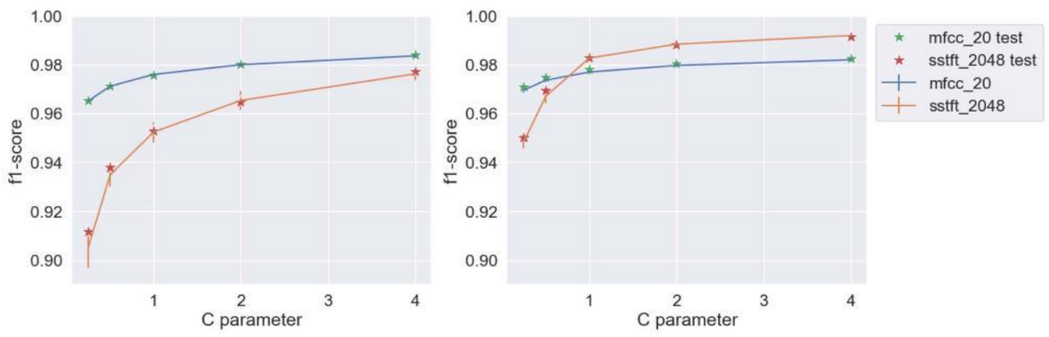

Fig. 6. Cross-validation and final test results changing the C parameter for the SVM classifier. On the left using only dataset 1 and on the right combining dataset 1 and dataset 2.

TABLE III
ACCURACY AND F1-SCORE OF SVM WITH DIFFERENT REGULARIZATION FACTORS

C	Feature	Accuracy	Accuracy test	F1-score	F1-score test
0.25	MFCC	0.9575 ± 1.41E-03	0.9598	0.9696 ± 1.00E-03	0.9710
0.5	MFCC	0.9633 ± 1.33E-03	0.9651	0.9737 ± 9.61E-04	0.9748
1	MFCC	0.9680 ± 1.42E-03	0.9698	0.9770 ± 1.05E-03	0.9781
2	MFCC	0.9718 ± 1.35E-03	0.9733	0.9798 ± 1.00E-03	0.9807
4	MFCC	0.9751 ± 1.39E-03	**0.9762**	0.9821 ± 1.05E-03	**0.9827**
0.25	STFT	0.9290 ± 3.48E-03	0.9320	0.9484 ± 2.57E-03	0.9501
0.5	stft	0.9544 ± 3.81E-03	0.9583	0.9670 ± 2.83E-03	0.9696
1	stft	0.9760 ± 2.18E-03	0.9764	0.9828 ± 1.60E-03	0.9829
2	stft	0.9840 ± 1.27E-03	0.9840	0.9885 ± 9.24E-04	0.9884
4	stft	0.9889 ± 9.01E-04	**0.9884**	0.9920 ± 6.61E-04	**0.9916**

The bold values are the best results obtained for that experiment, and for the two metrics "accuracy" and "f1-score".

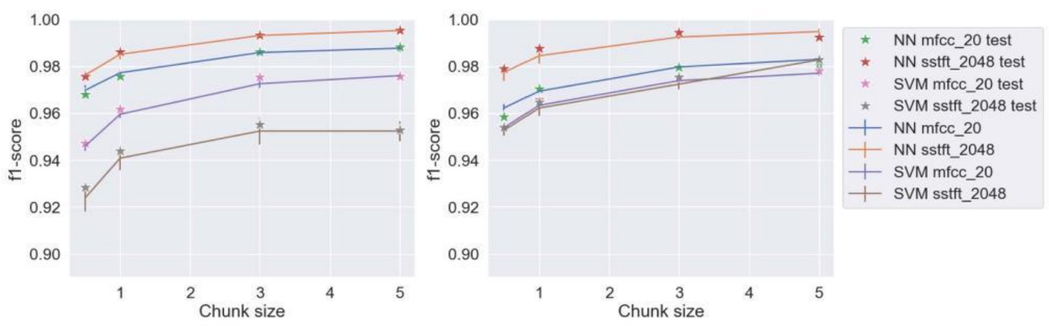

Fig. 7. Crossvalidation and final test results changing the size of the chunks but keeping the hop size constant. On the left using only dataset 1 and on the right combining dataset 1 and dataset 2.

C. SVM Regularization Parameter Influence

The second set of experiments was conducted on the SVM by changing the parameter C and applying this classifier first to the MFCC features and then also to the STFT features. The use of the two combined datasets in this case allows for higher accuracy for the STFT features with larger values of C, as highlighted in Fig. 6 and in Table III. On the other hand, for the MFCC features, the accuracy remains very similar. One reason of the higher accuracy in STFT case can be related to the SVM sensitivity and its robustness to high-dimensional data [37]. The nature of the STFT features, capturing detailed frequency information, might align well with SVM's sensitivity to the input space, resulting in improved classification performance with increased C.

D. Audio Chunks Length Influence

In these experiments, the size of the audio chunks was varied, from which both MFCC and STFT features were subsequently extracted. Then, both NN and SVM models were trained with the same inputs to generate the graphs in Fig. 7, and the numerical results are reported in Tables IV and V. The observed trend in accuracy, as previously documented in [23], regarding the influence of chunk size is reaffirmed. This trend indicates that

TABLE IV
SVM ACCURACY AND F1-SCORE WITH DIFFERENT CHUNK SIZES

Chunk size	Feature	Accuracy	Accuracy test	F1-score	F1-score test
0.5	mfcc	0.9352 ± 2.85E-03	0.9362	0.9537 ± 2.10E-03	0.9543
1	mfcc	0.9486 ± 1.76E-03	0.9525	0.9634 ± 1.18E-03	0.9659
3	mfcc	0.9633 ± 1.49E-03	0.9648	0.9739 ± 1.25E-03	0.9748
5	mfcc	0.9680 ± 1.42E-03	**0.9698**	0.9770 ± 1.05E-03	**0.9781**
0.5	stft	0.9337 ± 2.85E-03	0.9358	0.9526 ± 2.05E-03	0.9540
1	stft	0.9471 ± 4.89E-03	0.9511	0.9622 ± 3.47E-03	0.9649
3	stft	0.9614 ± 3.15E-03	0.9658	0.9724 ± 2.28E-03	0.9755
5	stft	0.9760 ± 2.18E-03	**0.9764**	0.9828 ± 1.60E-03	**0.9829**

The bold values are the best results obtained for that experiment, and for the two metrics "accuracy" and "f1-score".

TABLE V
NN ACCURACY AND F1-SCORE WITH DIFFERENT CHUNK SIZES

Chunk size	Feature	Accuracy	Accuracy test	F1-score	F1-score test
0.5	mfcc	0.9462 ± 2.55E-03	0.9419	0.9622 ± 1.78E-03	0.9586
1	mfcc	0.9566 ± 4.39E-03	0.9583	0.9694 ± 3.09E-03	0.9704
3	mfcc	0.9713 ± 2.20E-03	0.9714	0.9797 ± 1.59E-03	0.9797
5	mfcc	0.9761 ± 2.16E-03	**0.9732**	0.9830 ± 1.51E-03	**0.9806**
0.5	stft	0.9679 ± 5.09E-03	0.9707	0.9774 ± 3.55E-03	0.9791
1	stft	0.9781 ± 5.03E-03	0.9828	0.9845 ± 3.38E-03	0.9878
3	stft	0.9894 ± 1.48E-03	**0.9921**	0.9925 ± 1.01E-03	**0.9944**
5	stft	0.9927 ± 1.69E-03	0.9897	0.9948 ± 1.22E-03	0.9926

The bold values are the best results obtained for that experiment, and for the two metrics "accuracy" and "f1-score".

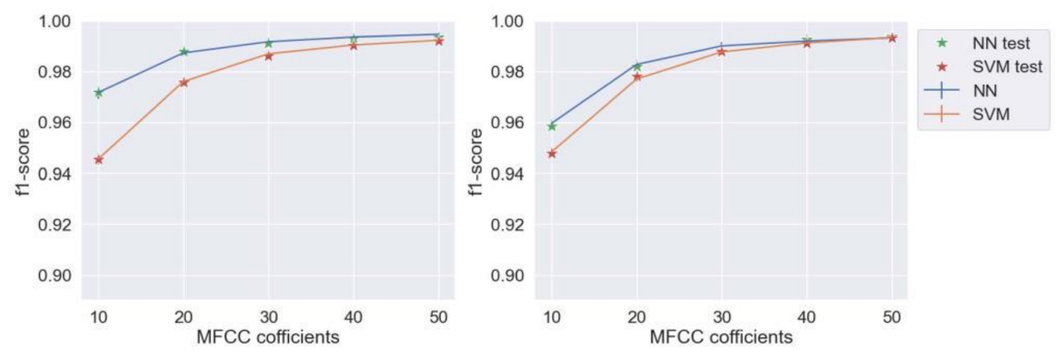

Fig. 8. Cross-validation and final test results changing the number of extracted MFCC. On the left using only dataset 1 and on the right combining dataset 1 and dataset 2.

longer audio segments contribute to improved results. However, it is crucial to acknowledge that longer audio chunks also necessitate higher computational resources, a factor that may pose challenges in low-power IoT systems where such resources are not always readily available. It is noteworthy that the results from the experiment involving 1-s chunks (AUC $=$ 0.9876 ± 8.08E-04) surpass significantly the outcomes achieved in [29], where MFCC_20 features and an SVM classifier were employed, resulting in an AUC of approximately 0.91 using only the dataset 2 [24].

E. Feature Resolution Influence

In the last group of experiments, the audio feature parameters were modified affecting the size of the features extracted from the audio files. This determines the number of inputs required in the classifier model. The results obtained and represented in Figs. 8 and 9 confirm the trend demonstrated in the previous work [23], where features with a greater number of coefficients benefit the classifier because they contain more details necessary for achieving higher accuracy. This is visible also in Tables VI and VII where the best accuracy is obtained with the higher resolution features. The downside is that higher number of features require more complex classifiers, for example the number of input nodes of the NN increases. This will require more memory to store the classifier model but also increases the complexity of