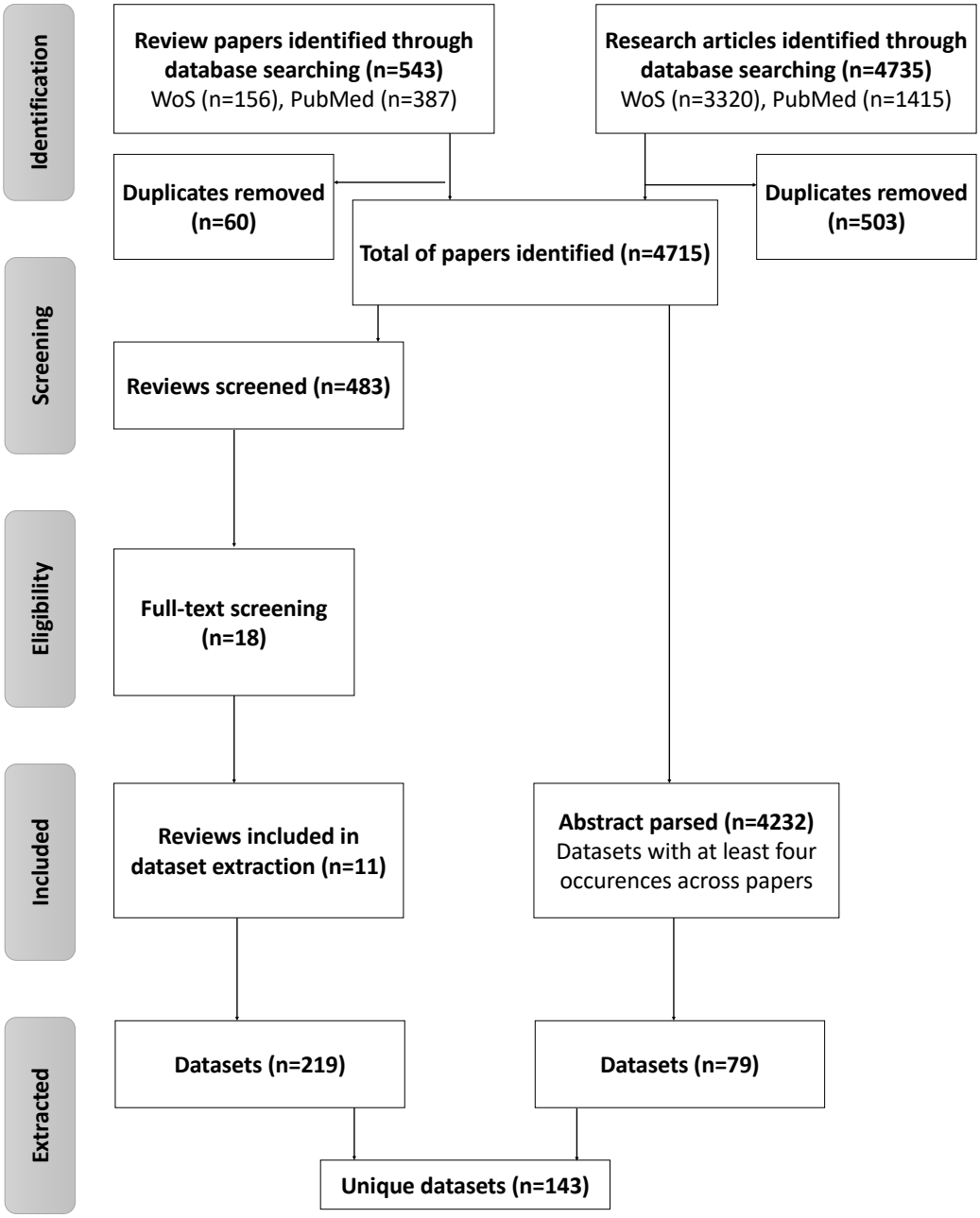
Review protocol

	Study question How biased are the existing emotion recognition datasets with respect to gender, age and ethnicity?
	Information sources for reviewed papers **Included:** Web of Science (WoS), PubMed Reasons: high compatibility of search and export options, wide coverage of scientific papers (~33000 Journals in each platform) **Excluded**: Google Scholar, Research Gate Reasons: no full export options
	Search Query (emotion OR affect) AND (recognition OR detection) AND (dataset OR database OR corpus) AND (video OR image OR audiovisual OR audio OR speech)
	Filters **Language:** English **Fields: WoS** Title, Abstract, Author Keywords, Keyword+ **PubMed**: all fields as the selection of multiple fields was not possible. PubMed fields included (Affiliation, Author, Author-Corporate, Author-First, Author-Identifier, Author-Last, Book, Conflict of Interest Statements, Date-Completion, Date-Create, Date-Entry, Date-MeSH, Date-Modification, Date-Publication, EC/RN Number, Editor, Filter, Grant Number, ISBN, Investigator, Issue, Journal, Language, Location ID, MeSH Major Topic, MeSH Subheading, MeSH Terms, Other Term, Pagination, Pharmacological Action, Publication Type, Publisher, Secondary Source ID, Subject - Personal Name, Supplementary Concept, Text Word, Title, Title/Abstract, Transliterated Title, Volume)
	Dataset selection strategy Our goal was to identify as many emotion/affect datasets as possible fullfilling the following criteria: **Inclusion criteria** **Modality**: audio, video, audio-visual **Labeling**: single or multiple emotions/affect dimensions **Exclusion criteria** **Modality**: physiology data (EEG, ECG, fMRI etc), body movements without face information **Labeling**: pain, stress and other emotion-related terms that do not belong to the basic emotion categories The results of the search on WoS and PubMed were exported in excel and csv files, respectively and combined into a single excel file. The review and research papers were saved into separate files. For the review papers, the relevance of each paper was assessed by two researchers independently and was discussed until an agreement rate of 90% was achieved (calculated automatically based on researcher annotations). The relevance was determined based on the review's likeliness to discuss datasets of the pre-defined inclusion and exclusion criteria. The full texts of the relevant reviews were then scanned for datasets.

To complement this approach, we also developed a tool for automatic parsing of the abstracts in the research articles of the search query. In particular, we took advantage of the fact that most dataset names contained at least two uppercase letters and their full phrases were explained in brackets. Hence, we developed an automated parsing tool, which extracts every sentence from an abstract with words that 1) contained at least two uppercase letters and 2) were written in brackets. Some of the common capitalized abbreviations that were not dataset names (e.g., EEG, ECG, etc.) were included in a blacklist (see next). After filtering out words with less than four occurrences (88th percentile) across the scanned papers, we inspected the sentences associated with each term (containing the uppercase letters) and extracted all dataset names.



PRISMA study flow chart for the systematic review detailing the database searches, the number of abstracts and full texts screened, and the datasets extracted.}

Blacklist

eeg, audiovisual, ser, fmri, cnn, svm, bovw, auc, hci, hog, swift, uk, usa, microexpression, ii, hrv, ecg, hmm, iot, youtube, imdb, softmax, vgg, ai, ml, dnn, dcnn, pca, lda, sift, mri, mfcc, cnns, pubmed, rgb, mlp,

	covid, methods, results, resnet, roi, knn, unet, shortterm, background, matlab, tv, iou, cca, bow, rnn, gru, fps, eegbased, densenet, svms, objective, db, ann, materials, melfrequency, lidar, mrmri, knearest, violajones, fscore, inceptionv, ssd, gan, bv, cnnbased, yolo, ct, lstm, conclusion, rgbd, alexnet, hmms, mfccs, conclusions, photooptical, adaboost, dl, lbp, dnns, ieee, dr, rcnn, rmse, roc, nlp, yolov, elm, cad, ci, ad, pd, au, modis, dbn, rois, spie, rf, ir, wm, md, asd, map, ga, fau, pet, cc, snr, ar, cd, lr, dti, gmm, ict, fp, eer, mes, me, er, si, hr, nir, surf, glcm, sar, asr, mr, aus, fa, tbi, ms, amd, dwt, mci, sd, oct, dct, us, uav, tm
	Information extraction from the datasets In the next step, we searched the descriptions of the identified datasets in the associated publications. The information source for each dataset is documented and can be viewed in our dataset repository. Two researchers and two research assistants participated in the information extraction process using clearly defined guidelines as follows: **Dataset**: str, dataset name **Dataset**: str, dataset name **Authors**: str, authors of the dataset, can be first author et al. **Link**: link where you get the information, paper (preferable) or dataset… **Year**: int, XXXX, year of the paper publication (preferable) or dataset **Notes**: str, optional **Voice**: boolean, 1 or 0, is there speech/voice modality or not **Image**: boolean, 1 or 0, is there image modality or not. Of course videos consist from images, by definition however if there is video modality image should be coded with 0 UNLESS the videos are short image sequences leading to the emotion (for example, participants are asked to gradually make an angry face) **Video**: boolean, 1 or 0, is there video modality or not **Other modalities**: str, optional (if easy to extract), delimiter is ‘,’. Example: EEG, EDA, Gestures **Samples**: int, number of videos, images, utterances, etc. **Setting**: str, lab or wild, dropdown **Setting type**: str, posing(make a sad face, raise eyebrows); acting(simulate emotions in context); spontaneous; mix (specify in notes); induced (E.g. participants were recorded when viewing/processing emotional material); dropdown **Posed**: str, what they are posing? Emotions, facial action units? Dropdown **Prof. Actors**: boolean, 1 or 0, if they are professional actors **Interaction**: boolean, 1 or 0, was it an interaction or not? **N of participants**: int, how many participants? **N-female**: int, how many female. Count, not percentage **N-male**: int, how many male. Count, not percentage **P-detailed age info:** boolean, 1 or 0. If there is standard deviation or age information for each participant, then 1, otherwise 0. **Age range**: int-int. Example: 18-76. **Age mean:** float. Mean of the age distribution

	Age median: float. Median of the age distribution. **Ethnicity**: str. Percentage also if available. For example: Caucasian(80%), East-Asian(10%) **Language**: str. If native speakers, L1-French, L2-English **Rater**: str, who is rater, [self, external, both] **Total raters**: int, number of external raters **RpI**: int, raters per item, if they rated only some subset **N-ER-female**: int, number of female raters **N-ER-male**: int, number of male raters **R-detailed age info:** boolean, 1 or 0. If there is standard deviation or age information for each rater, then 1, otherwise 0. **ER-Age range:** int-int. Example: 18-76. **ER-Age mean**: float. Mean of the age distribution **ER-Age median**: float. Median of the age distribution. **ER-ethnicity**: str **Affect labeling mode**: str, scale, method (SAM)… If it's range, use # **Emotion labels:** str, delimiter is ',' . For example: anger, happiness, calm **Emotion Label Mode**: str, e.g. FC (forced choice), Likert scale…