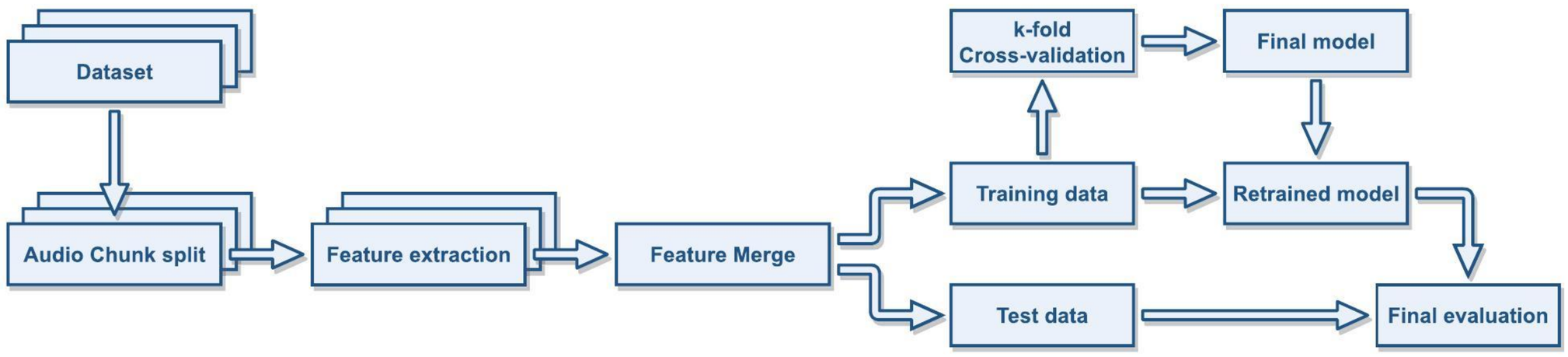

Fig. 3. Diagram with the step sequence followed during the experiments, each of these have different configuration parameters that can be set in a configuration file in order to automatize the process of producing the following results.

TABLE II
NN ARCHITECTURE

Layer	Size
Input	Depends on the feature size
Inner 1	16 fully connected (ReLU)
Inner 2	8 fully connected (ReLU)
Output	Sigmoid function

in [29], named dataset 2. This dataset presents a different sub-folders organization with respect to dataset 1 [4] that was already available. Subsequently, we have incorporated the ability to amalgamate extracted features from diverse datasets and store them in distinct folders. This facilitates the efficient reuse of data for conducting multiple experiments on audio chunks with identical settings, preventing the need to re-extract features that are already available. The complete dataflow is shown in Fig. 1, where it is visible that the features are merged only after the extraction from the relative dataset, and following operations are performed as they would be done for a single dataset.

III. RESULTS

A. Experiment Setup

All the experiments execute the complete sequences of steps described in Fig. 3. Starting from a base configuration, each group of experiments explored the effect of different settings on the classification performances of the model. The base configuration has the following settings.

1) *Audio chunk split:* Length of 5 s with a hop size of 5 s, so they are contiguous without overlapping.
2) *Feature extraction:* For each experiment two types of audio features are extracted: MFCC with 20 coefficients and STFT with a window size of 1024 samples.
3) *Training/test data split:* Training data are 80% and the remaining 20% are used for the final test.
4) *K-fold cross-validation:* It is performed dividing the training set into 10 folders.
5) *Classifiers:* The NN has two fully connected inner layers with an architecture resumed in Table II, and the SVM is trained with the C parameter set to one.

B. NN Size Influence

In the subsequent set of experiments, we explored the impact of varying the NN size by adjusting the number of layers from 1 to 4 and modifying the number of neurons per layer within the range of 2 to 8. The experiments were conducted separately using MFCC features and STFT features. The findings, presented in Figs. 4 and 5, align with the trend observed in the prior study [23], indicating that larger networks tend to achieve higher F1-score. It is important to note, however, that when utilizing combined datasets, a slightly lower accuracy is observed compared to the results obtained with individual datasets. The simplest network, with four layers and two neurons, shows a reduction of the F1-score of about 8% and 5% with MFCC and STFT, respectively. However, the gap is reduced to 1% or lower values when at least six neurons per layer are present. Colors in Figs. 4 and 5 are scaled to the same range to make them comparable. It is possible to visually observe that STFT features better performed over MFCC features in almost all cases, with exceptions for the smallest networks.

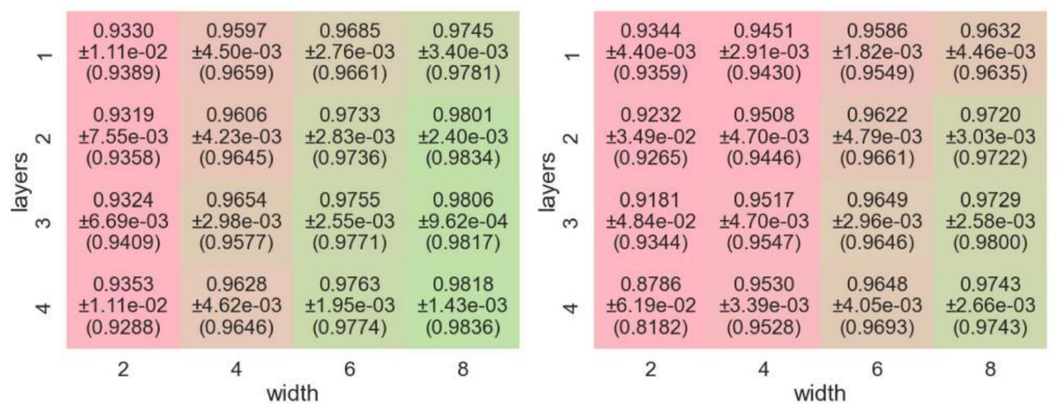

Fig. 4. Cross-validation and final test results changing the number of layers and the number of nodes in the NN using the MFCC features. On the left the F1-score using only dataset 1 and on the right combining dataset 1 and dataset 2. In these figures is reported the mean value ± the standard deviation and, between parenthesis, the final test.

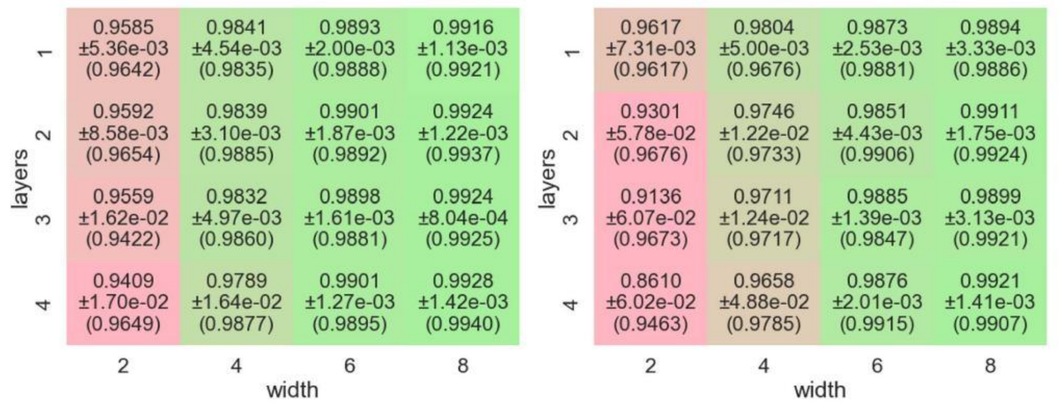

Fig. 5. Cross-validation and final test results changing the number of layers and the number of nodes in the NN using the STFT features. On the left the F1-score using only dataset 1 and on the right combining dataset 1 and dataset 2. In these figures is reported the mean value ± the standard deviation and, between parenthesis, the final test.