The National Contests Behind International Success: A Musical Comparison of the Eurovision Song Contest, the Festival di Sanremo and the Melodifestivalen

MSc Thesis *(Afstudeerscriptie)*

written by

Jasmijn van Harskamp
(born 10 November 1999 in Alphen aan den Rijn)

under the supervision of **dr. John Ashley Burgoyne**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
30 June 2022	Dr. Benno van der Berg *(Chair)*
	Dr. John Ashley Burgoyne *(Supervisor)*
	Prof. dr. Henkjan Honing
	Dr. David Baker

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

The Eurovision Song Contest is the longest-running annual international music competition. In recent years, especially Italy and Sweden have obtained many successes at the song festival. Both countries select their entries through a popular national contest, respectively the Festival di Sanremo and the Melodifestivalen. Though all three contests have been studied repeatedly in various research areas, very little is known about their music. This thesis analyses the music represented at the Eurovision and at the two national contests in the period between 2011 and 2021. The aim of this study is to investigate the differences and similarities between these contests with respect to the performed entries. In order to analyse this, different sets of XGBoost classifiers were trained on musical features to match songs to the contest they were entered in. The results show a greater similarity between Eurovision and Melodifestivalen songs, than between Eurovision and Sanremo songs. We argue that this is most likely caused by distinct national music styles. Additionally, we attempt to predict the outcome of the Eurovision final from the outcome of the national contests by using the musical features. This problem was approached by training a top ten classifier and by constructing a ranking model. Neither method shows promising results. However, a comparison of the actual voting behaviour of all countries participating in the Eurovision and the predictions based on the national competitions supports our earlier hypothesis about national styles.

Acknowledgements

First of all, I would like to express my sincerest gratitude to Ashley Burgoyne. Without your guidance, this thesis would not have been as it is now. Thank you for all your knowledgeable advice, support and infectious enthusiasm throughout the process of conducting this study.

Secondly, I would like to thank all members of the defense committee: Benno van den Berg, Henkjan Honing and David Baker. I appreciate you taking the time to read and assess my thesis and contribute to my defense. Moreover, a special thank you goes to professor Yde Venema for stepping in at very short notice and chairing my defense.

My time in the MoL has definitely been unusual, having taken almost all my courses online. Nonetheless, I am grateful for all the intelligent, kind and interesting people I got to work with and learn from. In particular, I would like to thank Mike. I am so happy that the Zoom-algorithm decided to put us together in a break-out room during our first seminar. Not only did you turn out to be a brilliant homework partner, but more importantly an amazing friend.

To my parents, sister and brother, thank you for always supporting me no matter what, for encouraging me to be my best version every day and most of all for the endless joy and love you give me. To Nina and Larissa, thank you for being there for me whenever I need it and for inspiring me with your (quite literally) borderless ambitions. And finally, to Sybrand. You make my world so much brighter. Thank you for your calming and loving support, for challenging and motivating me, and of course for watching the Eurovision Song Contest with me, “for the sake of science”.

Contents

Introduction

The *Eurovision Song Contest* is an annual songwriting contest held by the European Broadcasting Union. The first edition was held in 1956, making it the world's longest-running annual international TV music competition. Moreover, with around 180 million viewers each year, it is one of the most successful television shows worldwide [20]. Every member country of the European Broadcasting Union is allowed to take part in the Eurovision Song Contest, as well as specially invited associates. In the last few years this resulted in around forty participating countries. The contest consists of two semi-finals and a grand final. In each of the semi-finals, the ten best scoring entries win a place for the final. The host country and the countries in the 'Big 5' — France, Germany, Italy, Spain and the United Kingdom — are automatically placed for the final. Usually a total of twenty-six or twenty-five entries perform in the grand final.

Every participating country gets to send exactly one entry to the festival. The manner in which the competitor is chosen, however, varies from country to country. The official Eurovision website mentions three common ways of selecting the competitor [21]. Either the participant is chosen through a televised national competition in which the public is given the opportunity to take part in selecting the entry; or the artist and song are chosen through a full internal selection process; or through some mixed format. In the latter case, an artist is often chosen by an internal committee, while the public can take part in selecting the song during a televised show.

The voting system of the Eurovision Song Contest is a positional voting system. Under the current voting rules introduced in 2016, the performances are judged per country by a jury consisting of five music industry professionals and by the general public, via telephone, SMS or the official app. Both of these groups allocate a set of 1 to 8, 10 and 12 points to their top ten acts. That is, their favourite performance gets 12 points, their second favourite 10 points, and so on, the tenth favourite act receiving 1 point. An artist can thus receive at most 24 points from one country. Between 2011 and 2015, the rankings from the public and the jury were combined to determine the allocation of one set of 1 to 8, 10 and 12 points. For both systems, the following additional rules hold. First of all, it is not possible to vote for your own country. Secondly, each country is only allowed to vote in the semi-final in which their entry performs. Additionally, during each semi-final, a subset of the pre-qualified countries are permitted to vote. In the grand final, all participating countries can vote again, regardless of whether their entry has made it to the final.

Although the Eurovision Song Contest still has a reputation for being camp and hosting music that is not always judged as 'high quality', its public is growing annually. Especially young audiences have embraced the festival in recent years, with almost 53% of TV viewing 15- to 24-year-olds watching the 2021 Eurovision edition [23]. Also, the contest is becoming more global. Since 2015, Australia is allowed to participate in the Eurovision Song Contest, after having broadcast the festival already since 1983. Furthermore, an American version of the contest was held for the first time this year (2022), after an increased popularity of the European contest in the US.

In addition to the growing public, the countries and artists seem to take the contest more seriously and approach their participation as an opportunity to boost their international career. This is no wonder, after the recent successes of winning participants, and even of non-winning candidates. For example, the winning song of Eurovision 2019, *Arcade* by Duncan Laurence, was streamed over 2.83 million times on the day after the final and entered Spotify's Global Top 50 on the fourteenth place [46]. Even more popular were the 2021 winners, the Italian band Måneskin, who during the course of 2021 found four of their songs in Spotify's Global Top 50 list.

This shows that selecting the right artist and song might lead to great successes for the individual artists, as well as creating lucrative opportunities for the corresponding countries and broadcasters. Therefore, in this thesis, we will focus on two countries that have proven to be very successful in the recent editions of the Eurovision Song Contest, namely Italy and Sweden. Since Italy reentered the Eurovision in 2011, they occupy the second position in the overall ranking regarding the obtained amount of points in the contests between 2011 and 2021 (with the exception of 2020, when the international song contest was cancelled due to the COVID-19 pandemic) [26]. They won the contest in 2021 and finished in the top three in 2011, 2015 and 2019. Sweden, on the other hand, occupies the first position in this total ranking since 2011. They secured the win in 2012 and 2015, and finished third in 2011 and 2014. Therefore, it is safe to say that for both countries their selection method seems to be effective.

Each year in February, the Italian national *Festival della canzone italiana di Sanremo* is held, more commonly referred to as the Festival di Sanremo or the Sanremo Festival. This contest was first held in 1951 and formed the inspiration for the international Eurovision Song Contest. With over ten million viewers each year, it is very popular among the Italian public. Since 2015, the winning participant of the multi-day festival gets to represent Italy at the Eurovision Song Contest, if they agree to do so. From 2011 to 2015, an internal committee chose the Eurovision entry, often from the participants in the Sanremo Festival.

The exact format of the contest and the method used for voting vary slightly from year to year. Though, during all editions, voting occurs through public televoting. In addition, it is complemented by the votes from an expert jury, a demoscopic jury of fans, a press jury, a jury consisting of the singers and musicians of the orchestra, or a combination of these. For example, in 2021 the ranking in the final round was determined by public televoting for 34%, the press jury for 33% and the demoscopic jury for the remaining 33%; while in the 2019 edition public televoting counted for 50%, press jury voting for 30% and an expert jury for 20%.

In Sweden, the outcome of the annual music festival *Melodifestivalen* determines the Swedish entry for the Eurovision Song Contest. The festival was first held in 1959 and attracts around three million viewers each year.

The final result of the Swedish competition is determined by public televoting and points assigned by juries —until 2018 there were eleven juries, since 2019 there are eight. Each jury represents some country that participates in the Eurovision Song Contest. Until 2017, the juries awarded 1, 2, 4, 6, 8, 10 and 12 points to their top seven. The points obtained from public voting corresponded to a share of 473 points (the total amount of jury points) based on the percentage in the televote. In 2018, the juries awarded their top ten with 1 to 8, 10 and 12 points. Consequently, the total number of televoting points was increased to 638 that year. From 2019 onward, when the number of juries was reduced to eight, the televoting public was also divided into eight groups; seven groups based on age for the app voters and one group for telephone voters. Like the juries, the voting in each group determines the allocation of 1 to 8, 10 and 12 points.

With the apparent success of Italy and Sweden at the Eurovision Song Contest, several questions which we will address in this thesis arise.

First of all, we wonder how similar these international and national contests are with respect to the musical entries. We note that the Eurovision, the Sanremo and the Melodifestivalen all share similar characteristics, such as their large audiences, their voting systems and even their objective to showcase the best of what Europe, Italy or Sweden has to offer in terms of music. However, it is exactly the music that we know very little about. Are the same kinds of songs represented at both the national and the international festivals? Or is there a distinct dissimilarity that differentiates these contests? With six wins in total and the most popular winners ever in the form of ABBA, Sweden is one of the most successful countries in the history of Eurovision. They send potential winners almost every year and moreover many other countries send songs inspired by —if not completely produced by— Swedish musicians. We therefore expect the music at the Eurovision and the Melodifestivalen to be quite similar. Italy on the other hand, is known as slightly more unorthodox. For example, they are one of the few countries who always sing in their own language. While the recent editions have shown that this can still lead to success in the international contest, we do expect certain differences between the Eurovision music and the Sanremo music. These questions posed above and their results and implications will be addressed in chapter 3.

Secondly, we might wonder whether the results from the Sanremo and Melodifestivalen can be used to make predictions about the Eurovision Song Contest —after all, it seems like the Italian and Swedish juries and public have an eye for selecting successful entries. That is, can we use musical features to predict what other songs will be popular in the Eurovision, based on the popular songs from the national competitions. It is of course notoriously hard to predict the outcome of the Eurovision Song Contest, as is predicting the outcome of any competition. Moreover, it is clear that the entries are not only judged based on their music, but that also factors such as the visual stage performance and politics play a part in determining the final ranking. We therefore do not expect the predictions to be very successful. A more fruitful approach could be to compare the predicted outcome based on one of the national competitions to the actual voting behaviour of countries participating in the Eurovision. There we would expect the Italians to vote for the most part like the prediction based on the Sanremo, and similarly we expect the Swedes to vote in accordance with the prediction based on the Melodifestivalen. Furthermore, such an analysis might improve our understanding of national preferences and consequently their voting behaviour. All endeavours corresponding to predictions will be further explored in chapter 4.

The data and implementation of all models presented in this thesis can be found at `https://github.com/Jasmijn-vH/Thesis-MoL`.

Related work

With its long history and availability of detailed data, the Eurovision Song Contest has been of interest to researchers from various scientific areas. We give an overview of some of the research topics.

First of all, the contest has been used to study political and cultural developments in Europe [49]. With its large international audience, the Eurovision Song Contest is an appealing medium for nation branding [10, 40]. An artist performing at Eurovision is not just an individual, but a representation of their entire country. Therefore, the responsible national broadcaster determines the way in which their country will be perceived by millions of European viewers, through their choice of the artist, song and accompanying act. Especially Eastern European countries have used the contest to ameliorate their international image [32]. In [35], Kyriakidou et al. argue that the contest is an arena for 'playful nationalism', where fans wave their flags and paint their faces in the national colours, but that the contest allows for a celebration of Europe's cultural diversity at the same time. With its message of diversity and inclusiveness, the contest has also become a popular event among LGBTQ+ individuals and groups [29]. Most notably, the win of the Austrian singer Conchita Wurst in 2014 intensified the debates surrounding LGBTQ+ rights ahead of the European Parliament elections [17, 4].

A second field of research concerns the voting. As with any type of juried contest, the fairness of the Eurovision can and has been questioned. Most notably, its voting system has often been accused of being biased by strategic voting. This allegation was supported in [18], but it was claimed that the bias was only partially caused by geographical reasons. In [47], Spierdijk and Vellekoop also confirmed a bias in the voting structure, based on various factors such as geography, culture, language, religion and ethnics. However, they did not find strong evidence for the publicly debated alliances. In [8], Budzinski and Pannicke analysed the German national contest, the Bundesvision Song Contest, to show that voting biases not only occur in the international contest, but also in a national contest of a similar structure. Apart from deliberate strategic voting, the Eurovision results might also be influenced by other factors. For example, it has been shown that participants that appear later in the show score better on average [15]. In [28], Haan et al. confirmed this ordering bias, adding that also the very first act obtains significantly better results. Moreover, they showed that the outcome of contests judged by expert judges is less affected by the ordering of participants than contests judged by the public.

In this thesis, we will focus on the musical features of the songs represented in the Eurovision Song Contest. As far as we know, there is little quantitative research on the contest with respect to its musical content. There has been some work in which Eurovision songs were used to test methodologies, e.g. for the semiotic description of music structures in [5]. However these studies are not primarily interested in the song festival as a contest and are therefore not very useful for our research.

In addition, also the Sanremo Festival and the Melodifestivalen have been studied with respect to various research areas. We briefly discuss some of the studied topics. Note that most of the analyses regarding these national contests, have been conducted for theses of local universities. We are not aware of any quantitative research focusing on the musical features represented in either of the national contests.

First of all, the history of the Sanremo Festival has been analysed and linked to important historical and musical developments in contemporary Italy in [25]. Moreover, the festival was the subject of a study focusing on its cultural aspects in [2]. It was argued there that the Italian contest is a cultural product as well as a product of popular culture —meaning that it still meets the taste of the majority of the population. Finally, events and acts at the Sanremo have been used in case studies on various occasions; for example in [14], the participation of the alternative rock band Afterhours in the Italian music contest was discussed with respect to authenticity marketing.

For the Swedish Melodifestivalen, several cultural-political studies have been conducted. For example, in [31] an analysis of femininity in the Melodifestivalen from the 2000s concluded that women in the contest usually could be categorised as one of four groups —i.e., 'the mother and/or wife', 'the friend', 'the diva' and 'the female subject'. The jury voting and its relation to the order of performances in the Melodifestivalen have been studied in [37]. However, this research did not result in any convincing conclusions. Finally, the contest has been researched in relation to its social media impact. For example, in [33] it was attempted to predict the outcome of the festival based on the sentiment in tweets. Their most successful approach was to predict the top five from the sum of positive sentiment. In [19] the festival's place branding on Instagram was analysed. This study showed that the social media coverage of an event like the Melodifestivalen positively influences the image of the hosting city.

In this thesis, we will compare the international Eurovision Song Contest and national selection competitions. We are not aware of any research that compares the Eurovision, the Sanremo Festival and the Melodifestivalen. Moreover, as far as we know, there has been no scientific comparison of the Eurovision and any national contest with respect to its music.

Background

In this chapter we provide the theoretical background necessary for the conducted research. We introduce the research field of Music Information Retrieval and discuss its use of machine learning techniques. In particular, we focus on a specific classification and ranking algorithm, which will be the main algorithms used in this thesis. Moreover, we elaborate on the evaluation metrics used for our models.

2.1 Music Information Retrieval

Music Information Retrieval (MIR) is the interdisciplinary field that deals with extracting data from music. It combines topics from musicology, psychology, psychoacoustics, statistics, artificial intelligence, machine learning and other related disciplines. Apart from being an active research field, MIR has many applications outside of academia as well. For example, its techniques are used for music recommendation systems [38] or query-by-humming [43].

2.1.1 Musical Features

To retrieve relevant information, multiple representations of music can be used, such as sheet music or digital audio. Subsequently, one could work with these files directly, or with some representation of it. A common approach to represent the audio —and the one we will be using in this thesis— is to work with musical features. Musical features describe the characteristics of a song pertaining to various aspects such as the audio spectrum, melody, rhythm, etcetera. For this thesis we obtain our musical features by using a feature extractor on the (digital) audio fragment. There are many different feature extractors, all with their own advantages [41]. We will use the Python environment of the open-source C++ library Essentia [6]. Essentia is a toolkit collecting a reference standard of MIR features. Most notably, these features are used in the AcousticBrainz project [1]. Moreover, this extractor has a high computational efficiency. Due to these characteristics, Essentia has been used both in industrial applications, such as music education apps, and in academic research, e.g. for music classification, musical instrument detection and music recommendation[1].

The Essentia library provides access to many algorithms. For our purposes, we mainly use the MusicExtractor wrapper. In order to guarantee consistency between the songs, this algorithm resamples all signals to a 44kHz sample rate, summes it to mono and normalises it using the replay gain values. Subsequently, it extracts a number of features, which are divided into three categories.

[1] For a more elaborate overview of Essentia's practical and academic applications, see `https://essentia.upf.edu/applications.html` and `https://essentia.upf.edu/research_papers.html`.

Lowlevel. The lowlevel features are generally considered to have no direct human interpretation. They mainly consist of spectral characteristics, such as the zero-crossing rate and descriptors of the spectral shape. Additionally, this class contains features corresponding to three physical models of the human ear, namely the Bark scale, the mel scale and the ERB-rate scale. The mel scale is the oldest and most commonly used scale, while the ERB scale is a more recent development that tends to outperform the other scales [42]. For this study, we chose to only consider the ERB scale. Related to these scales are the mel frequency cepstral coefficients (MFCC) and gammatone frequency cepstral coefficients (GFCC). Both vectors are used in MIR to capture timbre. Following our previous choice, we omitted the MFCC score corresponding to the mel scale and worked with the GFCC corresponding to the ERB scale instead. Also here, MFCC is the older, more frequently used measure, while GFCC is a newer and often more accurate measure [50, 45].

Rhythm. These features describe the rhythm of the audio fragment. It includes the number of detected beats, properties corresponding to the beats per minute and an estimate of the danceability.

Tonal. The tonal features describe properties of the audio fragment such as tuning frequency, the key and the scale.

2.1.2 Machine Learning

The academic field of MIR is relatively young and its developments are strongly related to the increase of computational power and the advancing computational methods [44]. An important part of machine learning related research in MIR deals with classification. Over the years, many forms of classification have been developed. There are logic based techniques, such as decision trees and inductive programming; perceptron-based techniques, which encompass the neural networks; and statistical learning algorithms, including Bayesian networks and instance-based learning approaches [34].

In this thesis we will focus on the XGBoost framework [11]. This open-source library is currently one of the most successful machine learning frameworks. Therefore, it has recently been used in a lot of applied machine learning research. This includes the field of MIR, e.g. to classify different genres [27] or to automate mood recognition in classical music [36].

XGBoost, which stands for eXtreme Gradient Boosting, is an implementation of gradient tree boosting. This technique works with decision tree ensembles. Decision tree learning is probably one of the best interpretable machine learning techniques and works by sequentially splitting the data based on some variable. Repeating this procedure, a tree is built which can then be used to make predictions about new data points. Boosting is an ensemble technique based on the belief that multiple weak learners can be combined into one strong learner. It starts by applying a base learner to all data points with equal weight. Subsequently, it increases the weight of the incorrectly classified items, while decreasing the weight of the successfully classified ones. This process is repeated until no further improvements can be made and finally the models are combined into one large model. An important feature of XGBoost is that it then regularises its models. That is, to avoid overfitting, the complexity of the models is bound.

In addition to the XGBoost classifier, we will also use its ranker. This implementation is based on the LambdaMART ranking algorithm [9]. This algorithm works by performing a pairwise classification of the data points in order to construct a total ranking.

2.2 Evaluation Metrics

In recent years, machine learning has become an important part of research not just in MIR, but in many scientific areas. However, when it comes to evaluating machine learning models, a clear consensus on which metrics to use has unfortunately not yet been reached. In this section we discuss the metrics we will use and elaborate on their workings.

Most of the metrics we discuss are based on confusion matrices. These diagrams summarise the predictions of a classifier as compared to the true values. In a binary classification task, a confusion matrix shows four values: the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). An advantage of the use of confusion matrices is that in addition to showing that errors have been made, it also provides insight into what kind of errors were the most prevalent.

The most commonly used evaluation metric is accuracy. Given a binary confusion matrix, the accuracy of the classifier can be calculated using the formula

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}.$$

Hence, it is the number of correct classifications, divided by the total number of classified objects. Its value ranges between 0 —only incorrect predictions— and 1 —only correct predictions. Obvious advantages of the accuracy metric are its easy calculation and intuitive interpretation. Moreover, the accuracy measure is independent of class labels, which is to say that it treats the positive and negative class with equal importance. However, the use of accuracy as the only evaluation metric might give a distorted evaluation of the classifier, especially when it was trained on imbalanced data.

A second, fairly common evaluation metric is the F-score. To calculate it, we first define the precision and recall of a classifier. The precision indicates the number of correct positives among all items classified as positive and is calculated as

$$\text{precision} = \frac{TP}{TP + FP}.$$

The recall indicates the ratio of retrieved positive items against all actual positives and is calculated as

$$\text{recall} = \frac{TP}{TP + FN}.$$

The F-score is then calculated from the precision and recall in the following way,

$$\text{F-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + 0.5(FP + FN)}.$$

It can be seen that the F-score is not symmetric, which means that it is dependent on the class labelling. In cases where there is no natural intuition for what should be the positive or negative class, this causes a problem. Therefore, a weighted version of the F-score has been introduced. Suppose we have two true classes A and B consisting of n_A and n_B elements respectively. We then calculate the weighted F-score as

$$\text{Weighted F-score} = \frac{\text{F-score}_A \cdot n_A + \text{F-score}_B \cdot n_B}{n_A + n_B},$$

where F-score$_A$ and F-score$_B$ are obtained by setting A and B as the positive class respectively. In our study, we will use this weighted version of the F-score. From now on, whenever we mention the F-score, we intend the weighted version.

While accuracy and F-score are well-known and often-used metrics, they have repeatedly been criticised. Instead, Chicco and Jurman [12] argue to favour Matthews Correlation Coefficient (MCC). This class-symmetric evaluation technique was introduced by Matthews [39] in 1975 and indicates the correlation between the actual classification and the predicted one. Its value ranges from -1 (total disagreement), through 0 (only agreement due to chance), to 1 (perfect agreement). It can be calculated from the confusion matrix as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}.$$

In statistics, this measure is also known as the phi coefficient.

A similar measure is Cohen's Kappa, calculated as

$$\kappa = \frac{2\cdot(TP \cdot TN - FP \cdot FN)}{(TP+FP)\cdot(TN+FP)+(TP+FN)\cdot(TN+FN)}.$$

It compares the observed accuracy of the classifier with the expected accuracy. The value of Cohen's Kappa is always closer to zero than the value of MCC [13]. Therefore, it is argued that MCC is a more informative metric than Cohen's Kappa, when considering binary classification [13, 16]. However, since Cohen's Kappa is more widely known than MCC, we include it here for completeness.

As a final metric for the classifiers, we consider the area under the receiver operating characteristic curve (ROC AUC). The ROC curve is a plot showing the true positive rate against the false positive rate. The area under this curve then indicates how successful the classifier is at distinguishing the two classes. The values range between 0 (perfect misclassification), through 0.5 (no ability to distinguish the classes), to 1 (perfect classification). This measure has been used for a long time and is generally regarded as a good representative of the classifier's performance [7]. A clear disadvantage is the asymmetry of this metric.

All metrics we have discussed so far evaluate the performance of a classifier. For the ranking algorithm, these metrics will not be useful. Instead, we compute Spearman's Rank Correlation Coefficient to assess the quality of a predicted ranking. This metric indicates to what extent a monotonic function is able to describe the relation between two rankings. A Spearman correlation of 1 indicates perfect agreement between the ranks, while a coefficient of -1 points to perfectly inverted rankings.

Classifying the Eurovision, Sanremo and Melodifestivalen

In this chapter, we analyse the songs of the Eurovision Song Contest, the Festival di Sanremo and the Melodifestivalen with respect to the represented music. We aim to describe the differences and similarities between the considered national contests and the international festival. We proceed by presenting our research approach and subsequently the corresponding results. Finally, we discuss our findings and give possible interpretations.

3.1 Data and Analysis

In this section we describe the data that was used for this study, how our audio and audio features were obtained and in what way these features were analysed.

3.1.1 Audio Data

For this thesis we used the songs that participated in the Eurovision Song Contest, the Festival di Sanremo and the Melodifestivalen between 2011 and 2021. We omitted the year 2020, when the Eurovision Song Contest was cancelled due to COVID-19 restrictions. In total, we thus researched ten editions of the festivals.

Firstly, we collected all participating artists, songs and their ranking per year with YouTube links to the corresponding videos. For the Eurovision, most of this contestant data was already available from [48]. For the songs in the Sanremo Festival and the Melodifestivalen, we created a similar database using the official results published by the respective national broadcasters, RAI and SVT. All resulting files can be found in the GitHub repository corresponding to this thesis.

This study comprises 920 songs over ten years; 408 from the Eurovision Song Contest, 216 from the Sanremo Festival and 296 from the Melodifestivalen. A complete overview of the number of songs per year per festival can be found in Table 3.1. The audio of these songs was obtained from YouTube videos, using a Python-based audioscraper, partly taken from [48]. Given that some songs did not have a live performance available, we chose to use studio recorded versions of all songs. Following common practice, we consider a 29 second fragment of each song for further analysis. These fragments were manually selected to capture the essence of the song and, in most cases, correspond to (a part of) the chorus. By selecting the fragments in such a way, we ensure that the extracted features are meaningful. That is, if we were to compute the features over a full song, the dynamics in the structure of the song could

	ESC	SR	MF
2011	43	14	32
2012	42	14	32
2013	39	28	32
2014	37	28	32
2015	40	20	28
2016	42	20	28
2017	42	22	28
2018	43	20	28
2019	41	24	28
2021	39	26	28
	408	216	296

Table 3.1: Number of participants per year per contest.
ESC stands for the Eurovision Song Contest, SR for the Sanremo Festival and MF for the Melodifestivalen.

cause the features to average out and become futile.

We retrieved the audio features of all song fragments by using the MusicExtractor algorithm from the Essentia library [6]. All available statistics were computed —i.e., mean, variance, median, minimum, maximum, and the mean and variance of the first and second derivative. This resulted in a large number of features, some of which described very similar properties. Therefore, we trimmed our list of features down. As mentioned before, we chose to use the ERB scale over the comparable mel and Bark scale; and consequently the GFCC over the MFCC. The classification algorithm we chose to use, can only deal with numerical values. To this extent, we encoded various features describing the key and scale of a song using One-Hot Encoding.

After these modifications, we extracted all numerical features. This resulted in 467 musical features which were used for the classification. For a list of these features, see Appendix A.1.

3.1.2 Classification and Evaluation

Subsequently, we trained several groups of classifiers; one group to distinguish Eurovision songs from Sanremo songs, which we will refer to as the ESC-SR classifiers, and another group to distinguish Eurovision songs from Melodifestivalen songs, the ESC-MF classifiers. Songs that appeared in more than one contest were not considered during the classification. That is, in most cases, the winning songs from the national contests were also performed at the Eurovision. These songs were omitted from the classification task, in order to prevent noise caused by having to classify the same song in more than one class. All classifiers were trained using the aforementioned gradient boosting algorithm XGBoost. Both groups of classifiers consist of the following models.

General. The general classifier was trained on all songs of the corresponding contests from all years. Moreover, it had access to the full set of features.

Yearly. The yearly classifiers were trained only on the songs of the corresponding contests of one particular year. In training, all features were available.

Selection. The selection classifiers were also trained on songs of one particular year, however now the classifier only had access to a limited set of features. This set was constructed by selecting the features that were used at least once in training the corresponding general classifier.

Apart from these one-to-one classifiers that compare the Eurovision Song Contest to one of the national contests, we also trained a classifier comparing songs from all three contests. Here, we only trained the general model. We will refer to this classifier as the ESC-SR-MF classifier.

From the resulting models we then extracted the most important features, using a built-in function from XGBoost. Moreover, we cross-validated our models and evaluated their performance in the manners described hereafter. The evaluation metrics that were used, are accuracy, Matthew's Correlation Coefficient, Cohen's Kappa, F-score and ROC AUC.

For the overall cross-validation of the general classifier, we used the method 'Leave One Group Out', as the data is clearly structured into years. That is, we divided our data into groups based on the year the songs were entered in the competitions. Then, for each of these years i, we retrained the model on all other years $j \neq i$ and tested it on year i. The predictions that this model then made, were saved and used for constructing the confusion matrix and calculating the evaluation metrics mentioned before.

Additionally, we cross-validated the general classifier within the separate years. Here, we used a KFold cross-validation consisting of ten randomly shuffled folds on the songs from the corresponding year. To reduce noise, we repeated the cross-validation five times. Then, we assigned each song to the class it was classified in for the majority of the cross-validations. When considering the class-probabilities, we averaged the outcome of the five runs. Moreover, for each repetition of the cross-validation, we reported the average accuracy and its standard deviation over the ten folds. This same approach of a repeated KFold procedure was used for evaluating the yearly and selection classifiers.

3.2 Results

We now present the results of our classification models. We first show the results from the Eurovision-Sanremo classification, followed by the results from the Eurovision-Melodifestivalen classification. Finally, we present the outcome of the three-class classification.

3.2.1 Results of the Eurovision-Sanremo Classification

In this section we present the results of our classification of Eurovision and Sanremo songs. First, we focus on the important features of our models. Then, we evaluate the classifier and present its outcome.

Feature Importance

The feature importance for every classifier was displayed in an importance plot. These plots can be found in Appendix A.2.1. In Figure 3.1, we include the fifteen most important features of the general classifier. Here, the F-score indicates the number of times a feature was split on in the decision tree. The higher this number is, the more discriminative power the corresponding feature possesses and the more informative it is when trying to classify items.

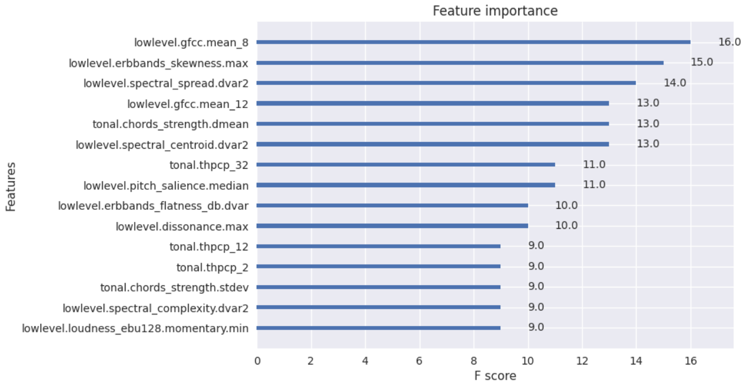


Figure 3.1: Top 15 feature importance plot for the general ESC-SR classifier.

In training the general classifier and building its decision tree, 275 features were split on at least once. These features were later used to train the selection classifiers with the yearly data.

Classifier Evaluation

As described in subsection 3.1.2, we evaluated our classifiers using cross-validation. Figure 3.2 shows the resulting confusion matrix from the Leave-One-Group-Out validation of the general classifier. We can easily calculate that 87% of the Eurovision songs was classified correctly, while for the Sanremo songs this was only 48%.

Figure 3.3 shows the confusion matrices from evaluating the general classifier per year using the repeated KFold validation. Recall that we performed five repetitions of a ten-fold cross-validation. The average accuracy and standard deviation per KFold repetition can be found in Table 3.2. This table shows that while the results vary between the repetitions, all are within one standard deviation.

In addition to the confusion matrices for the general classifier, we also constructed the confusion matrices for the repeated KFold validation of the yearly classifiers and the selection classifiers. These can be found in Appendix A.2.2.

Finally, using the information from the matrices, we evaluated the performance of all classifiers using the previously mentioned evaluation metrics. These results are presented in Table 3.3. The *‘Total’* row represents the results from the Leave-One-Group-Out cross-validation, while the other rows represent the results from the repeated KFold cross-validation. For each year and metric, we emphasised the highest scoring classifier.

Overall, our general classifier has an accuracy of 0.74 and an MCC of 0.38. The highest accuracy, MCC, Cohen’s Kappa score and F-score are obtained for the year 2011. The highest ROC AUC is achieved by the selection classifier for 2017. The lowest scores correspond to the year 2021, except for the ROC AUC, where the lowest score is found in 2018.

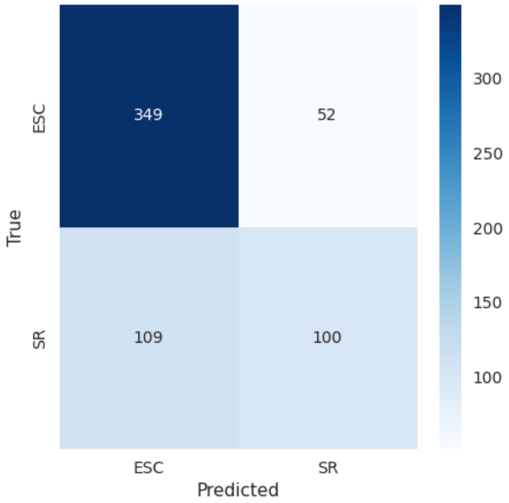
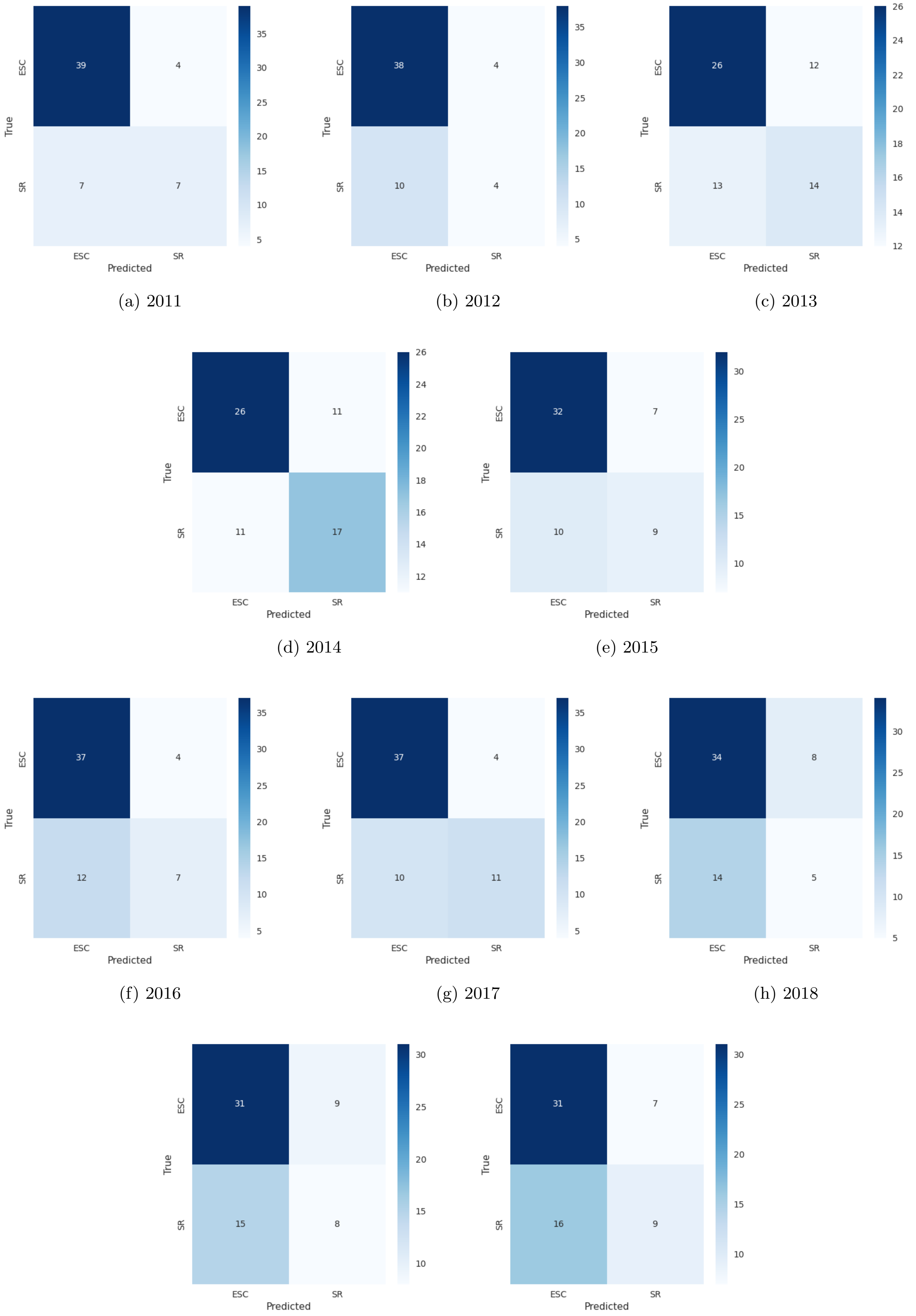


Figure 3.2: Confusion matrix of total data.
Constructed from evaluating the general ESC-SR classifier.

		1	2	3	4	5
2011	Accuracy	0.79	0.81	0.77	0.76	0.77
	St.dev.	0.12	0.13	0.21	0.20	0.12
2012	Accuracy	0.75	0.72	0.74	0.73	0.72
	St.dev.	0.20	0.13	0.23	0.13	0.15
2013	Accuracy	0.59	0.70	0.71	0.64	0.63
	St.dev.	0.21	0.24	0.17	0.25	0.19
2014	Accuracy	0.64	0.70	0.69	0.71	0.70
	St.dev.	0.16	0.17	0.20	0.11	0.22
2015	Accuracy	0.67	0.64	0.60	0.66	0.64
	St.dev.	0.19	0.19	0.19	0.25	0.19
2016	Accuracy	0.78	0.77	0.73	0.72	0.77
	St.dev.	0.15	0.15	0.11	0.20	0.11
2017	Accuracy	0.74	0.73	0.68	0.71	0.72
	St.dev.	0.16	0.13	0.19	0.11	0.18
2018	Accuracy	0.65	0.61	0.62	0.64	0.60
	St.dev.	0.22	0.15	0.25	0.20	0.19
2019	Accuracy	0.58	0.64	0.68	0.66	0.62
	St.dev.	0.23	0.28	0.21	0.23	0.20
2021	Accuracy	0.61	0.58	0.61	0.60	0.59
	St.dev.	0.16	0.22	0.20	0.16	0.14

Table 3.2: Accuracy and Standard Deviations per KFold repetition of cross-validating the general ESC-SR classifier.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure 3.3: Confusion matrices per year.
Constructed from evaluating the general ESC-SR classifier.

Year	Classifier	Accuracy	MCC	Cohen's Kappa	F-score	ROC AUC
Total	General	0.74	0.38	0.37	0.72	0.78
2011	General	0.81	0.44	0.44	0.80	0.73
	Yearly	**0.82**	**0.49**	**0.48**	**0.81**	0.72
	Selection	0.81	0.44	0.44	0.80	**0.75**
2012	General	**0.75**	**0.24**	**0.22**	**0.72**	0.59
	Yearly	0.68	0.00	0.00	0.65	**0.65**
	Selection	0.64	-0.20	-0.18	0.59	0.62
2013	General	0.62	0.20	0.20	0.61	0.69
	Yearly	0.60	0.17	0.17	0.60	0.69
	Selection	**0.65**	**0.26**	**0.26**	**0.64**	**0.72**
2014	General	0.66	0.31	0.31	0.66	0.75
	Yearly	**0.69**	**0.38**	**0.38**	**0.69**	**0.75**
	Selection	0.66	0.31	0.31	0.66	0.72
2015	General	**0.71**	**0.31**	**0.31**	**0.70**	0.71
	Yearly	0.66	0.18	0.18	0.64	**0.77**
	Selection	0.66	0.15	0.15	0.64	0.72
2016	General	0.73	0.33	0.31	0.71	**0.78**
	Yearly	**0.78**	**0.47**	**0.44**	**0.77**	0.77
	Selection	0.68	0.20	0.19	0.66	0.77
2017	General	**0.77**	**0.47**	**0.46**	**0.76**	0.79
	Yearly	0.73	0.36	0.35	0.72	0.78
	Selection	0.76	0.44	0.43	0.75	**0.80**
2018	General	0.64	0.08	0.08	0.62	0.55
	Yearly	**0.67**	**0.17**	**0.16**	**0.65**	0.54
	Selection	0.62	0.08	0.08	0.61	**0.57**
2019	General	0.62	0.13	0.13	0.60	0.62
	Yearly	**0.68**	**0.26**	**0.21**	**0.64**	**0.64**
	Selection	0.65	0.18	0.17	0.62	0.60
2021	General	0.64	0.20	0.19	0.61	0.61
	Yearly	**0.65**	**0.25**	**0.25**	**0.64**	0.61
	Selection	0.57	0.05	0.05	0.55	**0.63**

Table 3.3: Evaluation metrics of the three ESC-SR classifiers.

3.2.2 Results of the Eurovision-Melodifestivalen Classification

In this section we present the results of our classification of Eurovision and Melodifestivalen songs.

Feature Importance

The feature importance plots for all classifiers can be found in Appendix A.3.1. In Figure 3.4, we include the fifteen most important features of the general classifier. Again, the F-score indicates the number of times a feature was split on in the decision tree.

In training the general classifier and building its decision tree, 294 features were split on at least once. These features were later used to train the selection classifiers with the yearly data.

Classifier Evaluation

As before, we applied cross-validation to evaluate our classifiers. The resulting confusion matrix from the Leave-One-Group-Out validation of the general classifier is shown in Figure 3.5. We infer that 78% of the Eurovision songs was classified correctly, while for the Melodifestivalen songs this was 56%. Figure 3.6 shows the confusion matrices from evaluating the general classifier per year using the repeated KFold validation. The average accuracy and standard deviation per KFold repetition can be found in Table 3.4. We see that all but two values are within one standard deviation. The two exceptions are marked in the table. Still, these values are within two standard deviations.

In a similar fashion, we constructed the confusion matrices for the repeated KFold validation of the yearly classifiers and the selection classifiers. These can be found in Appendix A.3.2.

Finally, using the information from all the matrices, we evaluated the performance of the classifiers using the previously mentioned evaluation metrics. These results are presented in Table 3.5. For each year and metric, we emphasised the highest scoring classifier. If multiple values are highlighted, they were equal up to five decimals.

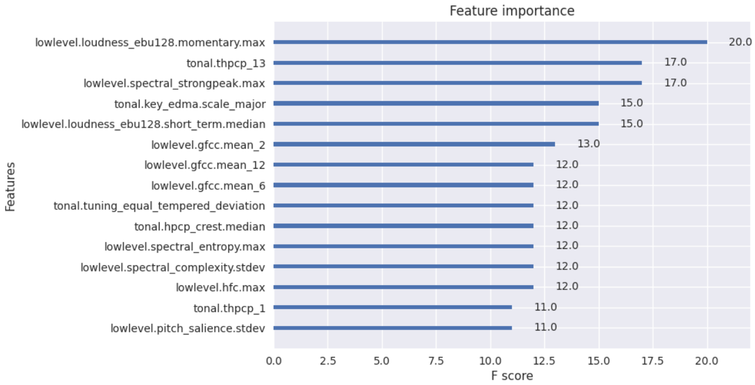


Figure 3.4: Top 15 feature importance plot for the general ESC-MF classifier.

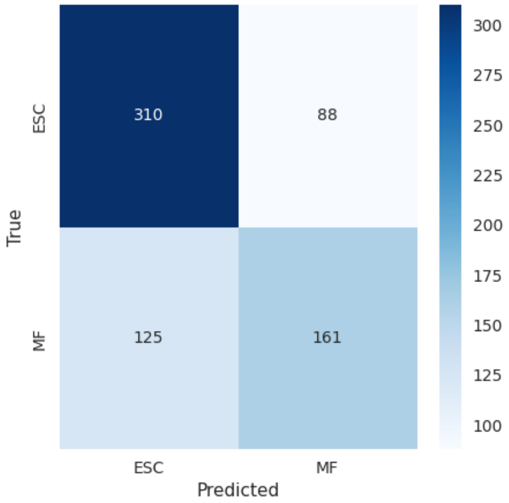
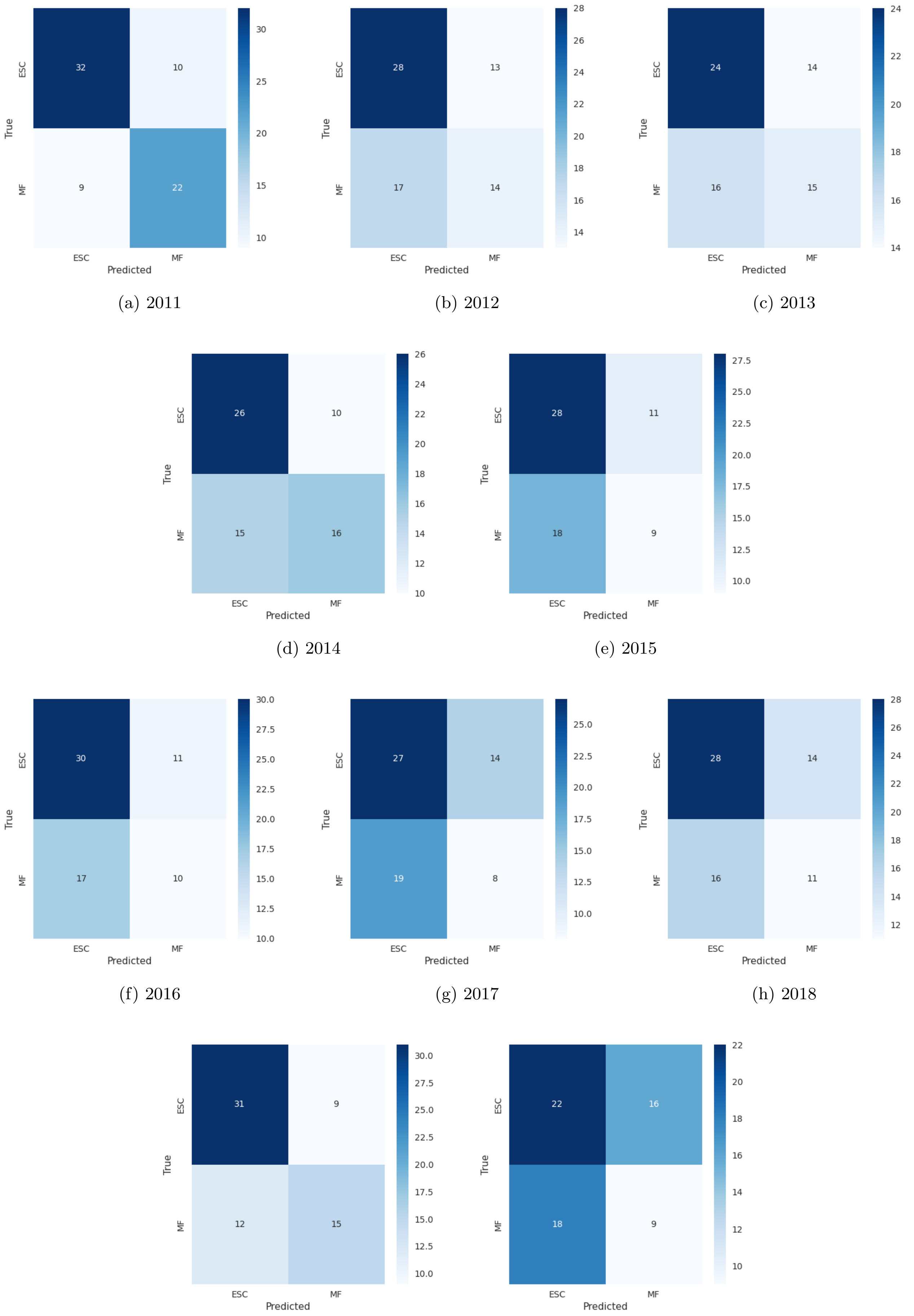


Figure 3.5: Confusion matrix of total data.
Constructed from evaluating the general ESC-MF classifier.

		1	2	3	4	5
2011	Accuracy	0.77	0.76	0.73	0.71	0.71
	St.dev.	0.17	0.10	0.18	0.15	0.16
2012	Accuracy	0.53	0.57	0.53	0.68	0.61
	St.dev.	0.17	0.19	0.19	0.21	0.15
2013	Accuracy	0.57	0.63	0.64	0.59	0.59
	St.dev.	0.21	0.23	0.11	0.18	0.14
2014	Accuracy	0.61	0.64	**0.72**	0.62	**0.58**
	St.dev.	0.16	0.12	**0.13**	0.27	0.17
2015	Accuracy	0.58	0.61	0.63	0.61	0.59
	St.dev.	0.15	0.18	0.17	0.14	0.13
2016	Accuracy	0.65	0.63	0.64	0.52	0.56
	St.dev.	0.14	0.10	0.15	0.18	0.21
2017	Accuracy	**0.49**	**0.59**	0.57	0.58	0.56
	St.dev.	0.21	**0.09**	0.17	0.17	0.20
2018	Accuracy	0.48	0.54	0.51	0.55	0.56
	St.dev.	0.22	0.15	0.21	0.20	0.19
2019	Accuracy	0.67	0.65	0.67	0.68	0.72
	St.dev.	0.15	0.19	0.20	0.22	0.21
2021	Accuracy	0.50	0.57	0.54	0.54	0.51
	St.dev.	0.20	0.12	0.10	0.16	0.20

Table 3.4: Accuracy and Standard Deviations per KFold repetition of cross-validating the general ESC-MF classifier.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure 3.6: Confusion matrices per year.
Constructed from evaluating the general ESC-MF classifier.

Year	Classifier	Accuracy	MCC	Cohen's Kappa	F-score	ROC AUC
Total	General	0.69	0.35	0.35	0.69	0.72
2011	General	**0.74**	**0.47**	**0.47**	**0.74**	0.77
	Yearly	**0.74**	0.46	0.46	0.74	**0.79**
	Selection	**0.74**	**0.47**	**0.47**	**0.74**	0.79
2012	General	0.58	0.14	0.14	0.58	0.66
	Yearly	0.58	0.14	0.14	0.57	**0.67**
	Selection	**0.61**	**0.20**	**0.19**	**0.61**	0.65
2013	General	0.57	0.12	0.12	0.56	**0.61**
	Yearly	0.58	0.15	0.15	0.58	0.57
	Selection	**0.63**	**0.23**	**0.23**	**0.62**	0.59
2014	General	0.63	0.24	0.24	0.62	0.67
	Yearly	0.61	0.21	0.21	0.61	0.67
	Selection	**0.64**	**0.28**	**0.27**	**0.64**	**0.72**
2015	General	0.56	0.06	0.05	0.55	**0.57**
	Yearly	0.58	0.08	0.08	0.56	0.55
	Selection	**0.65**	**0.26**	**0.26**	**0.64**	0.53
2016	General	0.59	0.11	0.11	0.58	0.57
	Yearly	**0.63**	**0.19**	**0.19**	**0.61**	**0.60**
	Selection	0.60	0.15	0.14	0.59	0.59
2017	General	0.52	-0.05	-0.05	0.50	**0.58**
	Yearly	**0.63**	**0.21**	**0.21**	**0.62**	0.51
	Selection	0.59	0.10	0.10	0.57	0.54
2018	General	**0.57**	**0.08**	**0.08**	**0.56**	0.50
	Yearly	0.54	0.01	0.01	0.53	0.51
	Selection	0.54	0.00	0.00	0.53	**0.56**
2019	General	0.69	0.34	0.34	0.68	0.73
	Yearly	**0.72**	**0.41**	**0.41**	**0.72**	0.70
	Selection	**0.72**	**0.41**	**0.41**	**0.72**	**0.76**
2021	General	0.48	-0.09	-0.09	0.47	0.48
	Yearly	**0.57**	**0.11**	**0.11**	**0.57**	0.52
	Selection	0.49	-0.05	-0.05	0.49	**0.52**

Table 3.5: Evaluation metrics of the three ESC-MF classifiers.

3.2.3 Results of the Three-Class Classification

After constructing the separate classifiers we trained one general classifier comparing all three contests. Its results are presented here.

Feature Importance

Figure 3.7 shows the fifteen most important features of the ESC-SR-MF classifier. The complete importance plot can be found in Appendix A.4. The decision tree was split on 359 features at least once.

Classifier Evaluation

We evaluated the classifier using the Leave-One-Group-Out cross-validation method. Figure 3.8 depicts the resulting confusion matrix. We find an accuracy of 0.59 and an MCC of 0.35; Cohen's Kappa is 0.35 and the F-score is 0.59.

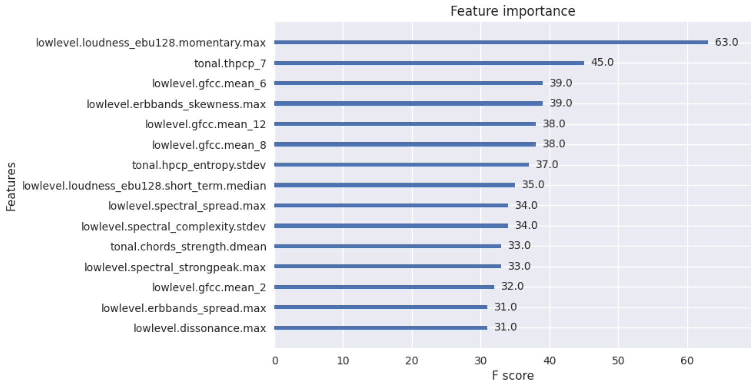


Figure 3.7: Top 15 feature importance plot for the ESC-SR-MF classifier.

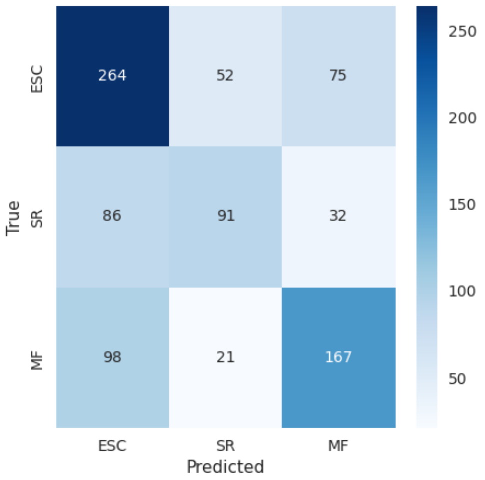


Figure 3.8: Confusion matrix of total data.
Constructed from evaluating the general ESC-SR-MF classifier.

3.3 Discussion

In this section we discuss the results and provide possible explanations for the observed phenomena.

3.3.1 Feature Importance

First, we discuss the most important features of the various classifiers.

Eurovision and Sanremo

We consider the fifteen most important features for the Eurovision-Sanremo general classifier. Unfortunately, providing a clear human interpretation is impossible for many of the features. However, for a couple of features an explanation is already available or can be extracted from an informal auditory analysis. The most important ones for our research are discussed here.

A first group of features that is essential in our plots, is the GFCC values. The GFCC is a 13-dimensional vector that as a whole is believed to capture the timbre of a song. However, there is no clear interpretation for the independent components of the GFCC vector. In our general ESC-SR classifier the features `lowlevel.gfcc.mean_8` and `lowlevel.gfcc.mean_12`, corresponding to the ninth and thirteenth component (note that the implementation counts from 0), are the most important GFCC features. In order to get a clearer understanding of these components, we performed an analysis of the values corresponding to these abstract features. That is, we sorted all songs by their GFCC component value and tried to discover a pattern by listening to the songs with high and low values.

For the thirteenth component, i.e. the feature `lowlevel.gfcc.mean_12`, we found that high values correspond to female voices, while lower values correspond to male voices. However, it does not seem to be the case that the higher (or lower) values necessarily correspond to high (or low) notes and tones. For example, the fifth highest value corresponds to the song *Il vento e le rose* by Patty Pravo from the 2011 Sanremo festival. While thus clearly sung by a female voice, the singing voice does not sound particularly high. On the other hand, the song *Tomorrow* by Gianluca from the 2013 Eurovision Song Contest scored the lowest value, even though the vocals are quite light and high pitched. Hence, the feature probably describes some other aspect of the male and female voice.

For the feature `lowlevel.gfcc.mean_8` we performed a similar analysis. Here, we presume that the feature gives information about the lightness and brightness of the singing voice. High values seem to correspond to voices that sound slightly nasal and pinched —such as in *Bagnati dal sole* by Noemi from the 2014 Sanremo Festival—, or tones that sound sharp —such as in *Un millione di cose da dirti* by Ermal Meta from the 2021 Sanremo Festival. Low values on the contrary correspond to songs with soft, bright and young sounding vocals —such as *Wars for nothing* by Boggie from the 2015 Eurovision, or *On my own* by Bishara from the 2019 Melodifestivalen.

Another group of features that appear more than once in the importance plot are the THPCP-values. This transposed harmonic pitch class profile consists of 36 values representing the intensities of pitch classes corresponding to the different notes. Here, the main values are `tonal.thpcp_2`, `tonal.thpcp_12` and `tonal.thpcp_32`, which correspond to D, C and A♭ respectively.

Among the remaining features, we point out the second most important feature, the maximum of the ERB-bands skewness. This indicates the most skewed moment, which most likely corresponds to a very high or very low note. Moreover, the classifier used multiple statistics that describe the chords strength. This feature tells us something about how easy it is to predict the sequence of chords used in the song. A high value indicates a harmonic chord progression without many riffs and licks; a low value on the other hand corresponds to progressions which are difficult to predict. Finally, the feature describing the minimum of the momentary loudness indicates the most silent moment.

Eurovision and Melodifestivalen

When we inspect the 15 most important features for the ESC-MF classifier, we again see multiple GFCC features.

For the feature `lowlevel.gfcc.mean_12` a possible interpretation was described before, namely that low values correspond to male singing voices, while high values correspond to female voices.

An informal analysis of the feature `lowlevel.gfcc.mean_2` seems to suggest that higher values correspond to sounds that are more whispery and airy. This is illustrated in the songs with the two highest values: *La notte* by Arisa from the 2012 Sanremo and *Calm after the storm* by The Common Linnets from the 2014 Eurovision Song Contest. It can be heard that the vocalists let more air out while singing these songs (or at least the considered fragments), which results in a certain lightness to the sound. On the opposite end one would then expect very heavy, perhaps even shout-like vocals. This however does not seem to be the case exactly. It is therefore likely that the feature partially corresponds to the observed quality of the vocals, but is more nuanced beyond our understanding.

For the final GFCC value, `lowlevel.gfcc.mean_6`, we perceive a difference in the mood of the songs. The low values are assigned to songs with lower intensity that feel sad. They are mostly dramatic ballads, such as *Why am I crying?* by Molly Sandén from the 2012 Melodifestivalen. The high values then seem to be assigned to songs with a higher intensity and an overall slightly angry mood. Here we find for example power ballads such as *Amen* by Ana Soklič from Eurovision 2021, or rock songs such as *Contagious* by Mustasch from the 2021 Melodifestivalen.

In the THPCP vector we observe that the features `tonal.thpcp_1` and `tonal.thpcp_13` are most important. Both correspond to a C♯.

Concerning the other features, we point to two that stand out. The most used feature, `lowlevel.loudness_ebu128.momentary.max`, considers the loudest moment of a fragment. The fourth most important feature, `tonal.key_edma.scale_major`, indicates that a distinction has been made based on whether songs were in the major scale as determined by the EDMA profile. This profile is trained on electronic dance music.

Thee-Class Classification

Unsurprisingly, among the most important features for our three-way classifier, we find many features that appeared in one or both of the separate general classifiers. Examples include the features `lowlevel.gfcc.mean_6`, `lowlevel.gfcc.mean_12` and `lowlevel.erbbands_skewness.max`. This indicates that our ESC-SR-MF model combines the two classifiers into one, as could have been expected.

The second most important feature, `tonal.thpcp_7`, does not occur in the top 15 of either separate classifier. This pitch class corresponds to a G.

General Discussion

The feature importance of the classifiers helps us to get an understanding of the difference between the songs from the various contests. However, as mentioned before, not many features come with a natural interpretation. This makes the plots rather abstract and complicates giving a direct explanation of the results.

Moreover, while the feature importance tells us something about which properties were meaningful for the overall classification process, we don't know anything about what songs they represent. For example, was some feature important because high values consistently corresponded to Eurovision songs? Or because certain ranges of values belonged to Sanremo songs? Questions like these cannot be answered by the importance plots alone and require further analysis.

For both groups of classifiers we found that the importance plots constructed by the yearly classifiers differ from year to year. This suggests that the contests evolve over time and that each year different musical features are prioritised. Further analysis of these results might provide insights into the evolution of the represented music at each of the festivals.

The importance plots of the selection classifiers are similar to the plots of the corresponding yearly classifiers. Given that they were trained on the same data, this outcome was expected.

3.3.2 Classifier Evaluation

Secondly, we discuss the evaluation results of our trained classifiers and provide possible interpretations for the results.

Eurovision and Sanremo

We start the evaluation of our ESC-SR classifier by pointing out that, overall, the results generated by our general classifier are similar to the results obtained by training a yearly classifier, both on the full set of features and on a selection. This indicates that our general classifier is representative not only on all years combined, but also when considering a specific year. Therefore, it justifies our focus on the general classifier as our main classifier.

Over the totality of considered years, our general classifier obtained an accuracy of 0.73 and an MCC of 0.38. While these scores certainly indicate that our classifier is successful to some extent, the confusion matrix in Figure 3.2 shows that it is remarkably better at correctly classifying Eurovision songs than it is at correctly classifying Sanremo songs, with success rates of 0.87 and 0.48 respectively.

A straightforward explanation would be the class imbalance. Each year, there are approximately twice as many Eurovision songs as Sanremo songs. Hence, a very naive classifier would classify the majority of the songs as belonging to the Eurovision Song Contest. However, if our classifier would behave in a similar manner as this naive model, the MCC value would be expected to be closer to zero. From its definition, we know that MCC corrects for imbalanced data sets. We thus argue that our classifier is more successful than a model just based on chance. The observed difference in success rate between classifying Eurovision songs and Sanremo songs is therefore likely caused by some other factors.

One possible explanation could be the 'Italianness' of the Italian songs. If there were a distinct Italian style, and consequently a distinct non-Italian style, this would make classifying Eurovision songs easier. That is, if a song does not conform to this Italianness, it must belong to the Eurovision Song Contest instead of the Sanremo Festival. On the other hand, the Italian songs would be much harder to classify, since they could belong to either festival.

The full name of the national competition, the *Festival of the Italian song in Sanremo*, might be seen as an affirmation of this hypothesis, since it emphasizes the Italianness of the participating songs. However, as far as we know, there is no scientific evidence of a distinct Italian music style as represented at the Eurovision Song Contest and Sanremo in the last ten years. A more elaborate study would be needed to test this hypothesis.

We do know that Italy is one of the few countries to almost always enter the international contest with a song in Italian. Since 1999, the participating countries are free to select the language or languages they want to perform their song in. This has resulted in many artists performing their songs in English. One of the few exceptions is Italy, who always sing in Italian and have used additional English lyrics in only three entries between 2011 and 2021. However, it is unlikely that our classifier is able to pick up on the language difference and use this to improve the classification. Even though the English and Italian languages possess different sounds when being spoken, this is already less apparent in sung lyrics. We therefore do not believe that the language differences have had an influence on the performance of our classifiers.

When looking at the confusion matrices in Figure 3.3, the years 2014, 2018, 2019 and 2021 stand out.

In 2014, the classifier is more accurate at classifying Sanremo songs than in any other year. This is supported by relatively high MCC and ROC AUC values. A possible explanation could be the setup of the 2014 Sanremo edition. In this edition fourteen artists participated, each with two different songs. During the contest one of the songs was chosen for each artist by the public and jury to advance to the final rounds. This alternative process might have influenced the choice of the songs. For example, an artist could now enter with both a more modern and a more traditional song, whereas they otherwise would have had to make this choice beforehand. This might have resulted in a more than average number of 'Italian' songs, which would explain the shift in classification. Alternatively, it might just be the case that the classifier 'recognises' an artist and their style. That is, if the classifier encountered one of the artist's songs during training, it might have learnt their style and be more likely to correctly classify the second song. We note, however, that the same contest setup was used in 2013, but that we do not see similar outcomes there. This possibly refutes the above explanations.

On the other hand, the results from 2018, 2019 and 2021 show a relatively low accuracy and MCC. This might suggest a trend towards a more 'Eurovision sound' and less Italianness at the Sanremo Festival during these last few years.

Most notable is the 2019 Sanremo Festival. Interestingly, this edition sparked a lot of controversy exactly around the topic of Italianness. The artist Mahmood, who has an Egyptian father, but was born and raised in Italy by his Italian mother, participated in the festival with his song 'Soldi', which features sentences in Arabic. After he had won the festival thanks to the votes of music journalists and the expert jury, he and his song were criticised for not being Italian enough, in particular by the far-right politician Salvini [30]. This sparked new discussions on the Italian identity and the Italian song [3].

Even though Mahmood was met with criticism, it is also said that he started a new era for the Festival di Sanremo[1]; one with more attention for diverse music genres and opportunities for young and upcoming artists. The 2021 edition seemed to fit this prediction, with the young band Måneskin winning the festival (and later also the Eurovision Song Contest) with a glamrock song.

To get a better understanding of the workings of our classifier, we analyse the misclassified songs. By closely listening to these songs, we found the following.

Recall that the general classifier incorrectly classified 52 Eurovision songs as belonging to the Sanremo, out of the considered 401 Eurovision songs, as was seen in Figure 3.2. Among these mistaken classifications, we find two notable groups. Firstly, we observe that multiple misclassified songs are light, acoustic love songs. Examples of this genre of songs classified as Sanremo include *Contigo hasta el final* by ESDM from 2013; *When we're old* by Ieva Zasimauskaitė from 2018; and *Together* by Ryan O'Shaughnessy from 2018. A special subgroup of these songs seems to be acoustic love duets, such as 2015's *Goodbye to yesterday* by Elina Born & Stig Rästa, or groups, such as in 2019's *Heaven* by D mol. A second observed group are songs that do not sound very Eurovision-like. We argue that these songs contain elements which are not often seen at the international festival and were therefore classified as Sanremo songs. One example is the song *No prejudice* by the band Pollapönk. With their objective to write punk-rock influenced children's songs, this was definitely unique for the Eurovision Song Contest. Another example is 2019's *Telemóveis* by Conan Osíris. With its electronic fado sounds combined with African world beat rhythms, this song is a mixture of elements from many genres that are not prevalent in the Eurovision. Therefore understandably, these songs were difficult to classify which could explain their misclassification. We note that the misclassifications were not limited to these groups. However, these were clusters of songs that were represented most often among the mistakenly classified songs.

On the other hand, we can look at Sanremo songs which were classified as Eurovision. Again from Figure 3.2, recall that 109 out of 209 Sanremo songs were classified as Eurovision. This indicates that it was more difficult to classify Sanremo songs, which is reflected also in the misclassified songs. While we attempt to highlight certain clusters, it should be mentioned that the distinctions were less clear and that we encountered songs from many different genres. Most notable are the Sanremo songs classified as Eurovision that are more up-tempo and up-beat. Examples include 2021's indie pop-rock song *Combat pop* by Lo Stato Sociale and 2015's dance influenced *Fatti avanti amore* by Nek. We also find that more electronically produced songs tend to be classified as belonging to the Eurovision, such as the 2021 songs *Chiamami per nome* by Fedez & Francesca Michielin, and *La genesi del tuo colore* by Irama. However at the same time slower, more subdued songs get the incorrect Eurovision label as well, such as *Portami via* by Fabrizio Moro from the 2017 Sanremo.

These observations could be interpreted as an affirmation of the proposed theory of Italianness. The misclassified songs indicate, in general, that the Sanremo Festival hosts more slow, acoustic-like pop-songs than the Eurovision does. On the other hand, the electronic dance music is less frequent in the Sanremo and more Eurovision-like.

Finally, we would like to note that the classifier trained in this chapter has no predictive power for the Sanremo. That is, it is not the case that songs classified as belonging to the Sanremo festival consistently obtain high (or low) rankings in the

[1] E.g. in `https://www.lacasadelrap.com/2022/02/02/festival-sanremo-urban/`. Accessed 13-4-2022.

national contest; nor is this the case for songs classified as belonging to the Eurovision Song Contest.

Eurovision and Melodifestivalen

Now, we discuss the classifiers for the Eurovision and the Melodifestivalen. First we note that the general, yearly and selection classifiers all have similar performance. Therefore, we again focus mainly on the general classifier.

Over the totality of years, the general classifier scores an accuracy of 0.69 and an MCC of 0.35. While this is certainly not bad, again our confusion matrix in Figure 3.5 shows that Eurovision songs are more often correctly classified than Melodifestivalen songs, with success rates of 0.78 and 0.56 respectively. Note however that the majority of Melodifestivalen songs is correctly classified. The MCC value indicates that this success is not merely due to chance.

As for the Eurovision-Sanremo classifier, we could explain the results by looking at the Swedish style. In the years we are considering, the Swedish Eurovision entry was an electronic pop song sung by a young man seven out of ten times. However, songs like these are quite common at the Eurovision Song Contest and so it is not unlikely that they get misclassified. Thus, we would again see that there are songs at the international festival that are definitely not Swedish and hence get classified correctly, while the songs performed at the Melodifestivalen could belong to either contest.

When looking at the confusion matrices for the separate years, we still find a successful classification of the Eurovision songs, but classification of the Melodifestivalen songs deteriorates towards a fifty-fifty split, or worse. Positive exceptions are 2011 and 2019, with relatively high accuracy and MCC. On the opposite end are the years 2015, 2016, 2017, 2018 and 2021, with accuracies around 0.5, MCC values close to 0 and a majority of the Melodifestivalen songs misclassified.

The MCC value of these latter years indicates that our classifier performs no better than a naive classifier based on chance. This suggests that the songs performed at the Eurovision Song Contest and Melodifestivalen are very similar and difficult to distinguish.

The year 2019 seems to break with this trend having a higher accuracy and MCC. However, we have not been able to find an explanation for this anomaly. By listening to the songs we did not find a significant difference from the songs performed at the other editions. Moreover, there was no impacting rule change for the 2019 festival, nor was there any controversy surrounding it.

Again, we analysed the misclassified songs to gain insights into the differences between the national and international festival.

Among the 88 Eurovision songs classified as belonging to the Melodifestivalen, we find many different musical styles. As a first group we encountered multiple Schlager songs, such as 2017's *Yodel it!* by Ilinca & Alex Florea. Given that the Melodifestivalen is sometimes also called the Schlagerfestivalen, it is no surprise that these songs are classified as Swedish. A second group comprises rock songs. Examples include the Scandinavian entries *Dark Side* by Blind Channel from 2021 and *I feed you my love* by Margaret Berger from 2013, but also the 2021 winner *Zitti e Buoni* by Måneskin. Finally, we also observed more generic misclassified songs. For one, we encountered several up-tempo dance numbers, such as 2016's *Walk on water* by Ira Losco. Moreover, a couple of ballads also got classified as Swedish, for example the song *Growing up is getting old* by Victoria. Interestingly, many of these songs are co-written by Swedish songwriters.

The other way around, we found 125 Melodifestivalen songs misclassified. This collection however, was very diverse and it was difficult to distinguish evident groups. For example, we encountered both high-pitched female voices singing acoustic songs such as *En himmelsk sång* by Ellinore Holmer from 2014, and louder rock songs such as *Runaways* by Eclipse from 2016. Moreover, also pop-dance ballads such as 2018's *Every single day* by Felix Sandman and 2019's *Hold you* by Hanna Ferm & Liamoo got misclassified as Eurovision songs.

To summarise, these observations show once again that it is difficult to distinguish Eurovision and Melodifestivalen songs. Apart from the Schlager songs, which consistently got classified as Swedish, different genres could be classified as belonging to both contests and were also confused in either case. This indicates that the Swedish national competition and the Eurovision indeed at least partially host very similar music. It is also good to note that Eurovision songs do not only often sound Swedish, but are actually (co-)written or produced by Swedish musicians. For example, in 2019 eight non-Swedish countries competing in the Eurovision, including that year's winner, entered a song co-written by Swedish composers. Given this wide-spread influence, it is not surprising that Eurovision songs tend to sound 'Swedish' and that they are hard to distinguish from Melodifestivalen songs.

Finally, also for this classifier we would like to point out that it has no predictive power for the Melodifestivalen. It is not the case that there is any relation between the classification of a song and its final ranking in the festival.

Three-Class Classification

For the three-way classifier we only trained a general model, evaluated by Leave-One-Group-Out cross-validation. From its confusion matrix we derive that it is most successful at classifying Eurovision songs. This is in line with the observation made in analysing the previous separate classifiers. We also see that incorrectly classified Eurovision songs (32%) are more often believed to belong to the Melodifestivalen with 59% than to the Sanremo festival with 41%. For both national festivals we note that an incorrect prediction is more often classified as Eurovision than as the other national festival. That is, of the incorrect guesses for the Sanremo (56%), 73% was classified as Eurovision, while the rest was classified as Melodifestivalen; for the incorrect classifications of the Melodifestivalen (42%), we found 82% classified as Eurovision and the rest as Sanremo.

These results suggest that there are more songs at the Eurovision Song Contest which resemble Swedish songs, than there are that resemble Italian songs.

General Discussion

To conclude, we discuss some general aspects of the presented method and results.

When inspecting the results in Table 3.3 and Table 3.5, we note that the best ROC AUC value for a particular year often belongs to a different classifier than the one that would be deemed 'best' based on the other metrics. Unlike the other metrics, the ROC AUC is biased towards one of the classes. Also, in the case of repeated KFold, the ROC AUC is calculated from the average values from all repetitions. This is different from the other metrics, which are calculated based on the confusion matrix constructed by assigning to each song their majority classification. Both these observations might explain the divergent best ROC AUC.

Moreover, we find that in some cases the selection classifier outperforms the yearly classifier. This is remarkable as both classifiers are trained on the same songs, with the yearly classifier having access to a larger number of features. The difference in performance is most likely explained by overfitting. By reducing the number of possible features, we avoid making the classifier too specific. Therefore, in some cases, the selection classifier actually reaches a better performance than the yearly classifiers.

Lastly, we mention some restraining choices that were made during the course of this research. Most notably, we only considered 29 seconds of each song. Even though (a part of) the chorus certainly gives a good feel for a song, it of course fails to capture various aspects of a song, such as its structure or key changes. To comprise also these facets and represent a song more precisely, one might include multiple fragments per song.

Moreover, we chose to use studio recorded versions of all songs. This choice was made for the sake of consistency, given that some songs did not have a live performance available. While the live performance is definitely an important aspect of the music festivals and the final ranking, we believe that for our goal of comparing the music represented at the contests, consistency is more important for training a valid classifier.

Predictions

In the previous chapter we have seen how musical features can be used in a classification task. In particular, we have presented a classifier that attempts to distinguish songs from the Eurovision Song Contest, the Sanremo Festival and the Melodifestivalen, based on their musical properties. A natural follow-up question would then be whether these features and classifications have any predictive power. That is, could we for example use the results from the Sanremo Festival and the Melodifestivalen to make a prediction about the outcome of the Eurovision Song Contest? Or perhaps we could predict how Italy and Sweden will vote? In this chapter, we will first discuss two approaches that attempt to predict the outcome of the Eurovision based on the results from the national competitions. Subsequently, we will look into a more specific prediction based on the points awarded in the Eurovision final.

It is of course notoriously hard to predict the outcome of the Eurovision Song Contest —and probably for the better, otherwise it would be quite boring to watch. It is especially difficult when only considering the musical features of the songs, while there are many more aspects that might influence the audiences votes, such as the overall performance, political circumstances, the popularity of an artist outside of the contest, etcetera. Therefore, we do not expect any of the predictive endeavours in this chapter to be overwhelmingly successful at forecasting the outcome of the Eurovision or the voting behaviour of certain countries. Nonetheless, we believe that by analysing the predictions we might still learn valuable lessons about the interplay between the national and international contests.

4.1 Predicting the Eurovision Ranking

When it is almost time for the Eurovision Song Contest, questions about who will win the festival keep many fans and bookmakers busy. In this section we will briefly discuss our attempt to predict (some part of) the outcome of the Eurovision Song Contest, by using the outcome of the Sanremo Festival and the Melodifestivalen.

4.1.1 Method

To predict (a part of) the outcome of the Eurovision Song Contest, we present two strategies. Both strategies use the same music features data as before. See subsection 3.1.1 for a description of how this data was obtained.

For our first approach, we assume that we are only interested in predicting which songs will end up in the top ten. To this extent we trained a classifier, again using the XGBoost algorithm. For all songs from all considered competitions we determined whether it obtained a top ten position (‘Yes’), or not (‘No’). We then proceeded per year; for each year we trained a classifier on the songs from the Sanremo and the Melodifestivalen. So, we provided it these songs' musical features and whether or not they reached the top ten. Subsequently, we fed it the features of the Eurovision songs from the corresponding year and computed the probabilities of each song belonging to the top ten. We did not include the Italian and Swedish entries. Finally, we selected the ten songs with the highest 'Yes' probability and classified them into the top ten. The rest got classified as 'No'. Note that this interference makes this approach slightly different from a standard classification, but that it ensures that we classify exactly ten songs as belonging to the top ten.

For the second approach we are interested in a full ranking of the Eurovision songs. To establish this, we used the XGBoost ranking algorithm. Again per year, we trained a ranker on the musical features and the outcome of the songs in the Sanremo Festival and the Melodifestivalen. Any song that did not receive a final rank in these competitions —because they were eliminated in one of the qualifying rounds—, was assigned rank 20. This number was approximately the average of the non-assigned ranks in both competitions. The model was then applied to the Eurovision songs from the corresponding year —once again omitting Italy and Sweden. Finally, the outputted predicted ranking was compared to the actual outcome by computing Spearman's correlation coefficient.

4.1.2 Results and Discussion

We present and discuss the results of both approaches.

Classification

As mentioned before, we trained the classifiers per year on the songs from the Sanremo Festival and the Melodifestivalen. A ten-fold KFold cross-validation gave an average accuracy of 0.56 and an average MCC of 0.03 over all years. These scores already indicate that the classifier does not perform much better than chance. The results from the Eurovision classification are presented in Table 4.1. We find an average accuracy of 0.63 and an average MCC of 0.04. Moreover, on average 2.9 songs that should have been classified as top ten were indeed classified as such. We find the best classification in 2011 and the worst in 2021.

This approach to predicting the Eurovision Song Contest is thus not very successful. We see that especially in more recent years, the accuracy, MCC and number of correct

	Accuracy	MCC	N
2011	0.71	0.21	4
2012	0.70	0.20	4
2013	0.68	0.18	4
2014	0.66	0.16	4
2015	0.63	0.05	3
2016	0.60	-0.07	2
2017	0.65	0.07	3
2018	0.66	0.07	3
2019	0.54	-0.21	1
2021	0.51	-0.23	1

Table 4.1: Performance top ten classifier. N is the number of correct 'Yes' instances.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021
SCC	-0.04	0.22	0.19	-0.11	-0.18	-0.10	0.13	0.21	-0.13	0.13

Table 4.2: Spearman's correlation coefficient (SCC) per year.

'Yes'-instances drop towards a performance similar to or even worse than chance. Moreover, note that for this classifier 'in the top ten' means that the song actually ended up in at least the top twelve, given that Italy and Sweden were omitted from our classification. While these are definitely the most interesting positions, there is still a large difference in popularity between the first and tenth position. This is illustrated by the phenomenon that in most years, betting auctions by the time of the Eurovision final suggest there are only two or three serious favourites for the victory.

Ranking

For the second approach we trained a model that outputs a complete ranking for each edition. We evaluated these predictions by computing Spearman's correlation coefficient between the actual ranking and the ranking based on the outcomes of the Sanremo and Melodifestivalen combined. The results are presented in Table 4.2. We find a positive coefficient in five out of ten years, with a median correlation of 0.05. For 2012 we obtain the most accurate prediction; for 2015 the worst.

Again, we can see that the model is not very good at predicting the final outcome, at least not consistently. Both the median and average correlation are close to zero, indicating a performance not better than chance.

4.2 Predicting Voting Behaviour

The previous section has shown that it is very hard to predict the outcome of the Eurovision based on the musical features of popular Sanremo and Melodifestivalen songs. Apart from the fact that we might argue about whether or not musical features alone give a good representation of the popularity of a song, it is also the case that Italy and Sweden are not the only two countries voting in the international contest. It can therefore be expected that we cannot deduce the entire outcome based on these two countries' assumed preferences.

So, if we cannot predict the entire outcome of the international music festival, can we then at least predict the Italian and Swedish preferences? That is, do Italy and Sweden award points to Eurovision songs that are musically similar to the high-scoring songs in their respective national contests? And moreover, are there perhaps countries that vote 'Italian'- or 'Swedish'-like? In this section we will try to answer these questions by analysing the points awarded by all countries in the Eurovision finals from 2011 to 2021.

4.2.1 Method

First, we collected all points awarded by all countries in the Eurovision finals between 2011 and 2021, with the exception of the cancelled 2020 edition [22, 24]. For this analysis we used the combined jury and public vote scores. Then, for each voting country we established a ranking of all entries participating in the final based on the amount of awarded points. Moreover, we worked once more with the music features as described in subsection 3.1.1.

Subsequently, we proceeded with the following steps, which we describe here for Italy and the Sanremo Festival; the applied method for Sweden and the Melodifestivalen is analogous. Per year, we trained an XGBoost ranker on the songs from the Sanremo Festival. We then fed it the songs from the Eurovision final of the corresponding year and let it produce a predicted ranking. We compared this ranking to the actual scoreboard of the international festival and computed Spearman's correlation coefficient. Thereupon, we calculated Spearman's correlation coefficient between the Italian prediction based on the Sanremo and the rankings based on the number of awarded points from all voting countries (including Italy itself). To avoid any irregularities due to a country not being able to vote for themselves, all points awarded to Italy and the country under consideration are omitted. Both the predicted ranking and the ranking based on the votes were updated accordingly.

4.2.2 Results

Italy

Here, we present the results that were obtained by comparing the voting to the predicted ranking based on the Italian Sanremo Festival. All results are depicted in Table 4.3.

The top row of Table 4.3 shows the correlation coefficients for comparing the predicted results and the actual results of the Eurovision editions per year. These coefficients show that in 2012 the Sanremo-based prediction was most in accordance with the actual outcome of the Eurovision Song Contest. On the other hand, in 2016 the prediction was the least accurate, with a strongly negative correlation coefficient.

The other rows show the correlation between the voting behaviour of all countries participating in the Eurovision and the prediction based on the outcome of that year's Sanremo Festival. The countries were sorted by their median over all years in descending order. The Italian results are highlighted in bold. We observe that Montenegro votes the most in accordance with the Italian prediction, whereas Italy comes fourth.

Sweden

For Sweden and the Melodifestivalen we performed a similar analysis. Its results are presented in Table 4.4.

The top row in Table 4.4 shows Spearman's correlation coefficient per year for the comparison of the final ranking predicted based on the Melodifestivalen and the actual results of the Eurovision Song Contest final. We see that the prediction was the most accurate in 2014, with a high correlation coefficient. The prediction was least accurate in 2021.

The comparison of the voting behaviour of all countries and the Swedish prediction is presented in the rest of the rows in Table 4.4. Again, the countries were sorted by the median over all years in descending order. The Swedish results are highlighted in boldface. We note that in only two out of ten years, the Swedish correlation coefficient is positive. We find the lowest correlation in 2015. Moreover, Sweden is positioned tenth to last in the ordering of all countries.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021
Actual outcome	-0.28	0.31	-0.19	-0.13	-0.08	-0.45	-0.20	-0.16	0.00	-0.18
Montenegro		0.55	0.04	0.28	0.08	-0.10	0.17	-0.13	0.14	
Slovakia	0.01	0.11								
Turkey	-0.23	0.34								
Italy	**0.14**	**0.10**	**0.01**	**0.16**	**-0.14**	**-0.03**	**0.06**	**-0.08**	**0.04**	**0.29**
Azerbaijan	0.06	-0.18	-0.16	0.31	0.18	-0.07	0.22	-0.19	0.04	0.01
North Macedonia	-0.03	0.36	0.10	-0.15	0.15	-0.30	-0.17	-0.25	0.08	0.24
Israel	0.15	0.12	-0.25	0.01	0.04	-0.47	0.02	-0.08	0.08	-0.03
Georgia	-0.38	0.10	-0.14	-0.02	0.07	-0.38	0.09	0.13	0.02	-0.33
Bosnia Herzegovina	-0.02	0.52				-0.07				
Armenia	-0.10		0.06	0.04	-0.31	-0.24	0.02	-0.26	0.03	
Serbia	-0.03	0.25	-0.01		-0.06	-0.09	0.04	-0.58	-0.10	-0.22
Ukraine	-0.11	-0.03	0.06	0.10		-0.26	0.15	-0.16		-0.13
Croatia	-0.33	0.29	-0.16			-0.12	-0.18	-0.03	0.09	0.00
Malta	-0.14	0.12	-0.28	-0.11	-0.21	-0.05	-0.03	-0.06	-0.30	0.09
Portugal	-0.07	-0.10		-0.22	0.02		-0.24	0.24	0.10	-0.42
Bulgaria	0.00	0.41	-0.13			-0.42	0.04	-0.09		-0.21
Poland	-0.34			-0.07	-0.12	-0.31	-0.28	0.19	0.23	-0.02
Cyprus	-0.33	0.19	-0.10		-0.28	-0.07	-0.19	-0.20	0.10	0.03
Switzerland	-0.18	0.41	-0.09	-0.33	-0.28	-0.33	-0.12	-0.06	0.14	-0.08
Latvia	-0.20	-0.06	0.05	-0.33	0.17	-0.31	-0.01	0.05	-0.15	-0.22
Lithuania	0.13	-0.08	-0.13	-0.05	0.04	-0.48	-0.08	-0.17	-0.38	-0.27
Moldova	-0.03	-0.10	-0.12	0.26	-0.15	-0.14	0.07	-0.04	-0.19	-0.25
Germany	-0.10	0.01	0.16	-0.23	-0.24	-0.50	-0.11	-0.15	0.17	-0.11
Sweden	-0.23	0.11	0.04	-0.24	-0.12	-0.25	-0.23	-0.10	-0.07	-0.00
Norway	0.05	-0.20	-0.13	-0.21	0.01	-0.33	-0.45	-0.11	-0.11	0.02
Greece	-0.44	0.24	-0.23	-0.02	-0.28	-0.27	0.02	-0.30	0.02	0.18
Russia	-0.37	0.04	-0.30	0.02	0.05	-0.13		-0.16	-0.30	0.01
Albania	-0.29	0.30	-0.28	-0.27	-0.17	-0.09	-0.02	0.06	-0.26	0.12
United Kingdom	0.33	-0.18	-0.05	-0.37	-0.14	0.10	-0.28	0.11	-0.12	-0.20
Ireland	0.12	-0.04	0.09	-0.41	-0.11	-0.32	-0.16	0.20	-0.41	-0.25
Slovenia	-0.02	0.29	0.18	-0.24	-0.40	-0.49	-0.32	-0.58	0.18	-0.04
Belarus	-0.04	-0.14	-0.18	0.18	-0.00	-0.40	-0.27	-0.14	-0.12	
France	-0.20	0.13	-0.04	-0.27	-0.22	-0.63	-0.35	-0.03	0.01	-0.09
Hungary	0.05	-0.00	-0.31	-0.15	-0.12	-0.40	-0.33	-0.12	-0.26	
Austria	0.05	0.40	-0.51	-0.10	-0.21	-0.35	-0.10	-0.20	-0.06	-0.40
Iceland	-0.08	-0.07	-0.14	-0.29	-0.12	-0.17	-0.41	-0.20	-0.09	-0.27
Australia					-0.13	0.03	-0.39	-0.24	-0.09	-0.19
Spain	0.10	-0.13	0.09	-0.21	-0.13	-0.38	-0.29	-0.24	-0.08	-0.23
Estonia	-0.06	-0.14	0.03	-0.25	-0.24	-0.18	-0.32	-0.17	-0.03	-0.33
Netherlands	-0.02	-0.11	-0.12	-0.24	-0.17	-0.56	-0.35	-0.27	-0.04	-0.19
Finland	0.15	-0.17	0.05	-0.41	-0.06	-0.20	-0.34	-0.17	-0.20	-0.40
Romania	-0.25	0.00	-0.13	-0.30	-0.40		-0.31	-0.20	-0.10	-0.15
Belgium	-0.23	0.15	-0.20	-0.26	-0.02	-0.49	-0.34	-0.16	0.17	-0.22
Denmark	0.17	-0.23	-0.20	-0.21	-0.08	-0.35	-0.28	-0.31	-0.33	0.10
San Marino	-0.24	0.18	-0.29	-0.30	-0.46	-0.23	-0.24	-0.24	-0.26	0.03
Czech Republic					-0.28	-0.36	-0.31	0.10	-0.23	-0.39

Table 4.3: Spearman's correlation coefficients computed from comparing the voting behaviour per country with the predicted outcome based on the Sanremo Festival. The top row shows the correlation between the actual outcome of the Eurovision final and the predicted outcome based on the Sanremo Festival.

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021
Actual outcome	-0.09	0.28	0.06	0.41	-0.28	0.10	0.08	-0.03	-0.02	-0.30
Russia	0.14	0.17	0.33	0.16	-0.25	-0.32		0.21	-0.35	-0.23
France	-0.11	0.20	0.03	0.08	0.19	0.15	0.30	0.22	0.12	-0.24
Israel	0.11	0.05	-0.01	0.20	-0.27	0.25	0.15	0.14	0.14	-0.03
North Macedonia	0.25	0.06	0.22	0.39	-0.35	0.19	0.13	-0.31	-0.01	-0.39
Montenegro		0.14	0.25	0.04	-0.19	-0.32	0.15	-0.32	0.15	
Armenia	0.44		0.09	0.09	-0.15	0.02	0.36	0.28	0.03	
San Marino	0.27	0.11	0.40	0.17	-0.18	-0.06	0.07	0.34	-0.15	-0.24
Romania	0.12	0.06	0.08	0.34	0.09		-0.00	-0.17	0.11	-0.15
Belgium	-0.18	0.29	0.15	0.19	-0.21	0.25	0.25	0.01	-0.18	-0.11
Ireland	-0.28	0.03	-0.33	0.41	0.13	0.38	-0.14	0.09	0.12	0.07
Germany	-0.21	0.12	0.20	0.21	-0.15	0.32	-0.09	0.04	0.17	-0.25
Lithuania	0.17	0.00	0.13	0.34	-0.31	0.38	0.05	0.09	-0.05	-0.00
Finland	0.09	0.17	0.24	0.47	-0.30	0.19	0.04	-0.25	0.02	-0.21
United Kingdom	-0.43	0.24	-0.36	0.19	-0.11	0.24	-0.23	0.01	0.08	0.21
Turkey	-0.19	0.27								
Georgia	-0.09	0.09	-0.02	0.21	0.12	0.13	0.46	-0.34	-0.30	-0.02
Ukraine	0.31	-0.04	0.13	0.05		0.01	-0.07	0.21		-0.10
Denmark	-0.21	-0.13	0.15	0.13	-0.10	0.27	-0.07	0.20	0.20	-0.16
Czech Republic					0.17	-0.00	-0.02	0.05	0.05	-0.27
Bulgaria	0.02	0.43	0.06			-0.22	0.22	-0.09		-0.43
Australia					-0.31	0.39	-0.09	0.31	0.10	-0.07
Moldova	0.14	-0.03	-0.13	0.10	0.10	-0.29	0.12	-0.09	0.05	-0.29
Switzerland	-0.12	0.07	0.40	0.07	-0.21	0.08	-0.08	-0.22	0.47	-0.19
Hungary	0.17	-0.02	0.01	0.10	-0.22	-0.07	-0.01	0.07	-0.35	
Austria	-0.37	0.06	0.02	0.37	-0.42	0.14	-0.16	-0.05	0.07	-0.09
Iceland	-0.01	-0.31	-0.03	0.18	-0.35	0.01	-0.15	0.19	0.04	-0.12
Cyprus	0.23	0.27	-0.03		-0.28	-0.27	0.22	-0.13	0.04	-0.13
Portugal	0.04	-0.03		0.54	-0.05		0.18	-0.36	-0.07	-0.18
Poland	-0.01			0.20	-0.31	0.26	-0.32	-0.15	0.15	-0.10
Slovenia	-0.06	-0.14	0.18	0.32	-0.17	0.19	-0.05	-0.17	0.05	-0.50
Azerbaijan	0.26	0.02	-0.12	0.12	-0.16	0.13	-0.01	-0.11	-0.18	-0.34
Malta	-0.14	-0.00	-0.05	0.22	-0.10	-0.38	0.15	0.12	-0.07	-0.22
Albania	0.30	-0.12	0.26	-0.09	-0.40	-0.04	0.12	-0.17	-0.05	-0.28
Italy	-0.25	-0.20	-0.13	0.04	-0.23	0.18	0.23	-0.03	0.37	-0.31
Estonia	0.16	-0.16	-0.22	0.39	0.09	-0.07	-0.23	-0.32	0.01	-0.12
Serbia	0.12	-0.10	-0.10		-0.28	-0.11	0.17	-0.06	0.06	-0.38
Sweden	**-0.28**	**-0.12**	**-0.06**	**0.17**	**-0.33**	**-0.12**	**-0.09**	**0.17**	**-0.03**	**-0.28**
Norway	-0.11	-0.13	-0.13	-0.00	-0.24	0.26	-0.05	0.23	-0.15	-0.11
Latvia	-0.15	-0.13	-0.26	0.43	-0.26	-0.02	-0.10	0.16	0.22	-0.12
Spain	-0.06	-0.17	-0.19	0.18	-0.19	-0.17	0.14	0.17	-0.27	0.12
Bosnia Herzegovina	-0.12	0.33				-0.14				
Greece	0.15	-0.10	0.01	0.48	-0.30	-0.30	0.31	-0.14	-0.28	-0.16
Slovakia	-0.28	0.03								
Belarus	0.03	0.05	-0.14	0.32	-0.27	-0.14	-0.14	0.18	-0.16	
Netherlands	-0.44	0.10	-0.15	0.38	-0.39	0.09	-0.25	0.18	-0.16	-0.31
Croatia	-0.18	-0.18	0.11			-0.14	-0.19	0.19	-0.00	-0.43

Table 4.4: Spearman's correlation coefficients computed from comparing the voting behaviour per country with the predicted outcome based on the Melodifestivalen. The top row shows the correlation between the actual outcome of the Eurovision final and the predicted outcome based on the Melodifestivalen.

4.2.3 Discussion

Italy

We have seen in Table 4.3's top row that the predictions for the Eurovision ranking based on the outcome of the Sanremo are not very successful. Apart from 2012, all correlation coefficients are smaller than or equal to zero. Given the results from section 4.1, this was to be expected. There, we established that using the musical features of popular songs from the Sanremo and Melodifestivalen does not bring us closer to predicting the outcome of the Eurovision. Thus, we did not expect that using only the results from the Sanremo Festival would yield better predictions.

The correlation per country in Table 4.3 first of all shows us that Italy, at least to some extent, votes according to what could have been expected given the Sanremo outcome. We find Italy in the fourth place when the countries are ranked by their medians. Two of the countries that ranked higher only participated in two editions, which might give a distorted impression of the correlation. Thus, if we were to not consider these countries, Italy would effectively rank second, only behind Montenegro. Moreover, the correlation coefficient between Italy's voting behaviour and their predicted ranking based on the Sanremo festival is positive in seven out of ten years. In five out of ten years, the Italian correlation coefficient belongs to the top six highest correlations found among all countries (namely in 2011, 2014, 2016, 2017 and 2021). Especially in 2021, the Italian voting stayed close to their Sanremo preferences, with the highest coefficient of all countries.

Furthermore, we see the results from the top row reflected in the voting behaviour of all countries. Take for example the year 2012, where we found a correlation coefficient of 0.31 between the actual and predicted ranking. In the 2012 column in Table 4.3, we see that a lot of countries show a positive correlation with the Italian prediction —many even higher than the Italian one. This implies that countries voted according to this prediction, which in turn resulted in a final outcome similar to the prognosis. On the contrary, in 2016 almost all countries show a negative correlation coefficient. They thus voted very differently from the Italian prediction, resulting in a divergent final outcome. Interestingly, even though the Italian correlation is also negative, it is one of the three highest values. This indicates that Italy still voted more 'Italian' than the other countries.

In addition to the observations about the Italian voting behaviour, we can also look at other countries. For example, it turns out that Montenegro votes the most in accordance with the outcome of the Sanremo Festival. They have the highest median and moreover six out of eight correlation coefficients are positive. On the other hand, San Marino which is enclaved by Italy, votes remarkably non-Italian, with the second lowest median and eight out of ten very low correlation coefficients.

Sweden

Also when using the musical features from the Melodifestivalen, we find variable success in predicting the Eurovision outcome. Again, this is not very surprising, given our earlier conclusions about predicting the final ranking of the international festival. The best correlation between the Swedish prediction and the actual outcome is obtained in 2014, with a correlation coefficient of 0.41. The worst correlation is found in 2021.

Table 4.4 implies that the Swedish voting behaviour in the Eurovision Song Contest is not really in accordance with the voting behaviour expected based on the outcome of the Melodifestivalen. It is positioned low in the ranking based on the median correlation and shows a positive correlation coefficient in only two out of ten years.

In all years, we find many countries that obtained a higher correlation coefficient than Sweden. That is, these countries vote more according to the Melodifestivalen-based prediction than the Swedes themselves. For example, France votes quite 'Swedish' with the second highest median and eight out of ten positive correlation coefficients. Also Israel votes according to the predictions, with the third highest median and seven out of ten positive correlations.

General Discussion

Before discussing our general interpretation of the results, we briefly note that again this study was conducted with only a 29 second fragment of the studio performances. Especially the latter aspect might have had an influence on the findings in this chapter. That is, the voting behaviour is of course influenced by the quality of the live vocals. However, as mentioned before, we believe that for this study consistency was more important.

The discussed results give an interesting representation of the voting behaviour of Italy and Sweden. While we have seen already in the previous section that it is difficult and probably impossible to predict the outcome of the Eurovision Song Contest from the musical features of high-scoring Sanremo and Melodifestivalen songs, we have now established that the Swedish national competition generally produces a more accurate prediction than the Italian one. However, the Italians vote more in line with their national competition than the Swedes do. That is, the actual voting of the Italians for the Eurovision final correlates more strongly to the prediction based on the Sanremo, than the voting of Sweden correlates to the prediction based on the Melodifestivalen.

A possible explanation for this last observation could be the interpretation of the distinct music styles in both countries, as was argued for in the previous chapter. There, we claimed that the Italian Sanremo songs have a certain 'Italianness' that differs more from the music represented at the Eurovision Song Contest than the Swedish songs from the Melodifestivalen do. This might also explain the varying correlation coefficients here. Namely, with their distinct music style, the Italians might stick to similar songs in the international contest, resulting in better predictable voting behaviour. The Swedes on the other hand already have songs similar to those in the Eurovision competing in their national contest. When voting for the Eurovision final then, there are many songs which could be considered sounding 'Swedish'. Therefore, they could choose any song as their favourite, which results in unpredictable voting behaviour and a lower correlation.

When looking at the voting behaviour of all countries compared to the predictions based on the national competitions, we have observed several countries which seem to vote 'Italian' or 'Swedish'. The median and the number of positive correlation coefficients indicate that for Italy mostly Eastern European and Caucasian countries showcase similar voting behaviour —e.g. Montenegro, Azerbaijan, Israel, Georgia. Compared to the Italian results, the Swedish top ten based on median shows more Western European countries —i.e. France, Belgium, Ireland— in addition to countries such as Israel and Armenia.

Considering the body of academic work concerning political voting in the Eurovision, we might also wonder whether our results show any signs of the alleged voting blocs. Often when these alliances are researched, there is mostly attention for political and cultural factors connecting the countries in question. However, it might well be the case that the countries' musical preferences simply align. While the analysis conducted in this section was not designed to answer these question, we might still gain some insights.

For example, it is often claimed that Scandinavian countries tend to favour each other. Even though the results presented in this section do not immediately confirm this hypothesis, we can infer that they at least seem to vote more 'Swedish' than 'Italian'. The voting behaviour of countries such as Finland, Denmark and Iceland show higher correlation with the Swedish prediction, than with the Italian one. Norway obtains correlation coefficients similar to the Swedish ones. This could suggest that apart from voting for each other, they also prefer the same acts overall. The idea that they vote for each other merely from political motivations might then be refuted by these observations.

For Italy, it is well-known that mainly Albania and Malta always give and receive many points. Our results however, do not show a clear similarity in the rest of their voting behaviour, with the majority of their correlation coefficients being negative. This might imply that in the case of this small voting bloc the favouritism is based on non-musical aspects.

As mentioned before, given that this was not our endeavour, we do not claim that these results concerning voting blocs are very strong. We merely intend to suggest that a similar approach concerning musical features might shed new light on the often debated political voting in the Eurovision.

Conclusion

In this thesis we presented a musical comparison of the Eurovision Song Contest, the Italian Festival di Sanremo and the Swedish Melodifestivalen. As far as we know, a similar study has not been done before for any national, Eurovision-qualification contest.

First, we aimed at analysing the musical features of the songs from the three contests and constructing classifiers that could distinguish between them. We found that these classifiers were quite successful at correctly classifying songs belonging to the Eurovision Song Contest. On the other hand, classifying songs from one of the national competitions turned out to be more problematic, with often only 50% correct classification. Moreover, we observed that the classifiers for the Eurovision-Melodifestivalen task had lower performance than those for the Eurovision-Sanremo task. The proposed interpretation states that these results are caused by distinct national styles. We concluded that Italy in particular exhibits its own style that often differs from the general music style at Eurovision, while the Swedish songs are more similar.

Secondly, we attempted to use the data from the national contests to predict the outcome of the Eurovision. As was expected, these predictions were not very successful. However, by comparing the voting behaviour of different countries to the predicted outcome of the Eurovision based on one of the national contests, we showed that Italy votes more according to their Sanremo preferences than Sweden votes according to their Melodifestivalen preferences. This supports our earlier interpretation of the different styles. Namely, it seems the Italians have a more distinct national style and are inclined to stick with these preferences also in voting for the Eurovision. The Swedes, on the other hand, have a music taste more similar to the general Eurovision sound and therefore their voting behaviour is less predictable.

This thesis contributes a first musical analysis of the Eurovision Song Contest and two similar national contests. It shows that musical features can be used for a classification task distinguishing the songs from the different competitions. Moreover, we proposed an interpretation of distinct national music styles, at least within the considered festivals. In particular, we claim that Italy has a style that is unique for the Eurovision Song Contest. This explanation was supported by both the classification and prediction results.

For future research, it would first of all be interesting to further explore the Eurovision in terms of its music. Until now, no study on the international festival had analysed this. However, we have shown that its songs can provide first insights, for example in national music styles. In future studies this path could be extended by including more different countries, or one could focus on a different aspect of the music entirely.

In addition, it would be of interest to analyse the voting blocs in terms of musical preferences. Although the political voting has been researched already many times, the emphasis of these studies often lies on geographical, political or cultural factors. An analysis of the music might add a new dimension to the debate.

Finally, further research could focus on the national music preferences. Our results indicate a distinct Italian style and it would be interesting to delve into this deeper. For example, do we find these preferences also beyond the Eurovision, e.g. in radio or Spotify charts? And what about other countries, do they show specific music styles? It would be interesting to analyse these questions in search of a musical map of Europe.

Bibliography

[1] AcousticBrainz. `https://acousticbrainz.org/`. Accessed 3-6-2022.

[2] G. G. Amore. Il Festival di Sanremo come prodotto culturale. Bachelor's thesis, Luiss Guido Carli, 2021.

[3] M. Ardizzoni. On Rhythms and Rhymes: Poetics of Identity in Postcolonial Italy. *Communication, Culture & Critique*, 13(1):1–16, 2020.

[4] C. Baker. The 'gay Olympics'? The Eurovision song contest and the politics of LGBT/European belonging. *European Journal of International Relations*, 23(1): 97–121, 2017.

[5] F. Bimbot, G. Sargent, E. Deruty, C. Guichaoua, and E. Vincent. Semiotic description of music structure: An introduction to the Quaero/Metiss structural annotations. In *AES 53rd International Conference on Semantic Audio*, pages P1–1, 2014.

[6] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, et al. Essentia: An Audio Analysis Library for Music Information Retrieval. In *14th Conference of the International Society for Music Information Retrieval (ISMIR'13)*, pages 493–498, 2013.

[7] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[8] O. Budzinski and J. Pannicke. Culturally biased voting in the Eurovision Song Contest: Do national contests differ? *Journal of Cultural Economics*, 41(4): 343–378, 2017.

[9] C. J. Burges. From RankNet to LambdaRank to LambdaMART: An Overview. *Learning*, 11(23-581):81, 2010.

[10] J. Carniel. Nation Branding, Cultural Relations and Cultural Diplomacy at Eurovision: Between Australia and Europe. In *Eurovisions: Identity and the International Politics of the Eurovision Song Contest since 1956*, pages 151–173. Springer, 2019.

[11] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[12] D. Chicco and G. Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[13] D. Chicco, M. J. Warrens, and G. Jurman. The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, 9:78368–78381, 2021.

[14] M. Corciolani. Il marketing dell'autenticità in condizioni critiche: il caso degli Afterhours al Festival di Sanremo. *Il marketing dell'autenticità in condizioni critiche: il caso degli Afterhours al Festival di Sanremo*, pages 54–79, 2010.

[15] W. B. de Bruin. Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta psychologica*, 118(3):245–260, 2005.

[16] R. Delgado and X.-A. Tibau. Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one*, 14(9):e0222916, 2019.

[17] R. Dunin-Wasowicz. The reaction to Conchita Wurst's victory at Eurovision highlights the polarisation over LGBTI rights across Europe. *LSE European Politics and Policy (EUROPP) Blog*, 2014.

[18] S. D'Angelo, T. B. Murphy, and M. Alfò. Latent space modelling of multidimensional networks with application to the exchange of votes in eurovision song contest. *The annals of applied statistics*, 13(2):900–930, 2019.

[19] M. Eneberg and J. Scortea. Brand new image: a case study of engagement on Instagram during Melodifestivalen in Kristianstad. Bachelor's thesis, Högskolan Kristianstad, 2018.

[20] European Broadcasting Union. `https://eurovision.tv`, 2002-2022. Accessed 21-2-2022.

[21] European Broadcasting Union. National Selections. `https://eurovision.tv/about/in-depth/national-selections/`, 2002-2022. Accessed 21-2-2022.

[22] European Broadcasting Union. `https://eurovision.tv/events`, 2002-2022. Accessed 13-5-2022.

[23] European Broadcasting Union. 183 million viewers welcome back the Eurovision Song Contest. `https://eurovision.tv/story/183-million-viewers-welcome-back-the-eurovision-song-contest`, 2021. Accessed 21-2-2022.

[24] Eurovisionworld. `https://eurovisionworld.com/eurovision`. Accessed 13-5-2022.

[25] S. Facci, P. Soddu, and M. Piloni. *Il Festival di Sanremo. Parole e suoni raccontano la nazione.* Carocci, 2011.

[26] M. Flecht. Eurovision Song Contest Database. `https://eschome.net`, 2021. Accessed 8-2-2022.

[27] S. O. Folorunso, S. A. Afolabi, and A. B. Owodeyi. Dissecting the genre of Nigerian music with machine learning models. *Journal of King Saud University-Computer and Information Sciences*, 2021.

[28] M. A. Haan, S. G. Dijkstra, and P. T. Dijkstra. Expert judgment versus public opinion–evidence from the Eurovision song contest. *Journal of Cultural Economics*, 29(1):59–78, 2005.

[29] J. Halliwell. 'All Kinds of Everything'? Queer Visibility in Online and Offline Eurovision Fandom. *Westminster Papers in Communication and Culture*, 13(2), 2018.

[30] T. Hawlin. Take that, Salvini! How musical justice was served in Sanremo. `https://www.theguardian.com/commentisfree/2019/feb/13/salvini-music-sanremo-music-italian-mahmoud-eurovision`, 2019. Accessed 24-03-2022.

[31] S. Holmdahl. "Seriöst, det känns som att du har glömt genusanalysen?": En kulturstudie av femininitet och respektabilitet i Melodifestivalen. Master's thesis, Stockholms Universitet, 2018.

[32] P. Jordan. *The modern fairy tale: Nation branding, national identity and the Eurovision Song Contest in Estonia*. University of Tartu Press, 2014.

[33] A. Koski and J. Persson. And the winner is...: Predicting the outcome of melodifestivalen by analyzing the sentiment value of tweets. Bachelor's thesis, KTH Royal Institute of Technology, 2017.

[34] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3): 159–190, 2006.

[35] M. Kyriakidou, M. Skey, J. Uldam, and P. McCurdy. Media events and cosmopolitan fandom: 'Playful nationalism' in the Eurovision Song Contest. *International Journal of Cultural Studies*, 21(6):603–618, 2018.

[36] S. Lee, H. Jeong, and H. Ko. Classical Music Specific Mood Automatic Recognition Model Proposal. *Electronics*, 10(20):2489, 2021.

[37] J. Lindgren. Melodifestivalen: Tur Eller Ren Skicklighet. Bachelor's thesis, Uppsala Universitet, 2016.

[38] P. Magron and C. Févotte. Leveraging the structure of musical preference in content-aware music recommendation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585, 2021.

[39] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[40] A. Meijer. Be my guest: Nation branding and national representation in the Eurovision Song Contest. Master's thesis, Uppsala Universitet and Universiteit Groningen, 2013.

[41] D. Moffat, D. Ronan, and J. D. Reiss. An evaluation of audio feature extraction toolboxes. In *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[42] R. Rehman, G. C. Hazarika, and D. Kardong. Auditory Scale Analysis and Evaluation of Phonemes in MISING Language. *International Journal of Computer Applications*, 113(15), 2015.

[43] J. Salamon, J. Serra, and E. Gómez. Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, 2(1):45–58, 2013.

[44] M. Schedl, E. Gómez, and J. Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2-3):127–261, 2014.

[45] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan. An auditory-based feature for robust speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4625–4628, 2009.

[46] Showredactie AD. Duncans Arcade verbreekt record op Spotify. `https://www.ad.nl/show/duncans-arcade-verbreekt-record-op-spotify~a238e56e/`, 2019. Accessed 10-5-2022.

[47] L. Spierdijk and M. Vellekoop. The structure of bias in peer voting systems: lessons from the Eurovision Song Contest. *Empirical Economics*, 36(2):403–425, 2009.

[48] J. Spijkervet. The Eurovision Dataset. `https://zenodo.org/badge/latestdoi/214236225`, 2020. Accessed 8-2-2022.

[49] G. Yair. Douze point: Eurovisions and Euro-Divisions in the Eurovision Song Contest – Review of two decades of research. *European Journal of Cultural Studies*, 22(5-6):1013–1029, 2019.

[50] X. Zhao and D. Wang. Analyzing noise robustness of MFCC and GFCC features in speaker identification. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7204–7208, 2013.

Appendix A

A.1 Musical Features

We include the list of all musical features that were used for the classification tasks. If multiple statistics were available, we only mention the mean value. That is, for any feature name ending with `.mean`, we also computed the standard deviation (`.stdev`), variance (`.var`), minimum (`.min`), maximum (`.max`), median (`.median`), mean of the first derivative (`.dmean`), mean of the second derivative (`.dmean2`), variance of the first derivative (`.dvar`) and variance of the second derivative (`.dvar2`).

```
lowlevel.average_loudness
lowlevel.dissonance.mean
lowlevel.dynamic_complexity
lowlevel.erbbands_crest.mean
lowlevel.erbbands_flatness_db.mean
lowlevel.erbbands_kurtosis.mean
lowlevel.erbbands_skewness.mean
lowlevel.erbbands_spread.mean
lowlevel.gfcc.mean_0
lowlevel.gfcc.mean_1
lowlevel.gfcc.mean_2
lowlevel.gfcc.mean_3
lowlevel.gfcc.mean_4
lowlevel.gfcc.mean_5
lowlevel.gfcc.mean_6
lowlevel.gfcc.mean_7
lowlevel.gfcc.mean_8
lowlevel.gfcc.mean_9
lowlevel.gfcc.mean_10
lowlevel.gfcc.mean_11
lowlevel.gfcc.mean_12
lowlevel.hfc.mean
lowlevel.loudness_ebu128.integrated
lowlevel.loudness_ebu128.loudness_range
lowlevel.loudness_ebu128.momentary.mean
lowlevel.loudness_ebu128.short_term.mean
lowlevel.pitch_salience.mean
lowlevel.silence_rate_20dB.mean
lowlevel.silence_rate_30dB.mean
lowlevel.silence_rate_60dB.mean
lowlevel.spectral_centroid.mean
lowlevel.spectral_complexity.mean
lowlevel.spectral_decrease.mean
lowlevel.spectral_energy.mean
lowlevel.spectral_energyband_high.mean
lowlevel.spectral_energyband_low.mean
lowlevel.spectral_energyband_middle_high.mean
lowlevel.spectral_energyband_middle_low.mean
lowlevel.spectral_entropy.mean
lowlevel.spectral_flux.mean
lowlevel.spectral_kurtosis.mean
lowlevel.spectral_rms.mean
lowlevel.spectral_rolloff.mean
lowlevel.spectral_skewness.mean
lowlevel.spectral_spread.mean
lowlevel.spectral_strongpeak.mean
lowlevel.zerocrossingrate.mean
rhythm.beats_count
rhythm.beats_loudness.mean
rhythm.bpm
rhythm.bpm_histogram_first_peak_bpm
rhythm.bpm_histogram_first_peak_weight
rhythm.bpm_histogram_second_peak_bpm
rhythm.bpm_histogram_second_peak_spread
rhythm.bpm_histogram_second_peak_weight
rhythm.danceability
rhythm.onset_rate
tonal.chords_changes_rate
tonal.chords_key_A
tonal.chords_key_Ab
tonal.chords_key_B
tonal.chords_key_Bb
tonal.chords_key_C
tonal.chords_key_C#
tonal.chords_key_D
tonal.chords_key_E
tonal.chords_key_Eb
tonal.chords_key_F
tonal.chords_key_F#
tonal.chords_key_G
tonal.chords_number_rate
tonal.chords_scale_major
tonal.chords_scale_minor
tonal.chords_strength.mean
tonal.hpcp_crest.mean
tonal.hpcp_entropy.mean
tonal.key_edma.key_A
tonal.key_edma.key_Ab
tonal.key_edma.key_B
tonal.key_edma.key_Bb
tonal.key_edma.key_C
tonal.key_edma.key_C#
tonal.key_edma.key_D
tonal.key_edma.key_E
tonal.key_edma.key_Eb
tonal.key_edma.key_F
tonal.key_edma.key_F#
tonal.key_edma.key_G
tonal.key_edma.scale_major
tonal.key_edma.scale_minor
tonal.key_edma.strength
tonal.key_krumhansl.key_A
tonal.key_krumhansl.key_Ab
tonal.key_krumhansl.key_B
tonal.key_krumhansl.key_Bb
tonal.key_krumhansl.key_C
tonal.key_krumhansl.key_C#
tonal.key_krumhansl.key_D
tonal.key_krumhansl.key_E
tonal.key_krumhansl.key_Eb
tonal.key_krumhansl.key_F
tonal.key_krumhansl.key_F#
tonal.key_krumhansl.key_G
tonal.key_krumhansl.scale_major
tonal.key_krumhansl.scale_minor
tonal.key_krumhansl.strength
tonal.key_temperley.key_A
tonal.key_temperley.key_Ab
tonal.key_temperley.key_B
tonal.key_temperley.key_Bb
tonal.key_temperley.key_C
tonal.key_temperley.key_C#
tonal.key_temperley.key_D
tonal.key_temperley.key_E
tonal.key_temperley.key_Eb
tonal.key_temperley.key_F
tonal.key_temperley.key_F#
tonal.key_temperley.key_G
tonal.key_temperley.scale_major
tonal.key_temperley.scale_minor
tonal.key_temperley.strength
tonal.thpcp_0
tonal.thpcp_1
tonal.thpcp_2
tonal.thpcp_3
tonal.thpcp_4
tonal.thpcp_5
tonal.thpcp_6
tonal.thpcp_7
tonal.thpcp_8
tonal.thpcp_9
tonal.thpcp_10
tonal.thpcp_11
tonal.thpcp_12
tonal.thpcp_13
tonal.thpcp_14
tonal.thpcp_15
tonal.thpcp_16
tonal.thpcp_17
tonal.thpcp_18
tonal.thpcp_19
tonal.thpcp_20
tonal.thpcp_21
tonal.thpcp_22
tonal.thpcp_23
tonal.thpcp_24
tonal.thpcp_25
tonal.thpcp_26
tonal.thpcp_27
tonal.thpcp_28
tonal.thpcp_29
tonal.thpcp_30
tonal.thpcp_31
tonal.thpcp_32
tonal.thpcp_33
tonal.thpcp_34
tonal.thpcp_35
tonal.tuning_diatonic_strength
tonal.tuning_equal_tempered_deviation
tonal.tuning_frequency
tonal.tuning_nontempered_energy_ratio
```

A.2 Italy

A.2.1 Importance plots

Here we include all importance plots generated by the Eurovision-Sanremo classifiers.

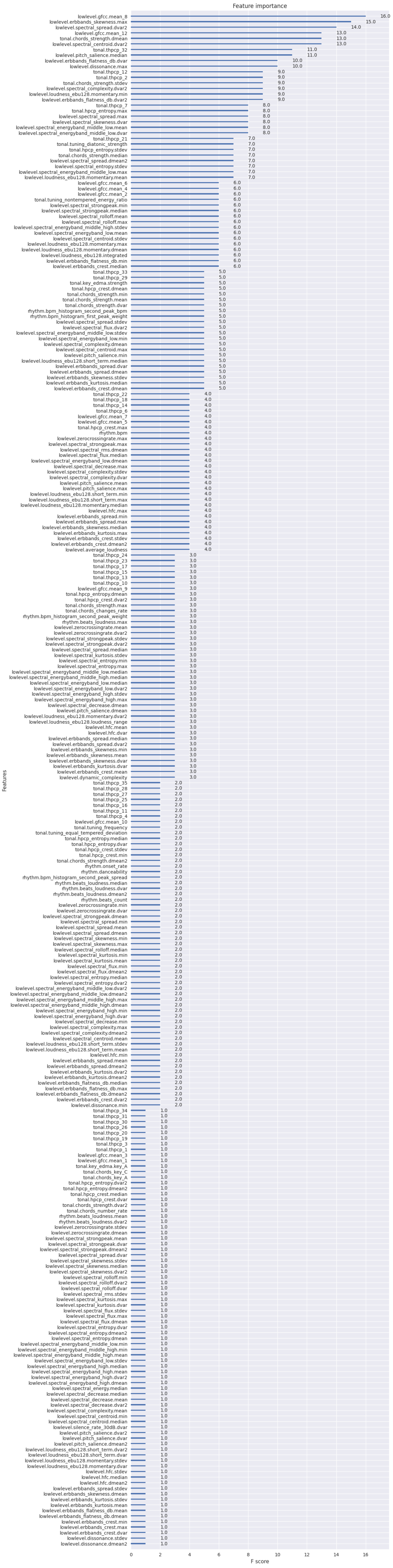


Figure A.1: Importance plot general ESC-SR classifier.

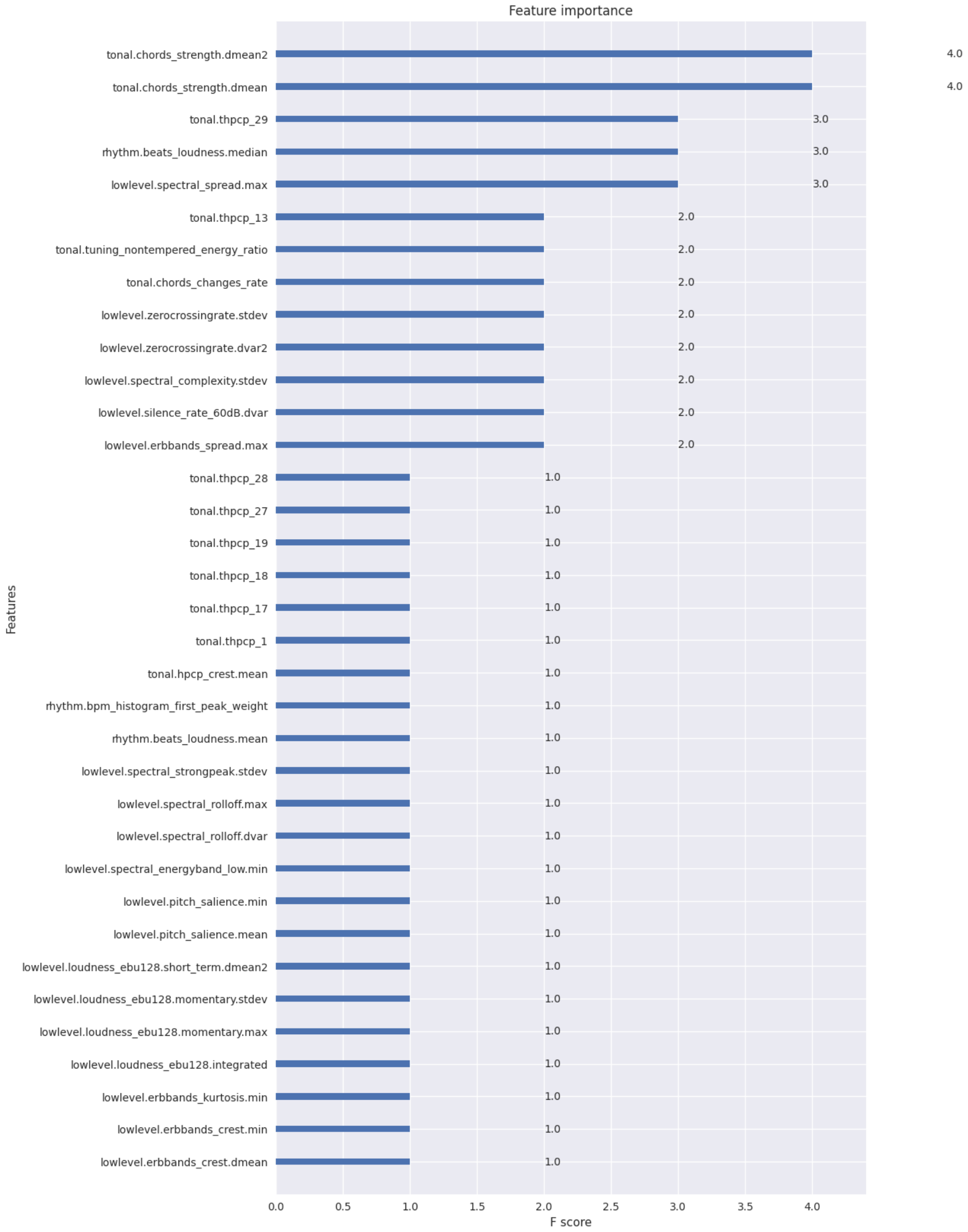


Figure A.2: Importance plot yearly ESC-SR classifier 2011.

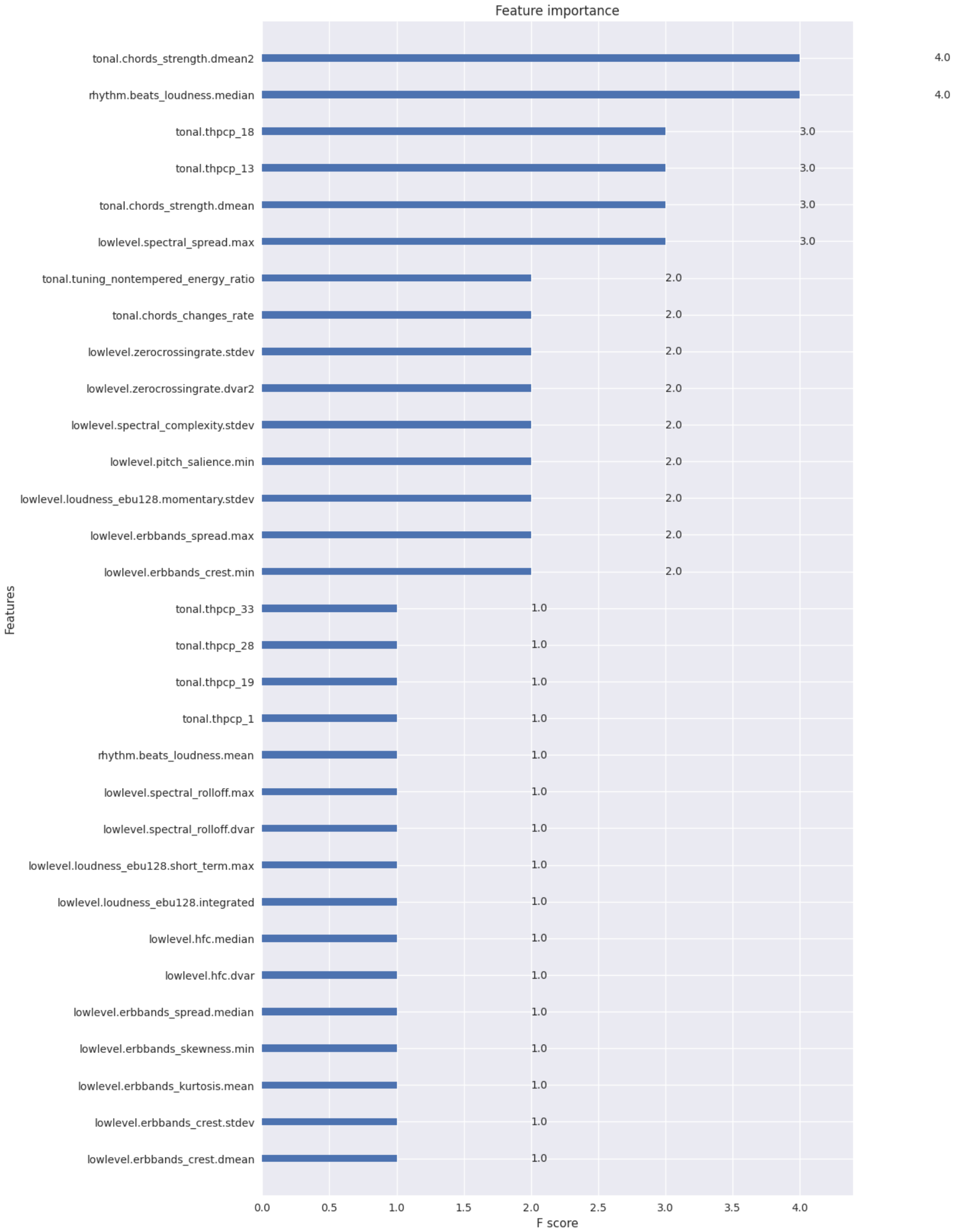


Figure A.3: Importance plot selection ESC-SR classifier 2011.

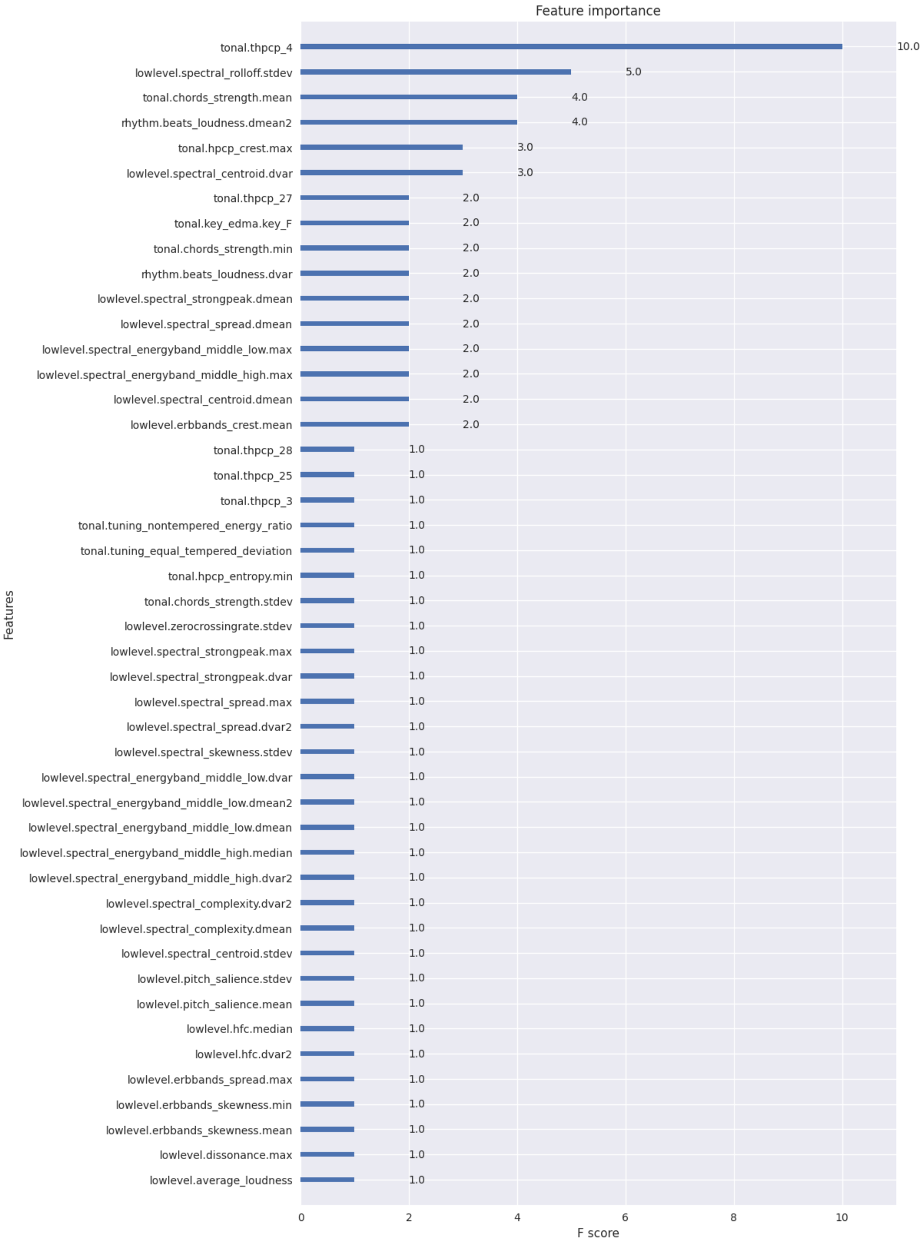


Figure A.4: Importance plot yearly ESC-SR classifier 2012.

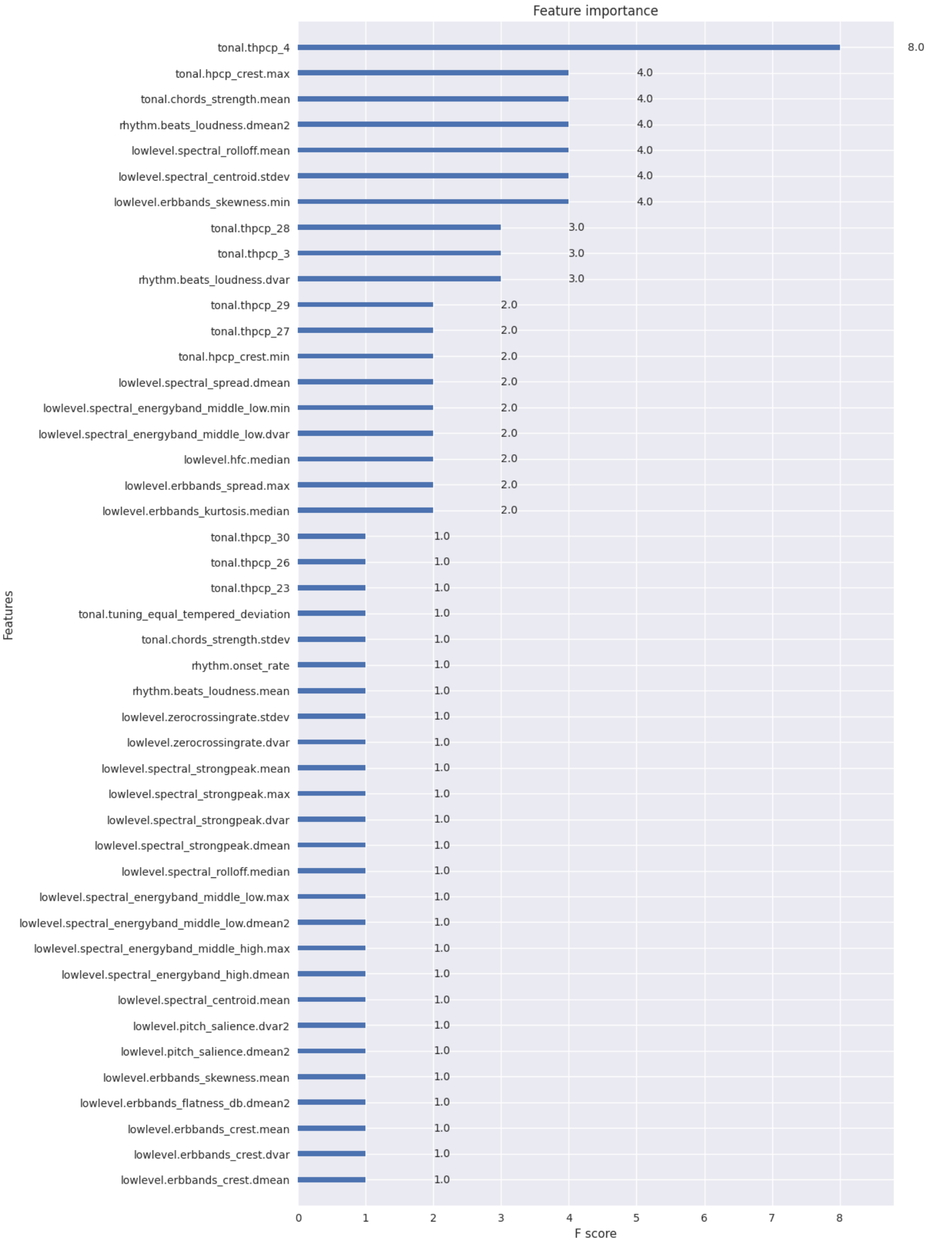


Figure A.5: Importance plot selection ESC-SR classifier 2012.

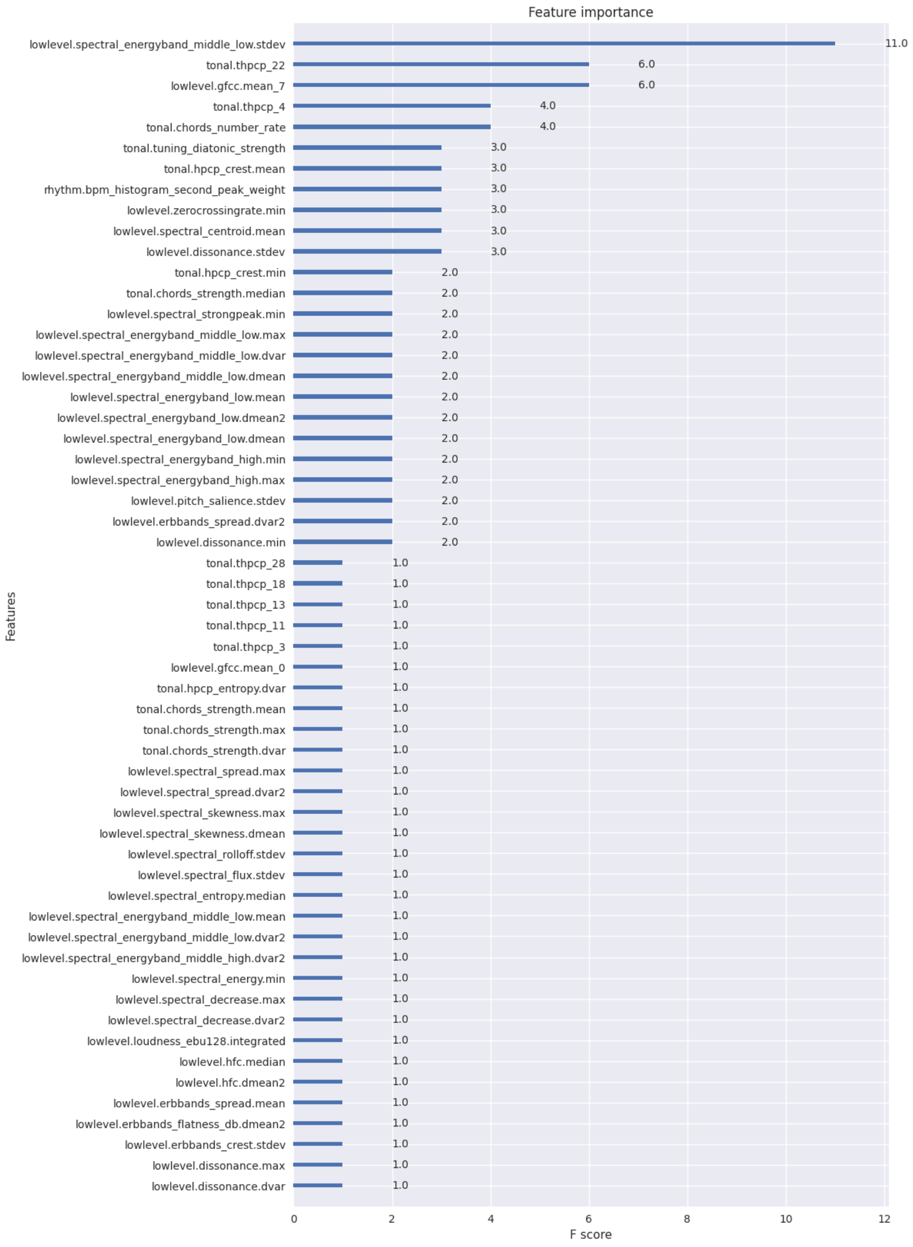


Figure A.6: Importance plot yearly ESC-SR classifier 2013.

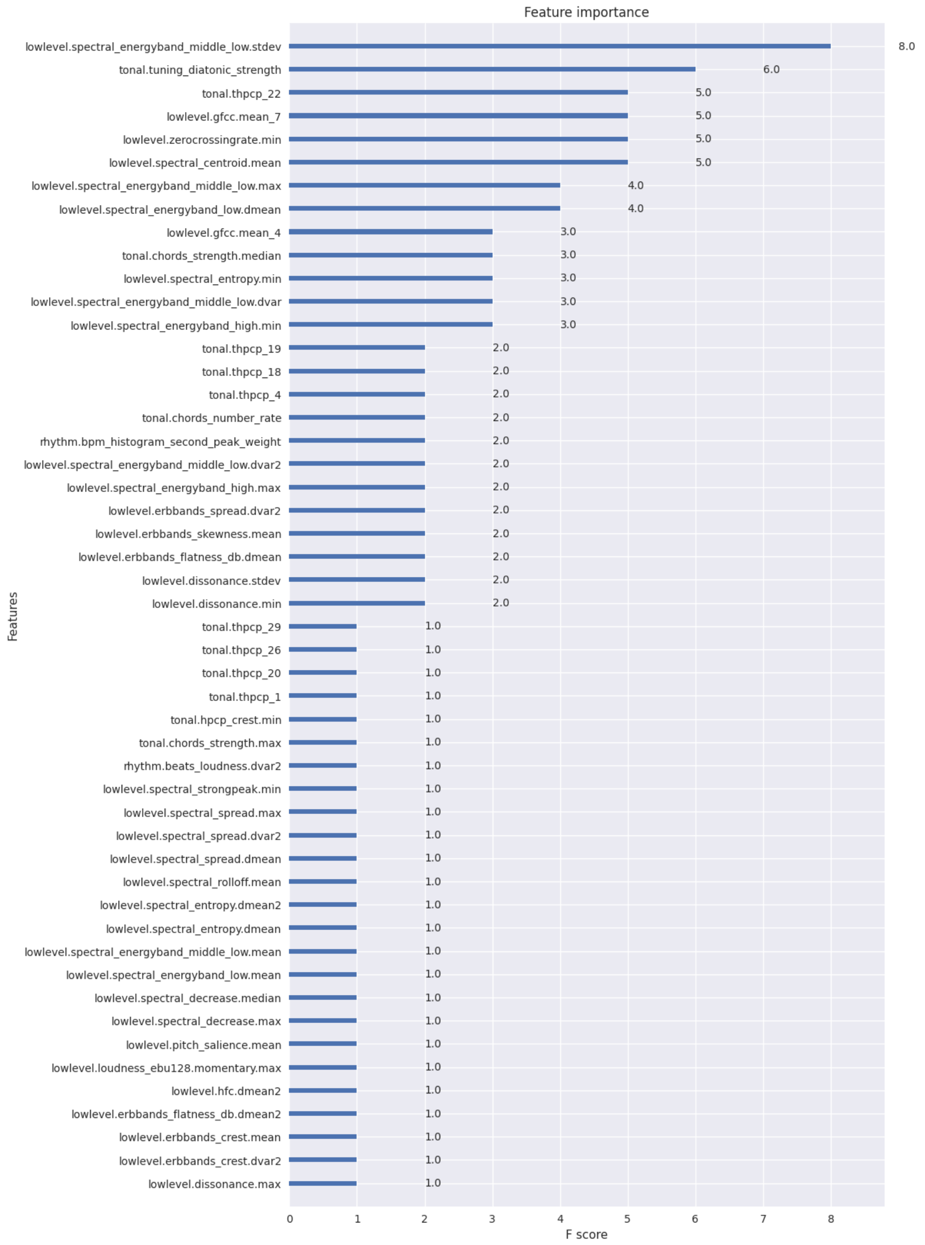


Figure A.7: Importance plot selection ESC-SR classifier 2013.

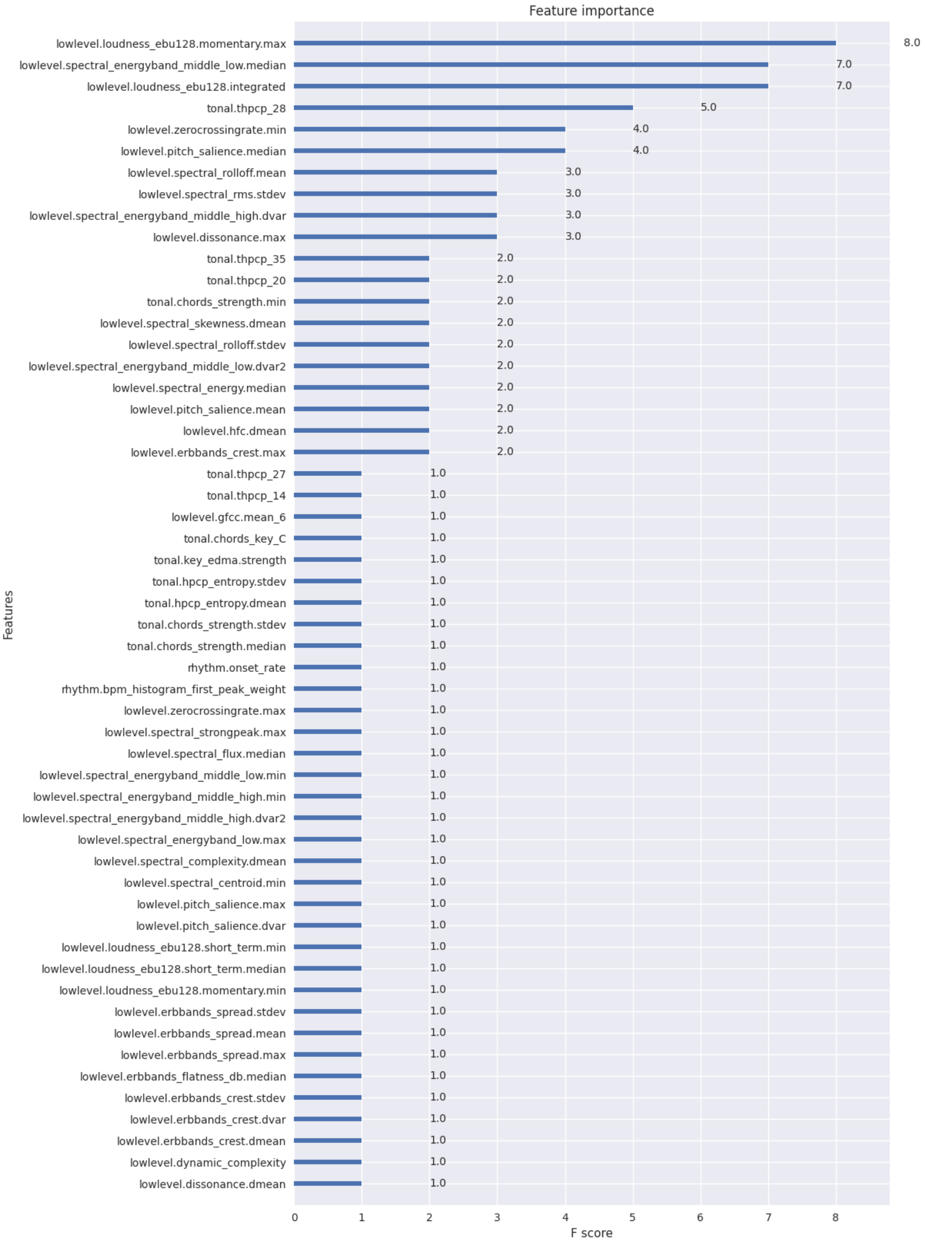


Figure A.8: Importance plot yearly ESC-SR classifier 2014.

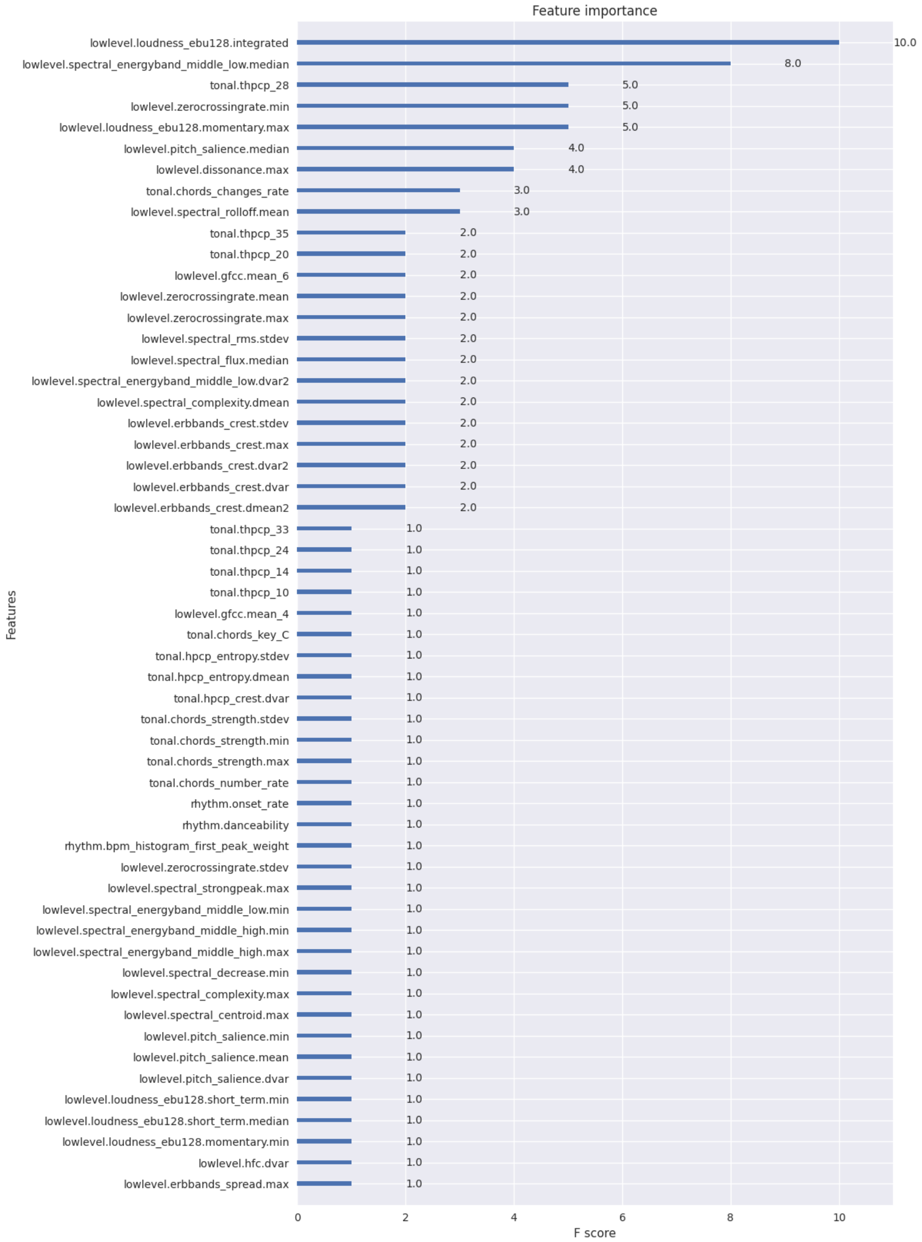


Figure A.9: Importance plot selection ESC-SR classifier 2014.

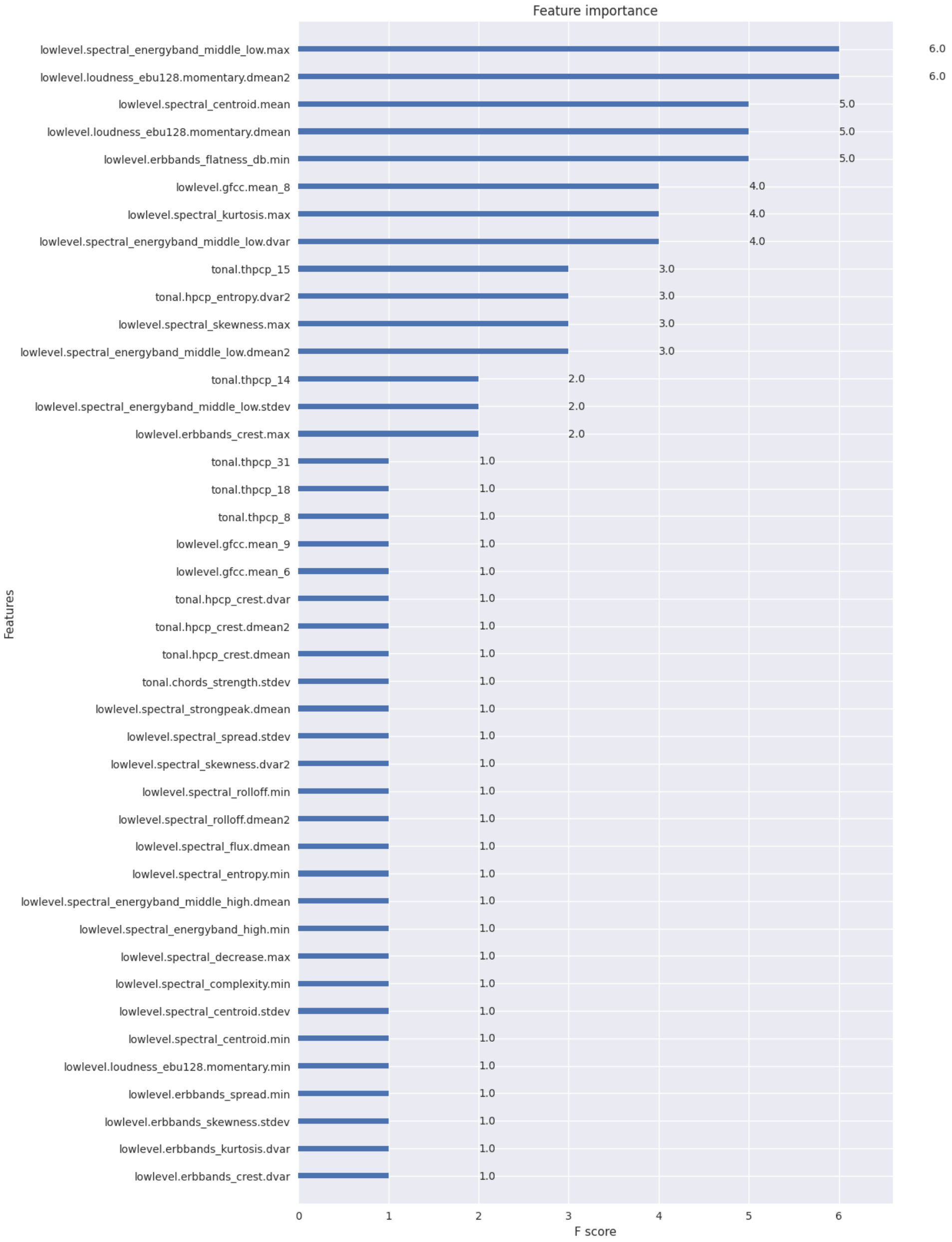


Figure A.10: Importance plot yearly ESC-SR classifier 2015.

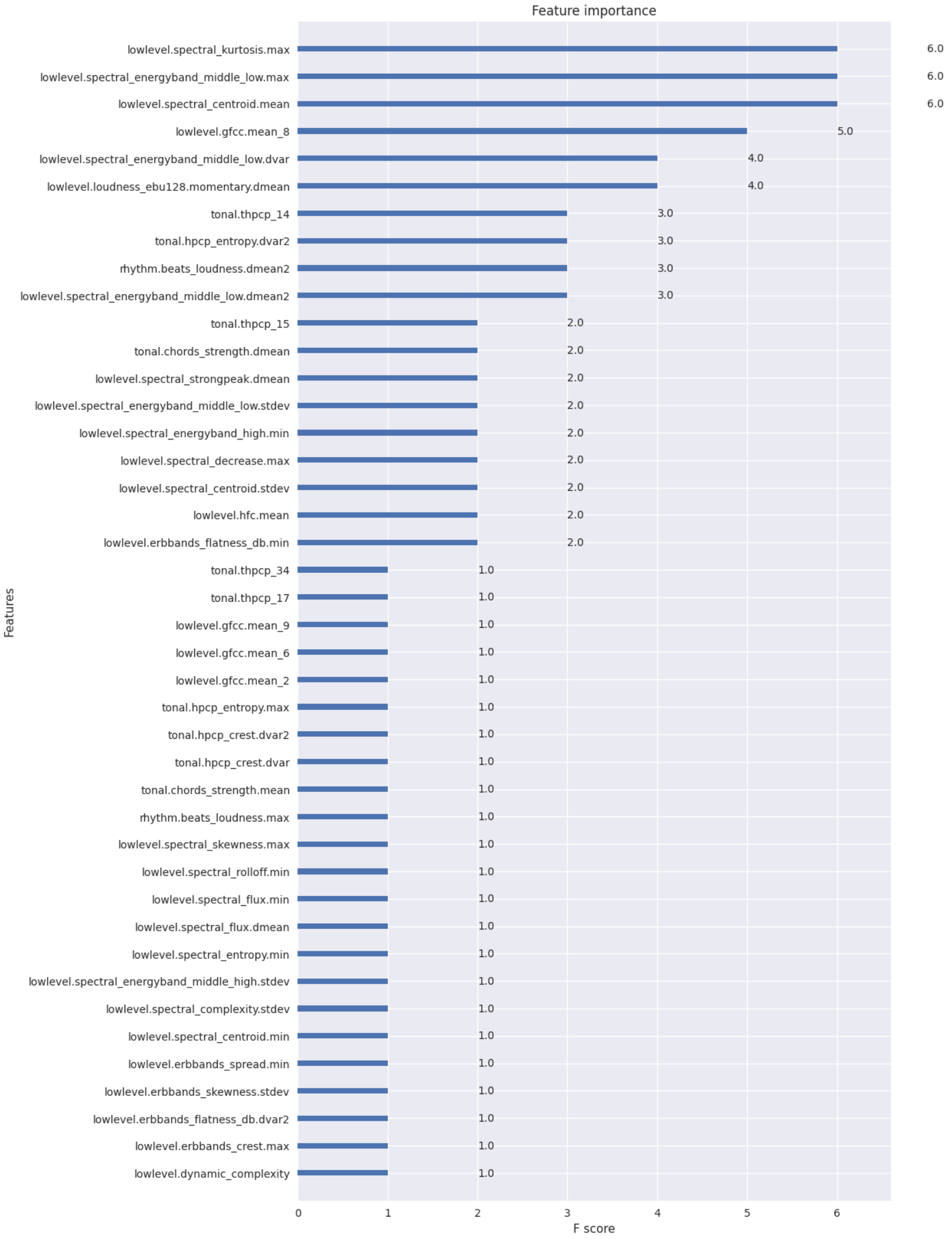


Figure A.11: Importance plot selection ESC-SR classifier 2015.

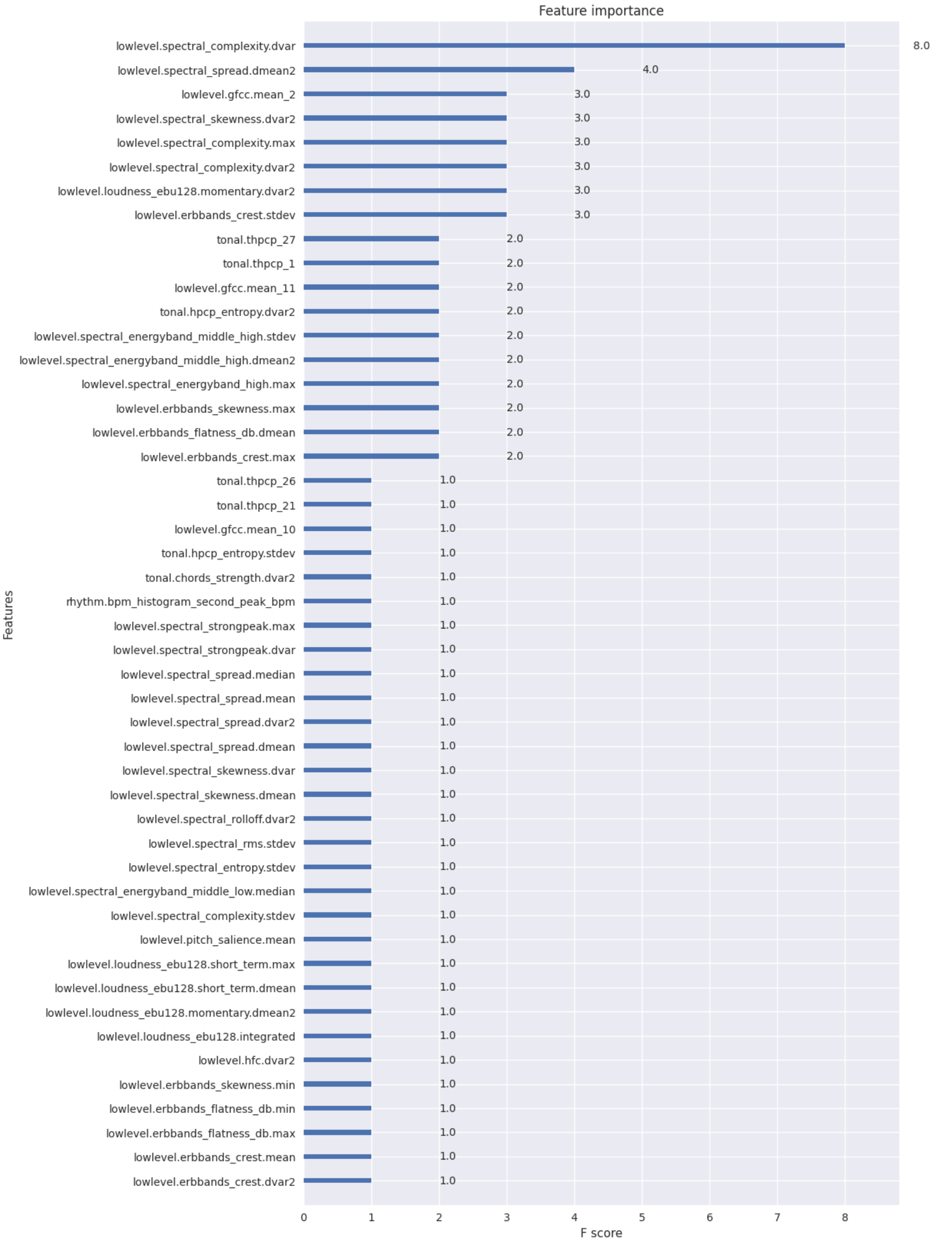


Figure A.12: Importance plot yearly ESC-SR classifier 2016.

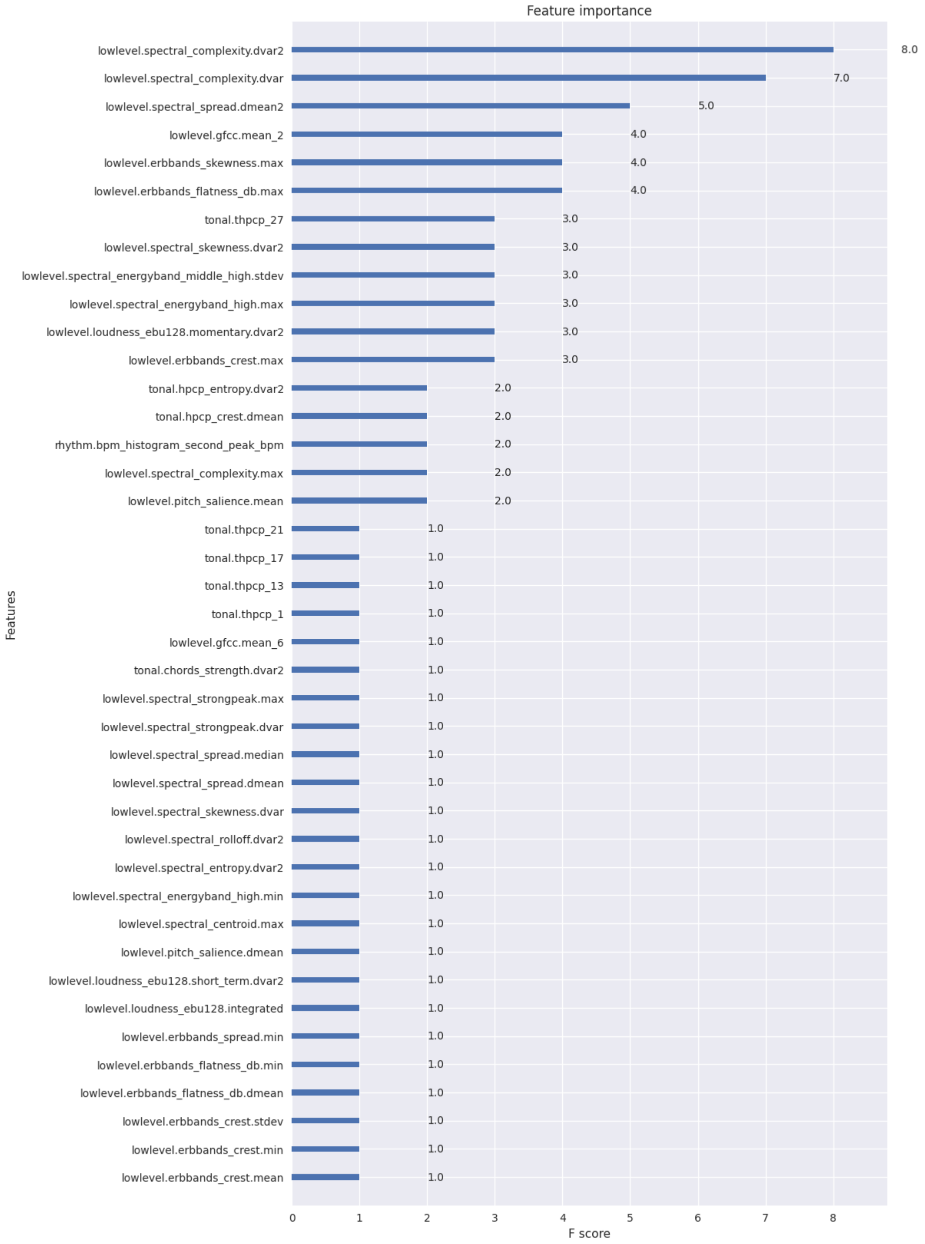


Figure A.13: Importance plot selection ESC-SR classifier 2016.

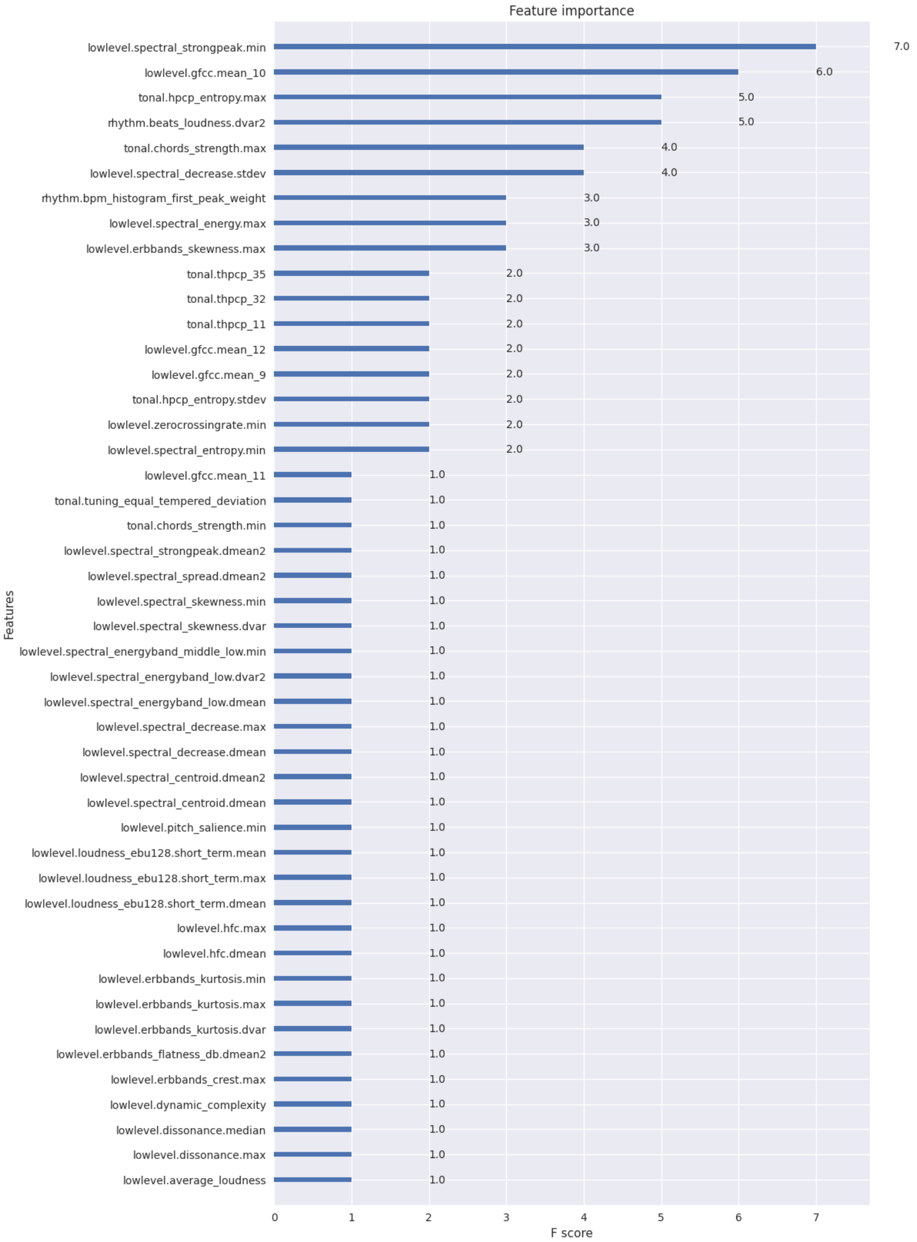


Figure A.14: Importance plot yearly ESC-SR classifier 2017.

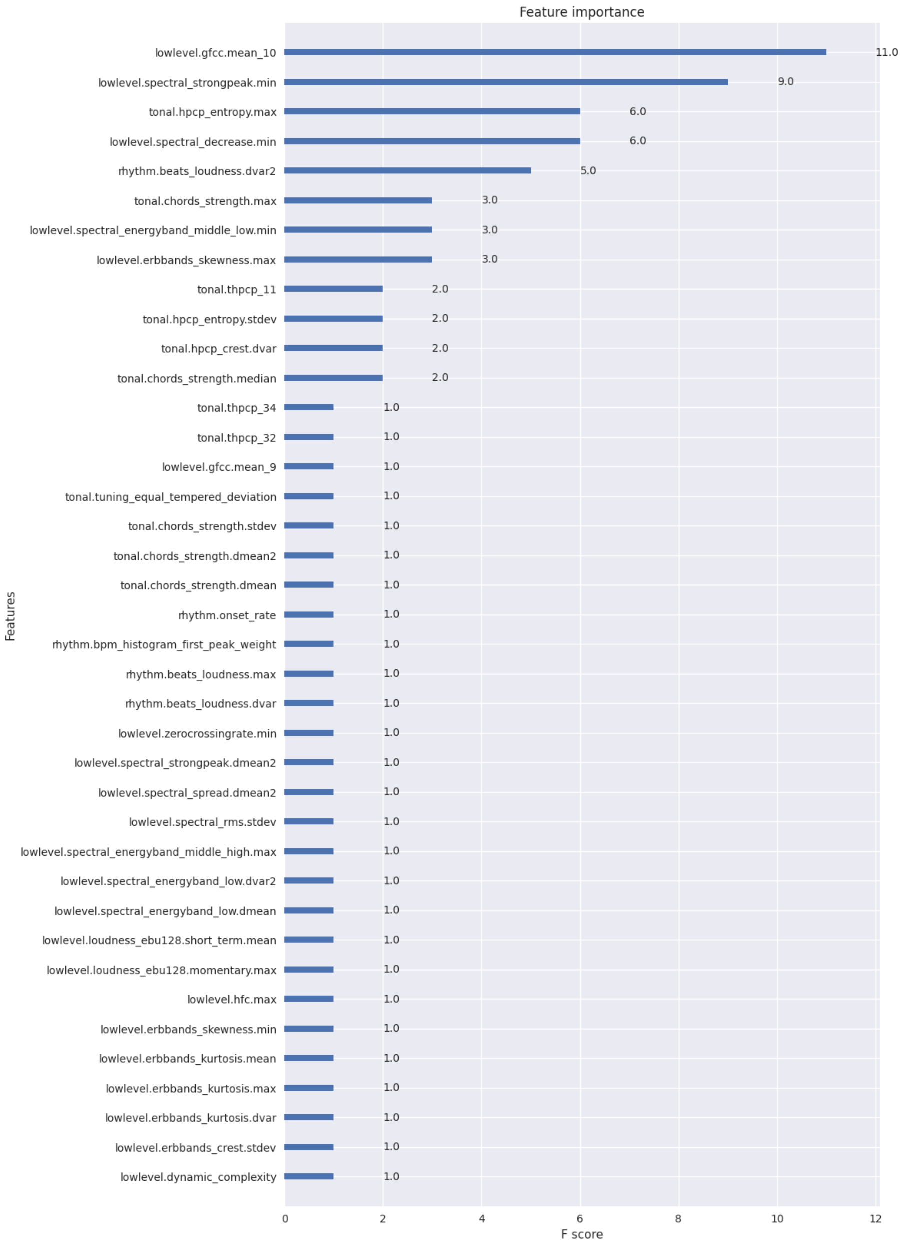


Figure A.15: Importance plot selection ESC-SR classifier 2017.

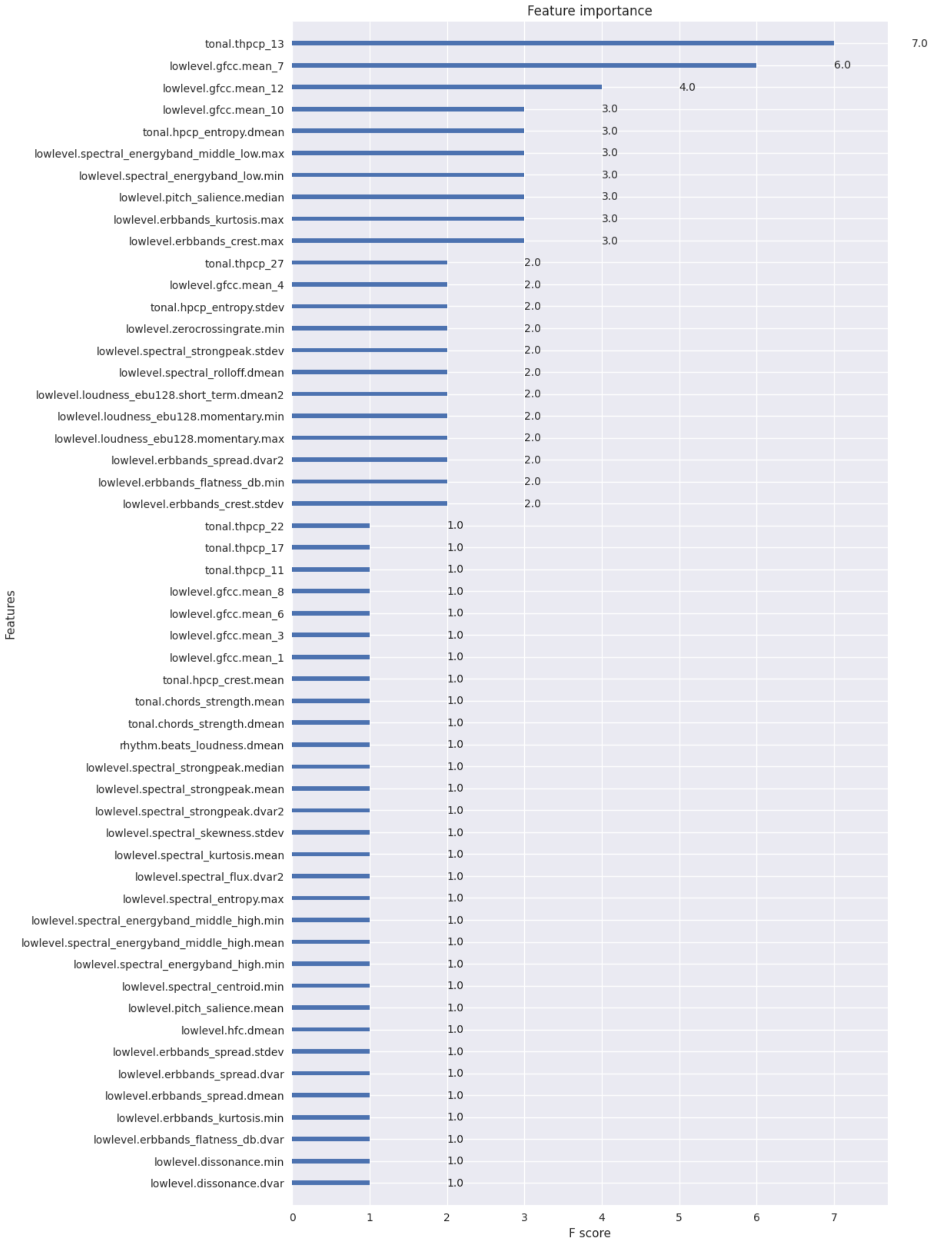


Figure A.16: Importance plot yearly ESC-SR classifier 2018.

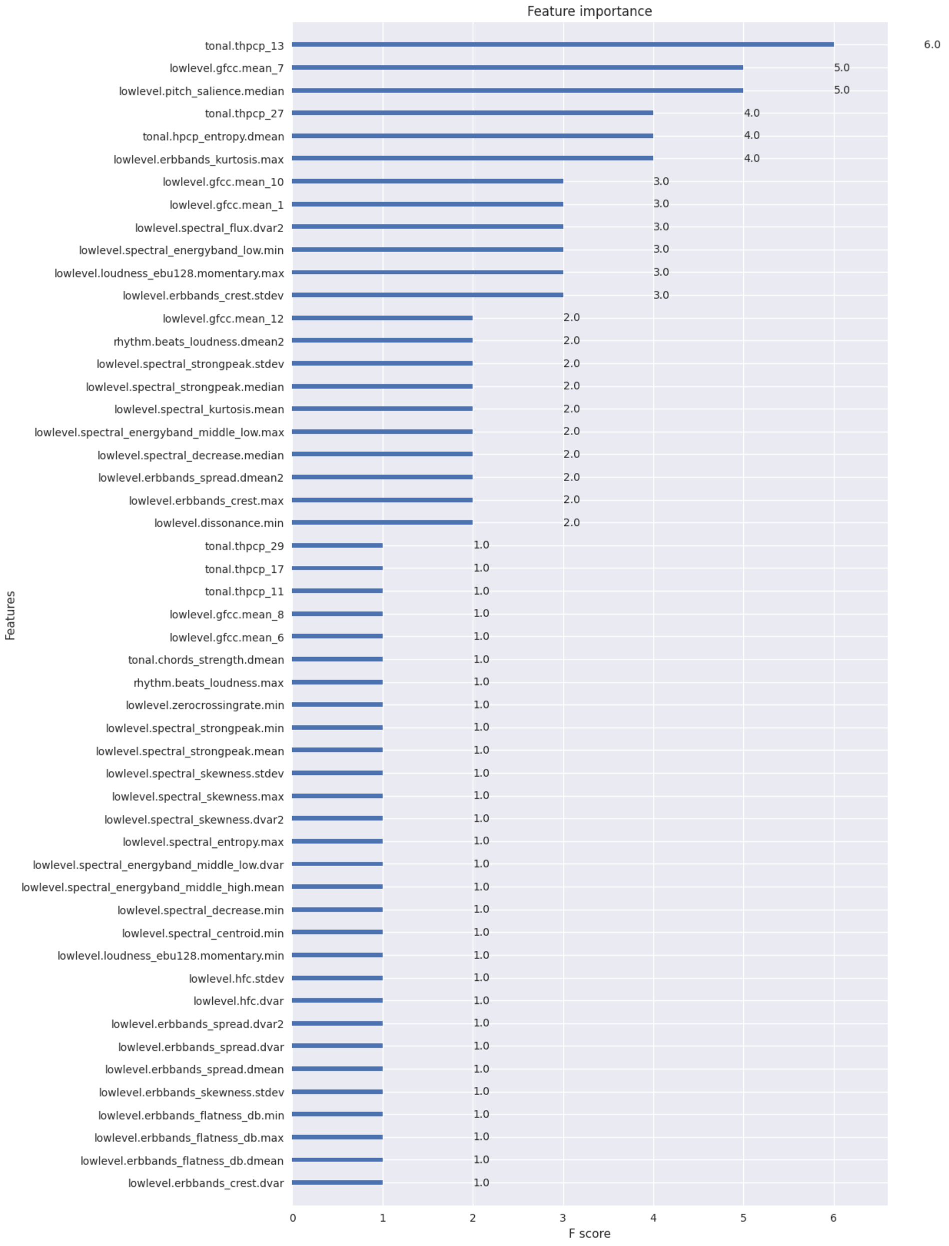


Figure A.17: Importance plot selection ESC-SR classifier 2018.

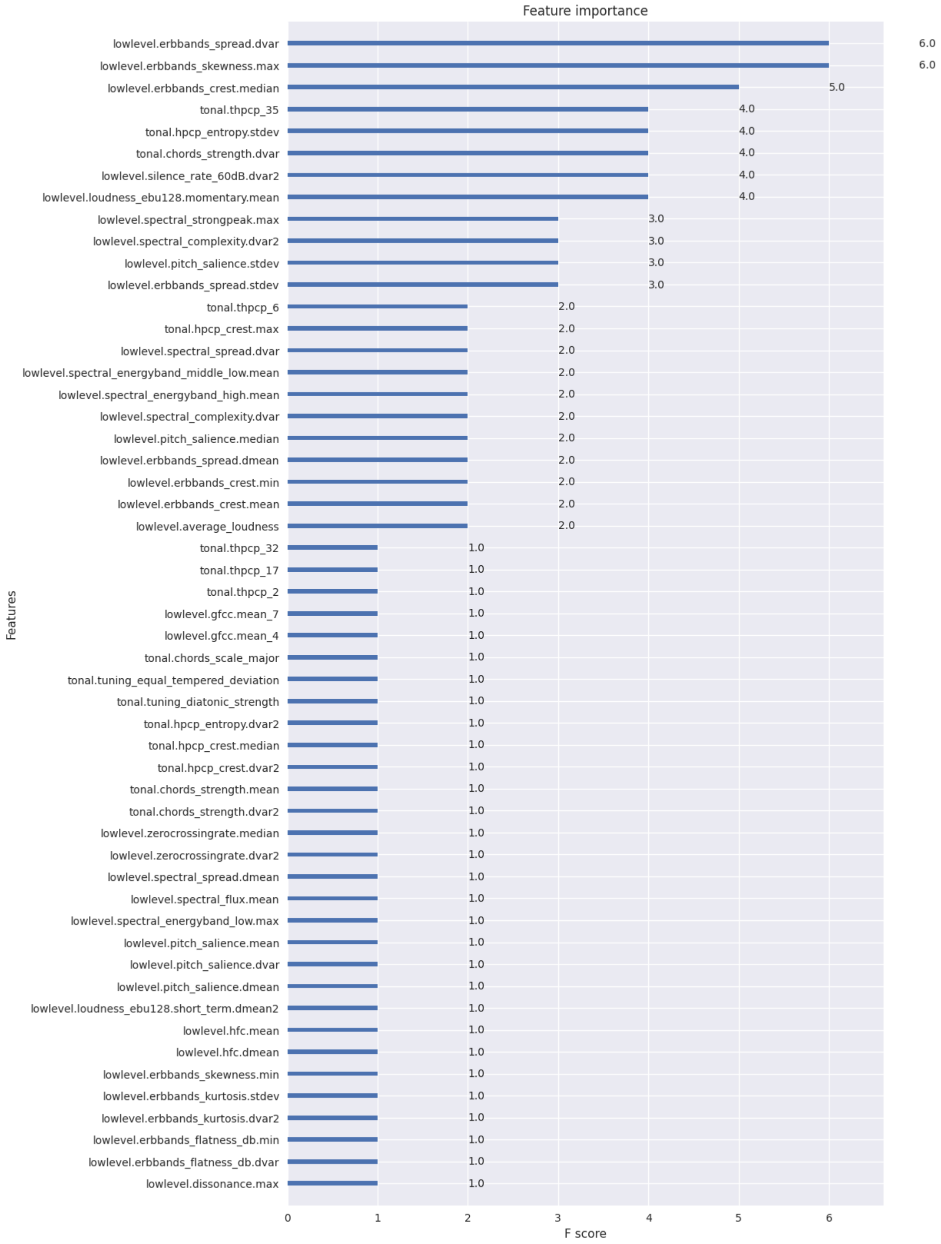


Figure A.18: Importance plot yearly ESC-SR classifier 2019.

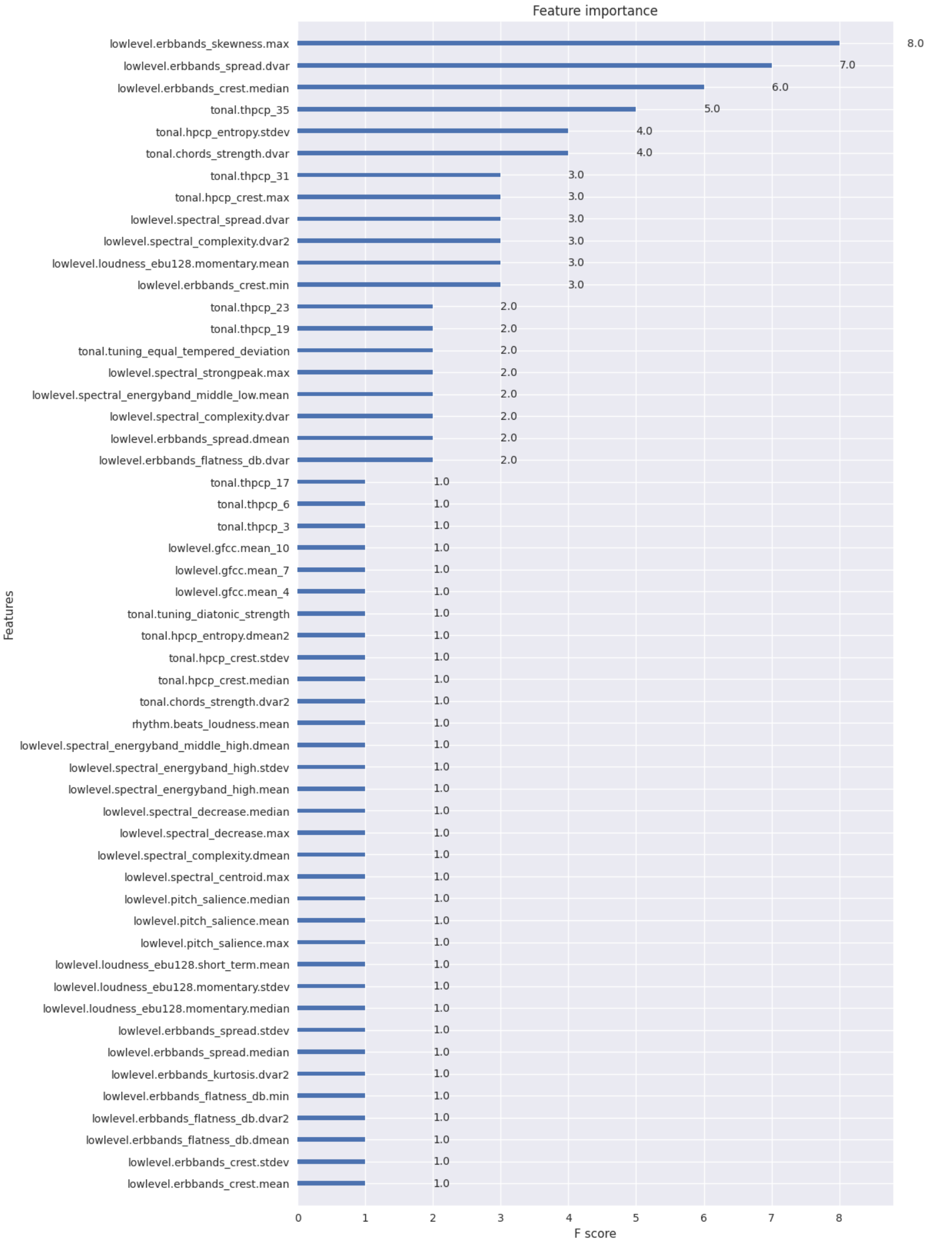


Figure A.19: Importance plot selection ESC-SR classifier 2019.

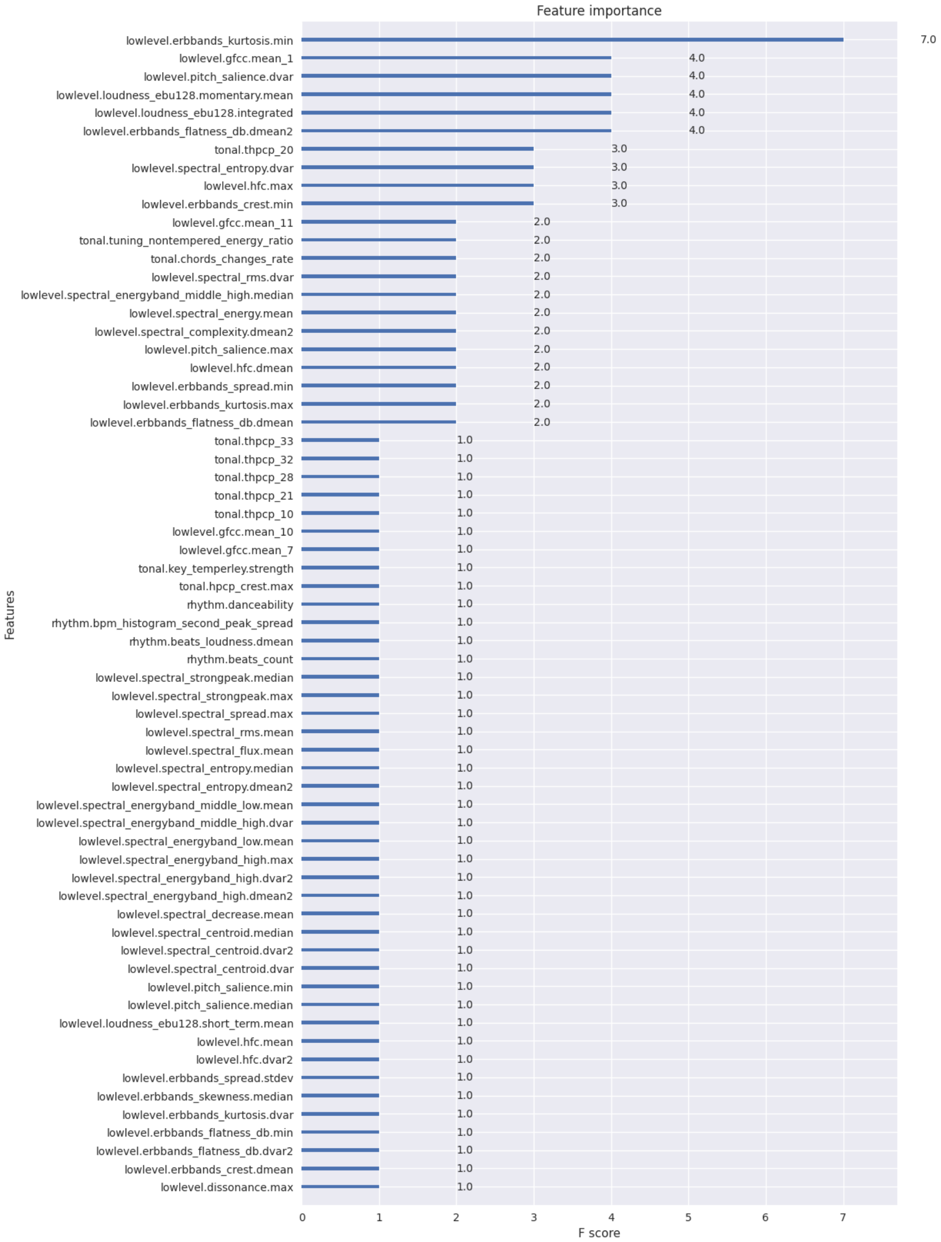


Figure A.20: Importance plot yearly ESC-SR classifier 2021.

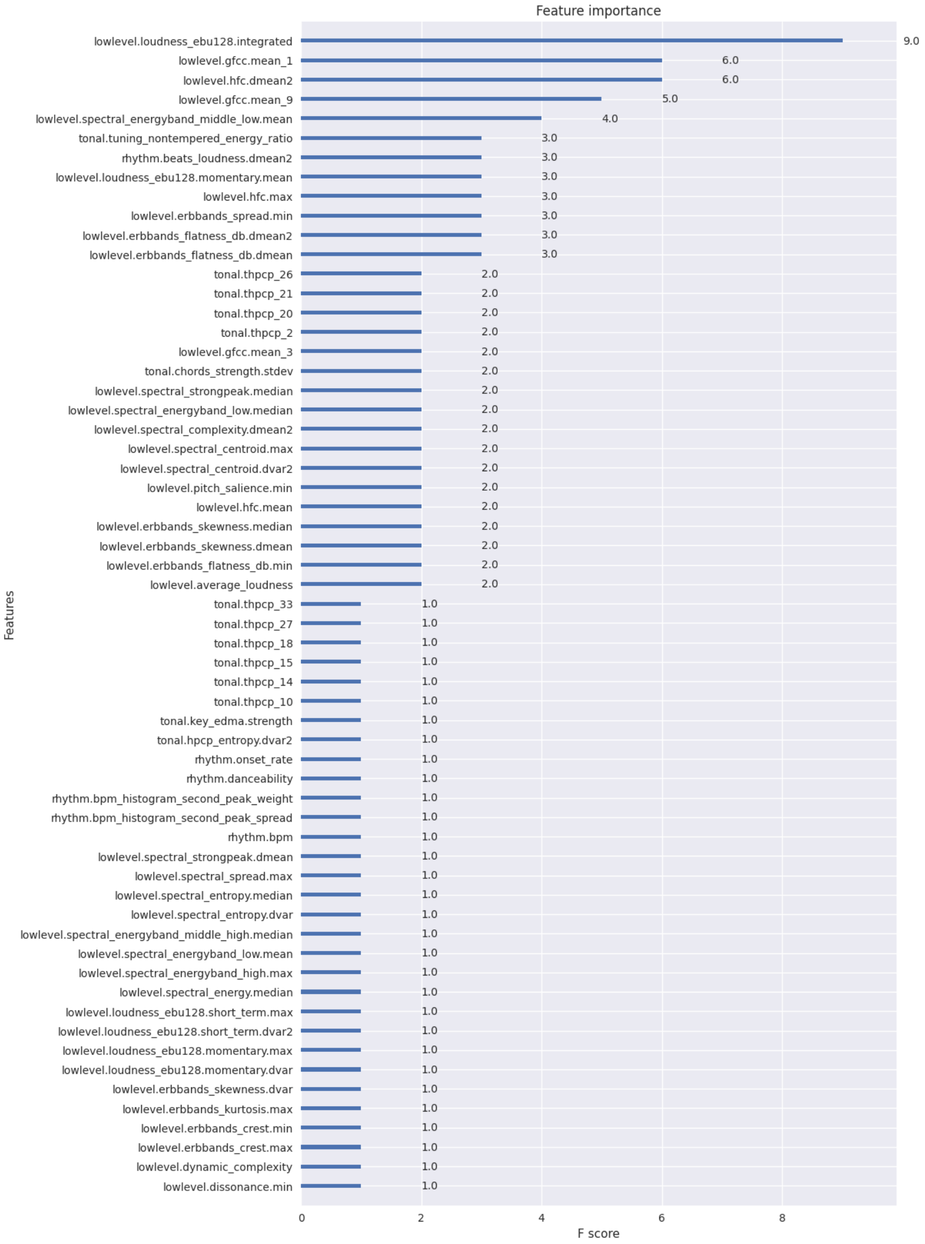
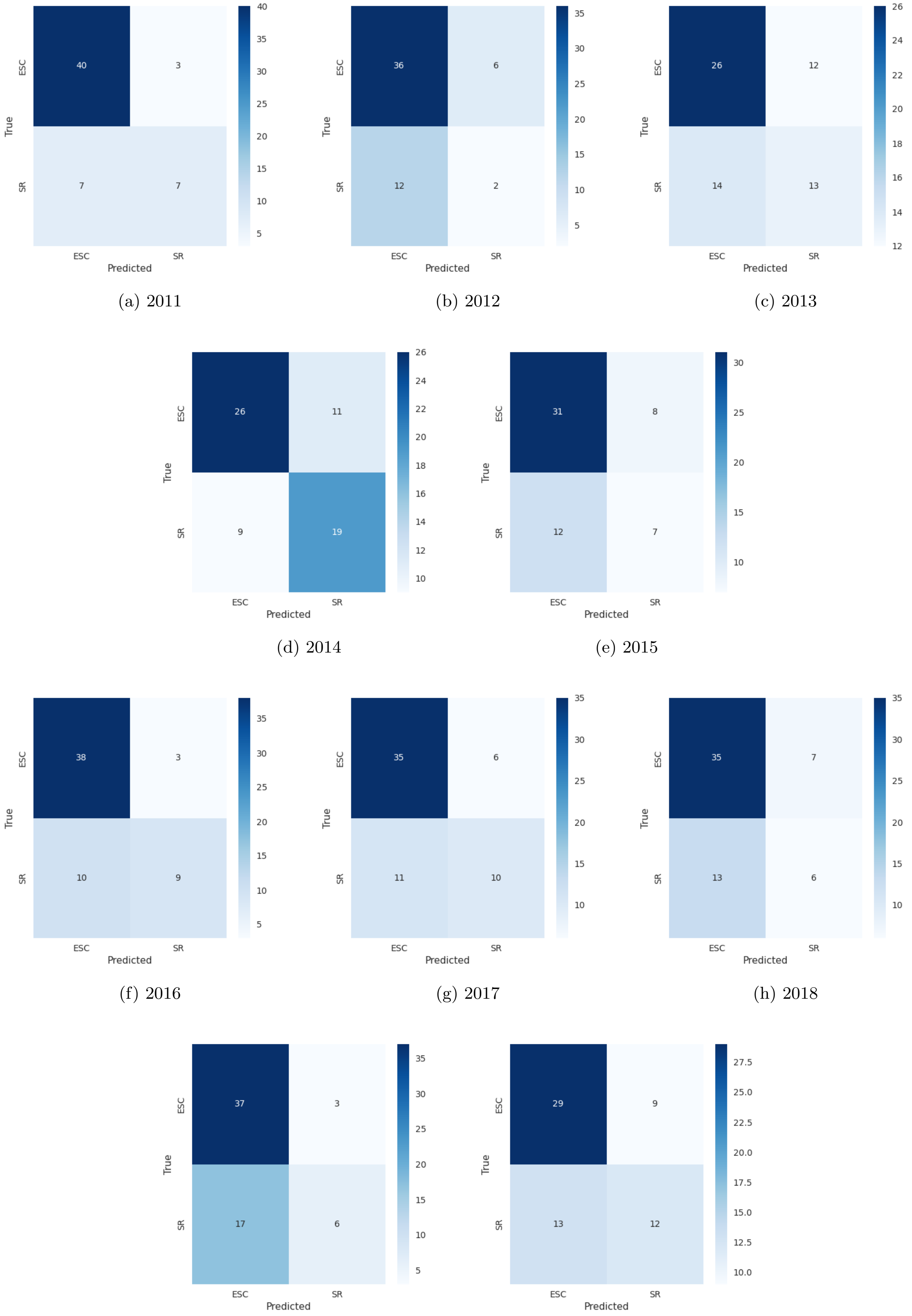


Figure A.21: Importance plot selection ESC-SR classifier 2021.

A.2.2 Confusion Matrices

Here we include the confusion matrices generated by the yearly and selection classifiers.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

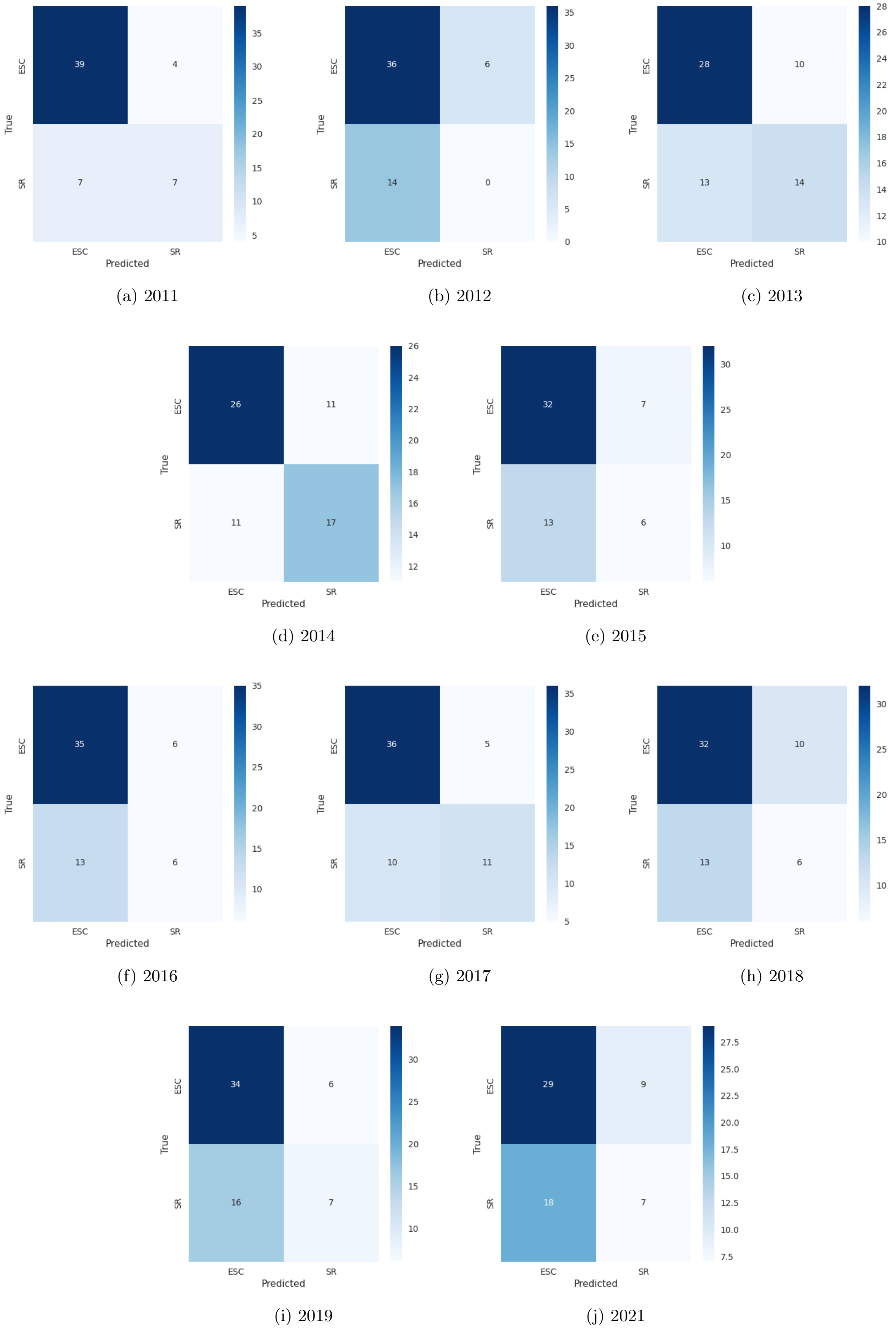
(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure A.22: Confusion matrices per year.
Constructed from evaluating the yearly ESC-SR classifiers.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure A.23: Confusion matrices per year.
Constructed from evaluating the selection ESC-SR classifiers.

A.3 Sweden

A.3.1 Importance plots

Here we include all importance plots generated by the Eurovision-Melodifestivalen classifiers.

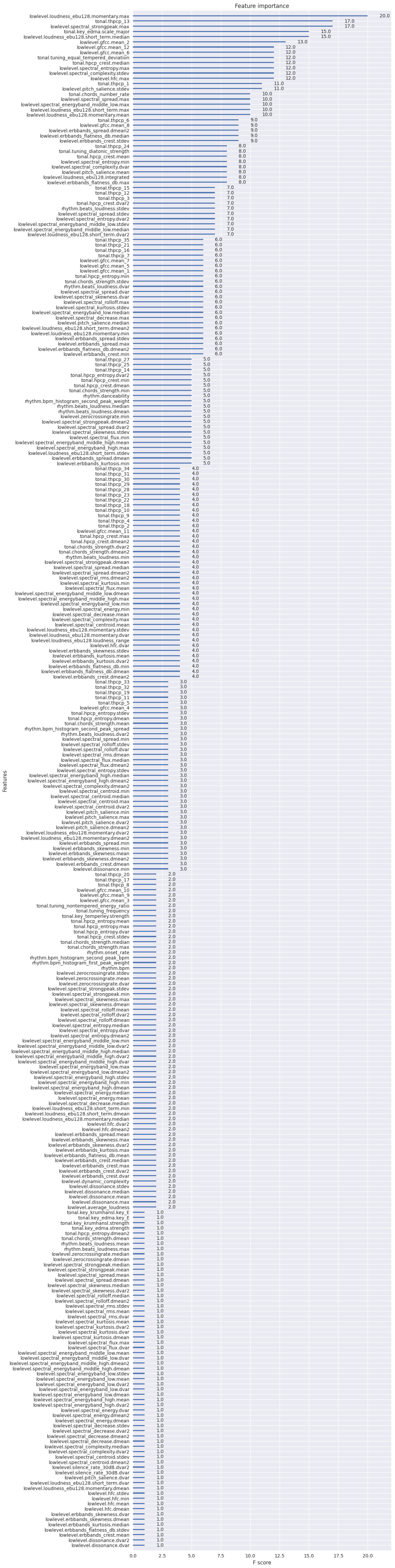


Figure A.24: Importance plot general ESC-MF classifier.

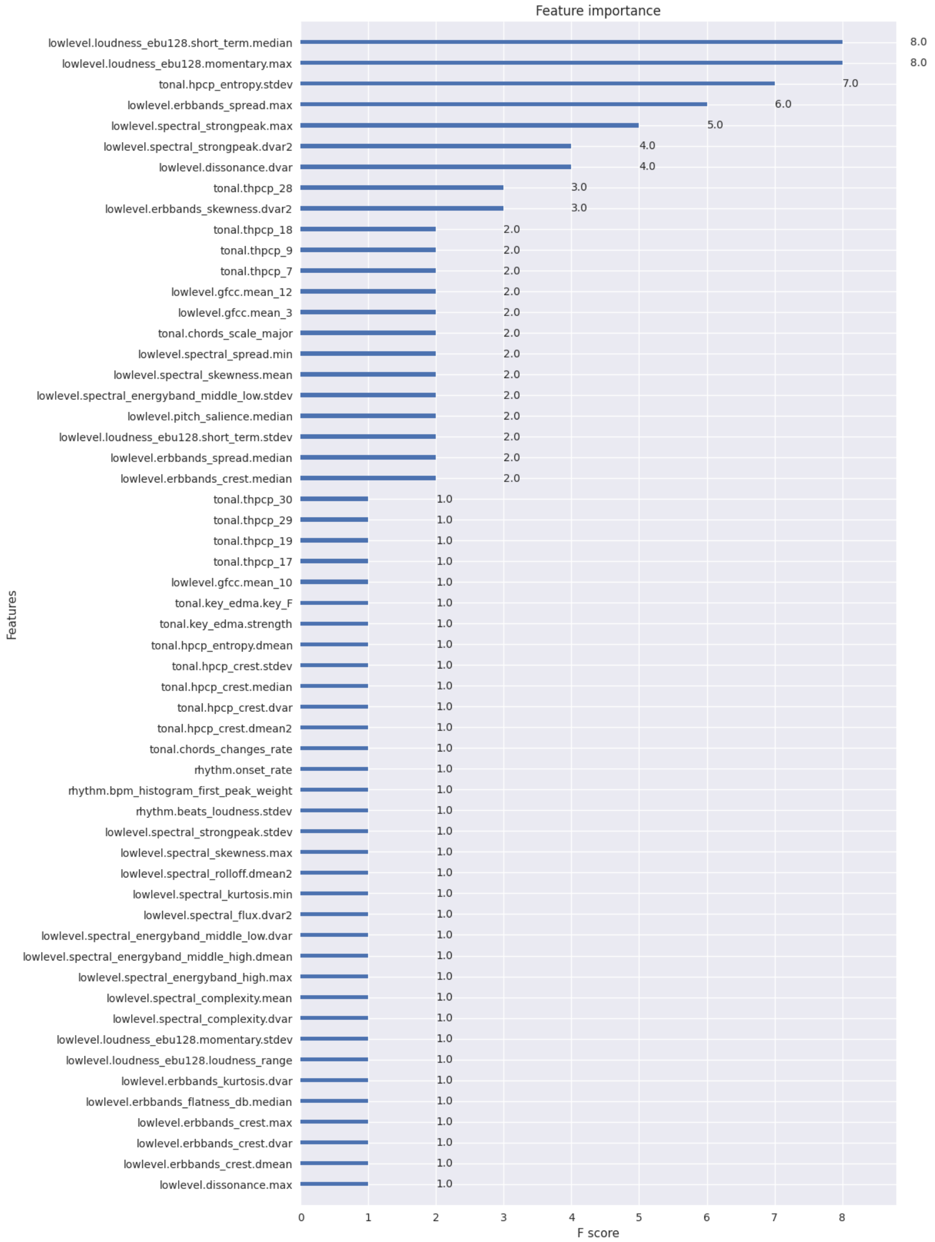


Figure A.25: Importance plot yearly ESC-MF classifier 2011.

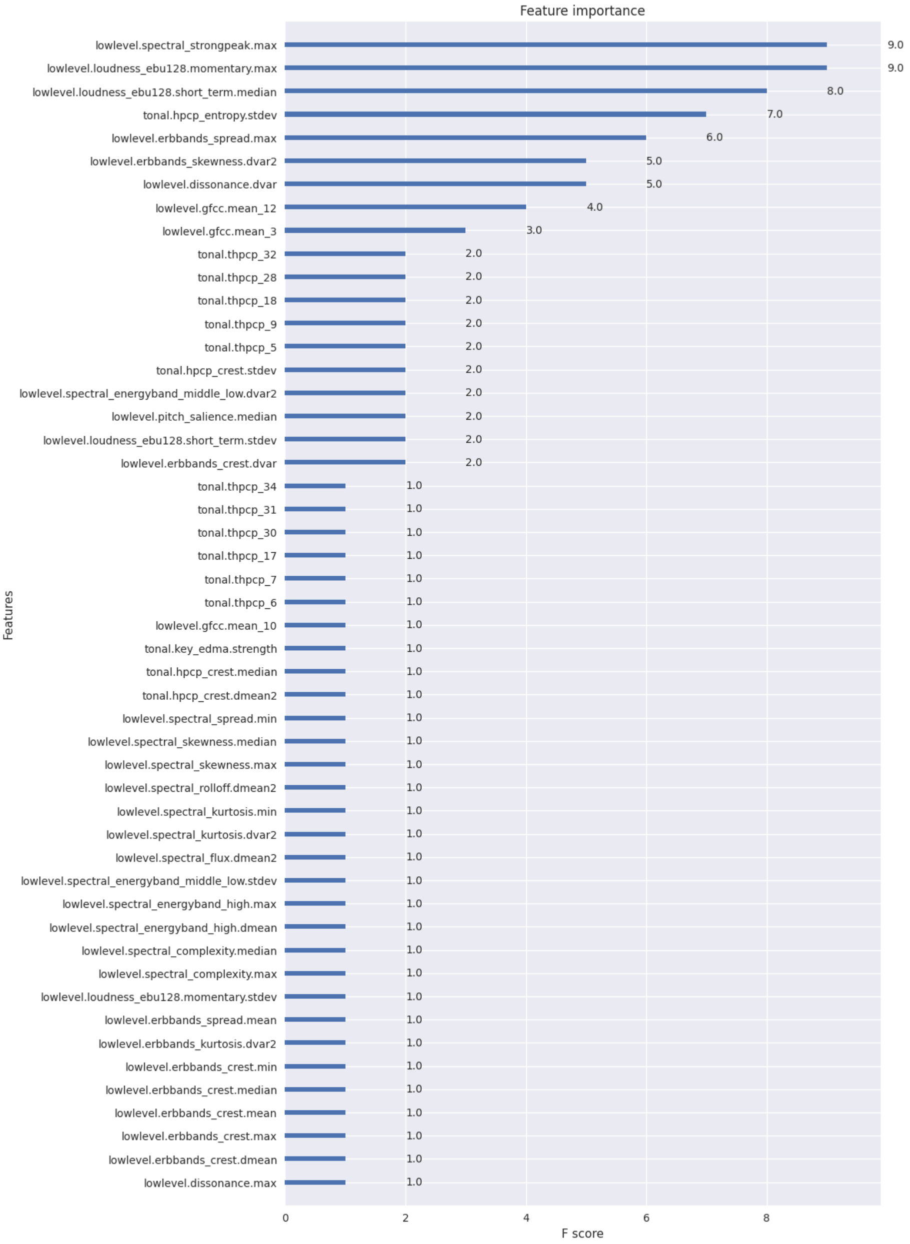


Figure A.26: Importance plot selection ESC-MF classifier 2011.

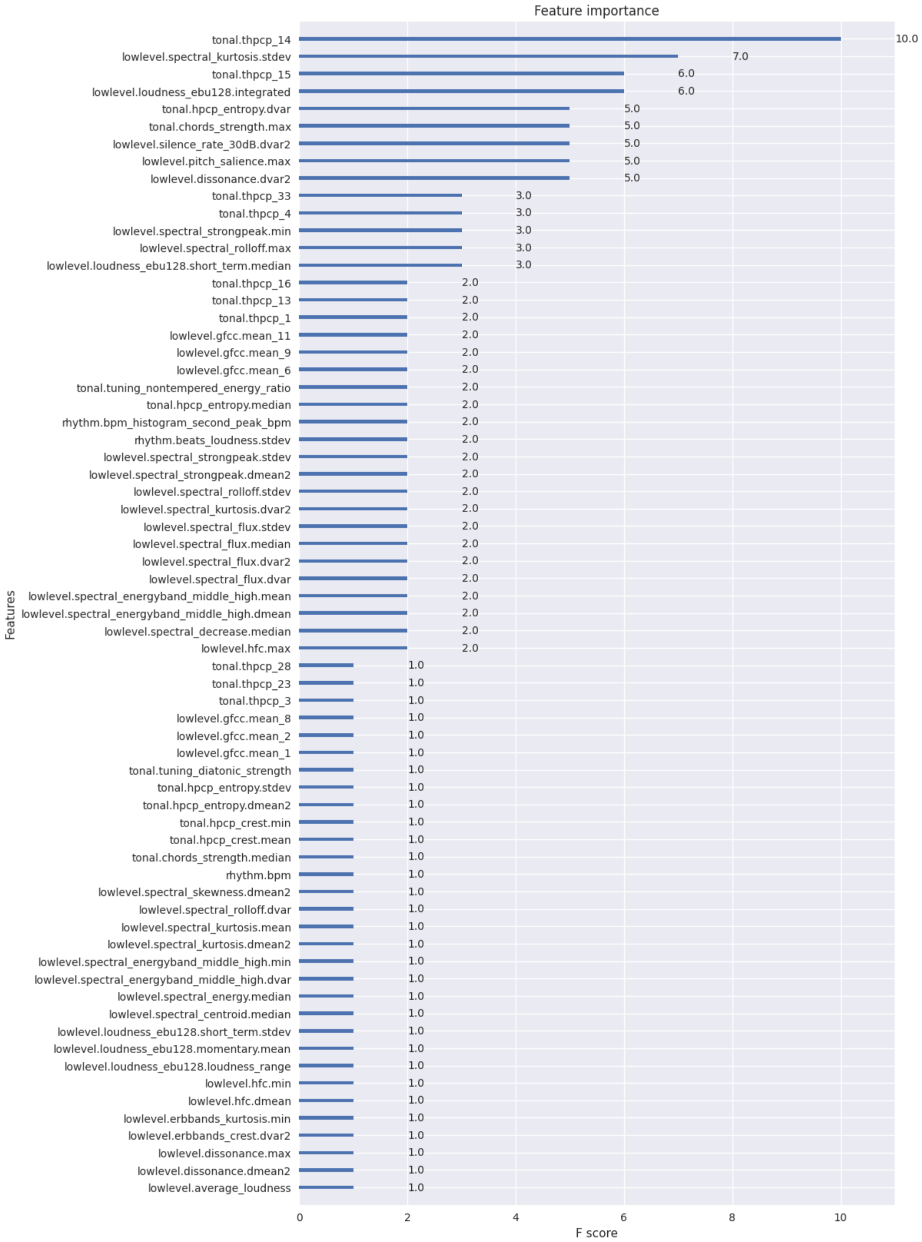


Figure A.27: Importance plot yearly ESC-MF classifier 2012.

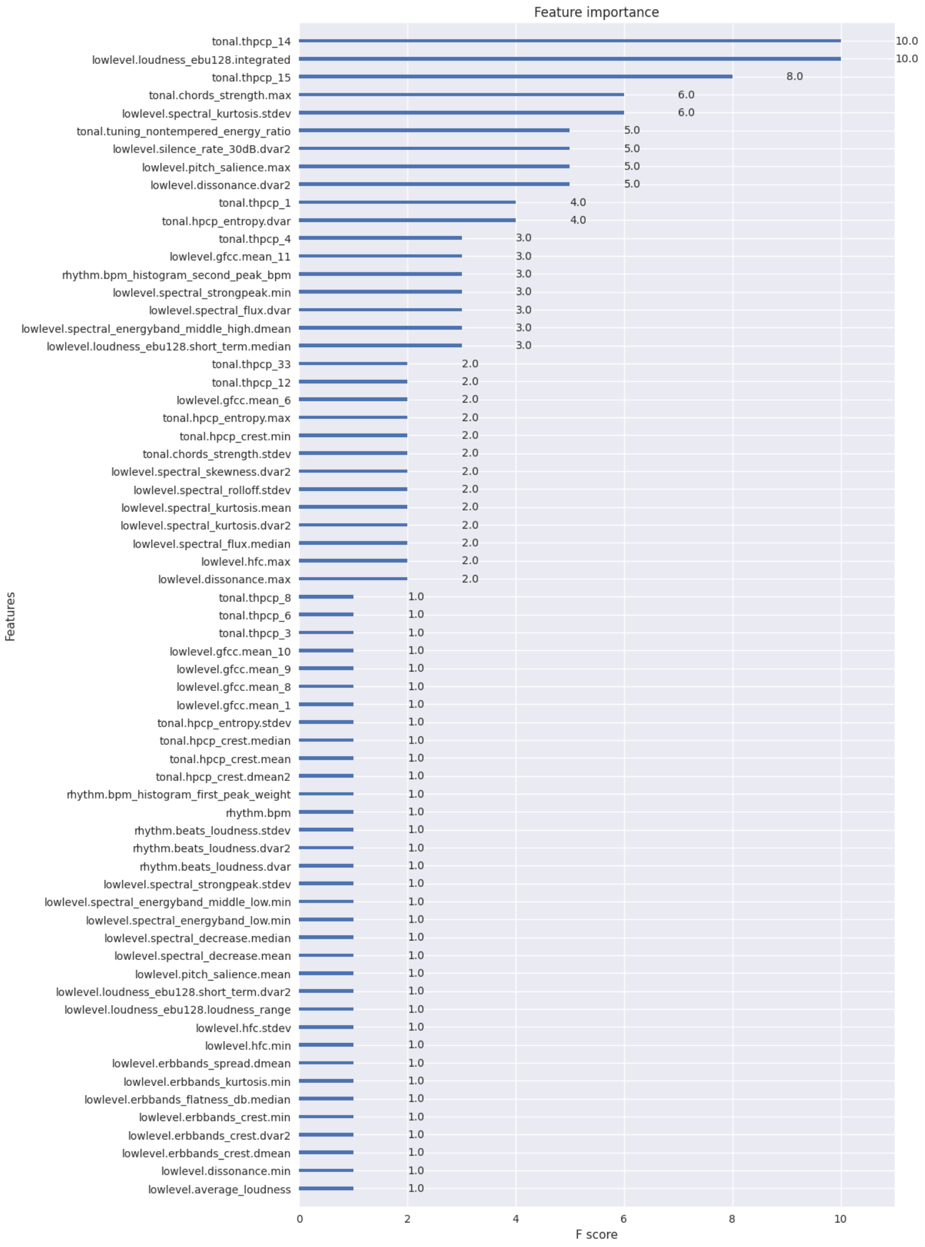


Figure A.28: Importance plot selection ESC-MF classifier 2012.

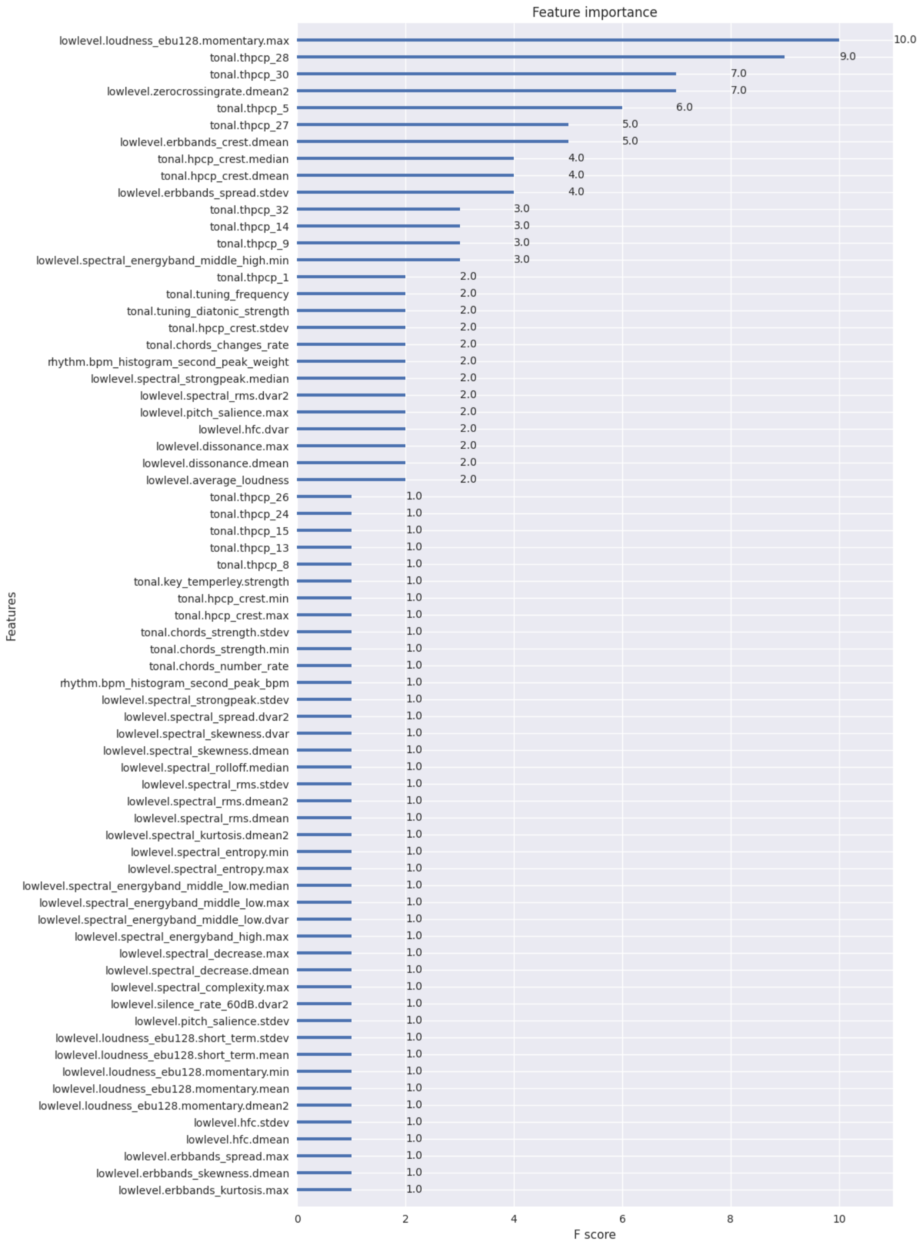


Figure A.29: Importance plot yearly ESC-MF classifier 2013.

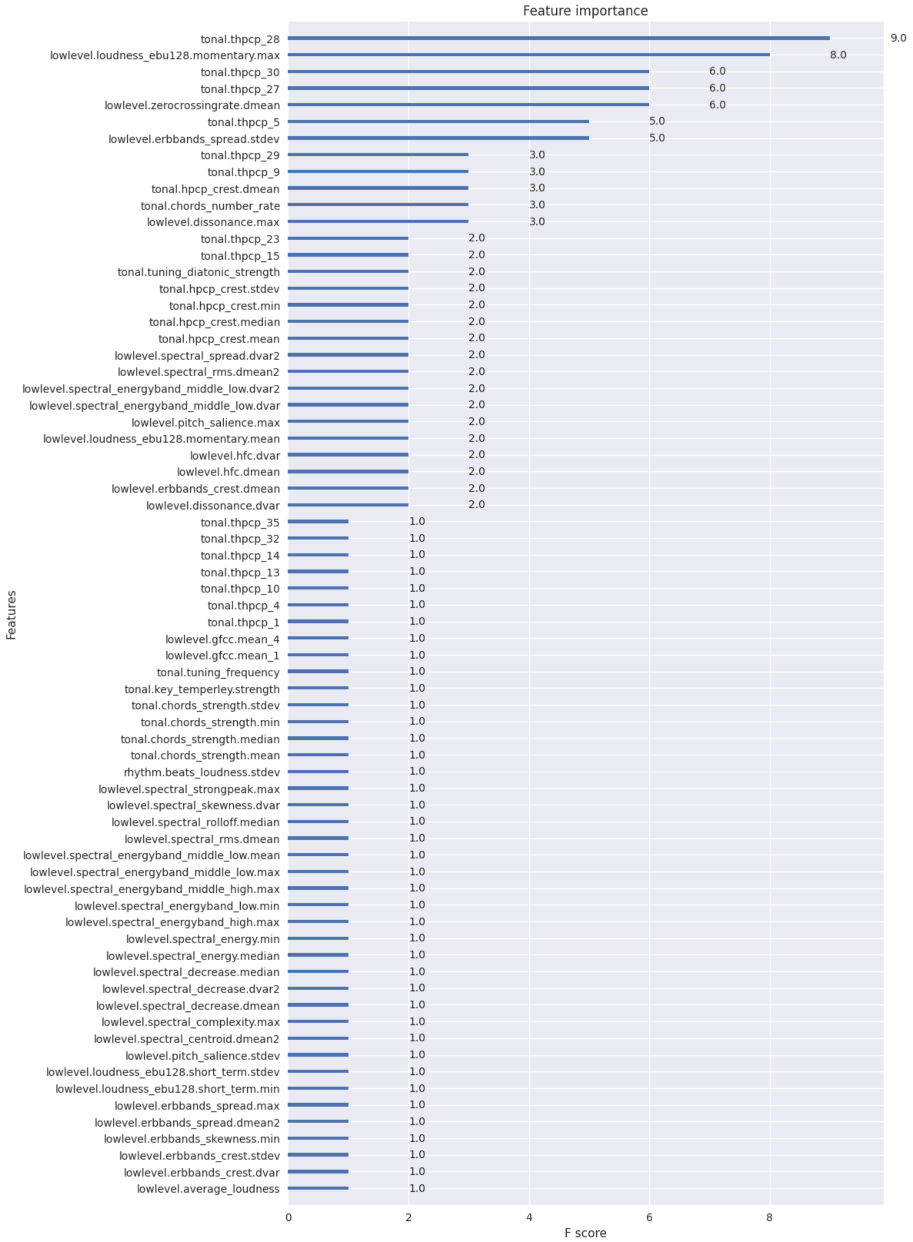


Figure A.30: Importance plot selection ESC-MF classifier 2013.

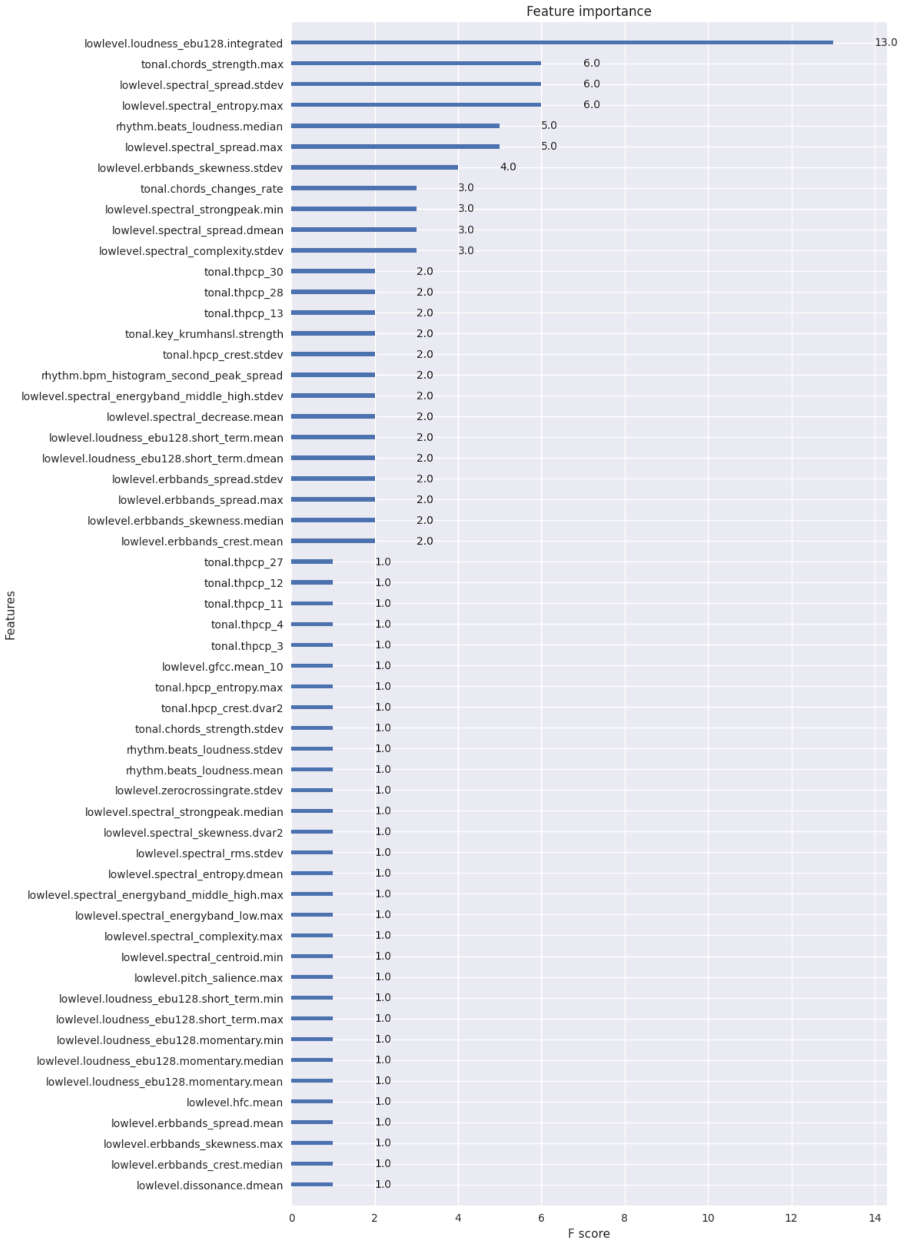


Figure A.31: Importance plot yearly ESC-MF classifier 2014.

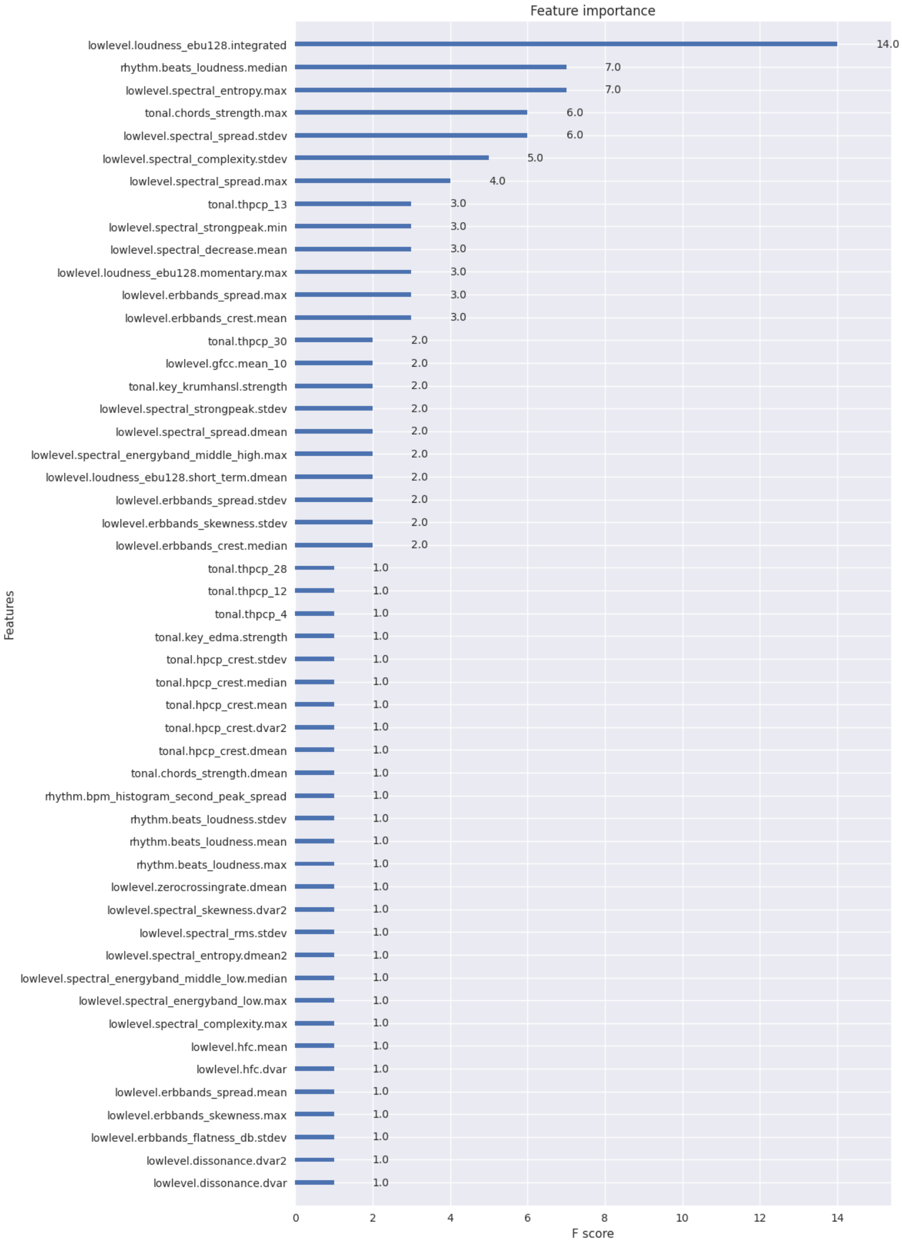


Figure A.32: Importance plot selection ESC-MF classifier 2014.

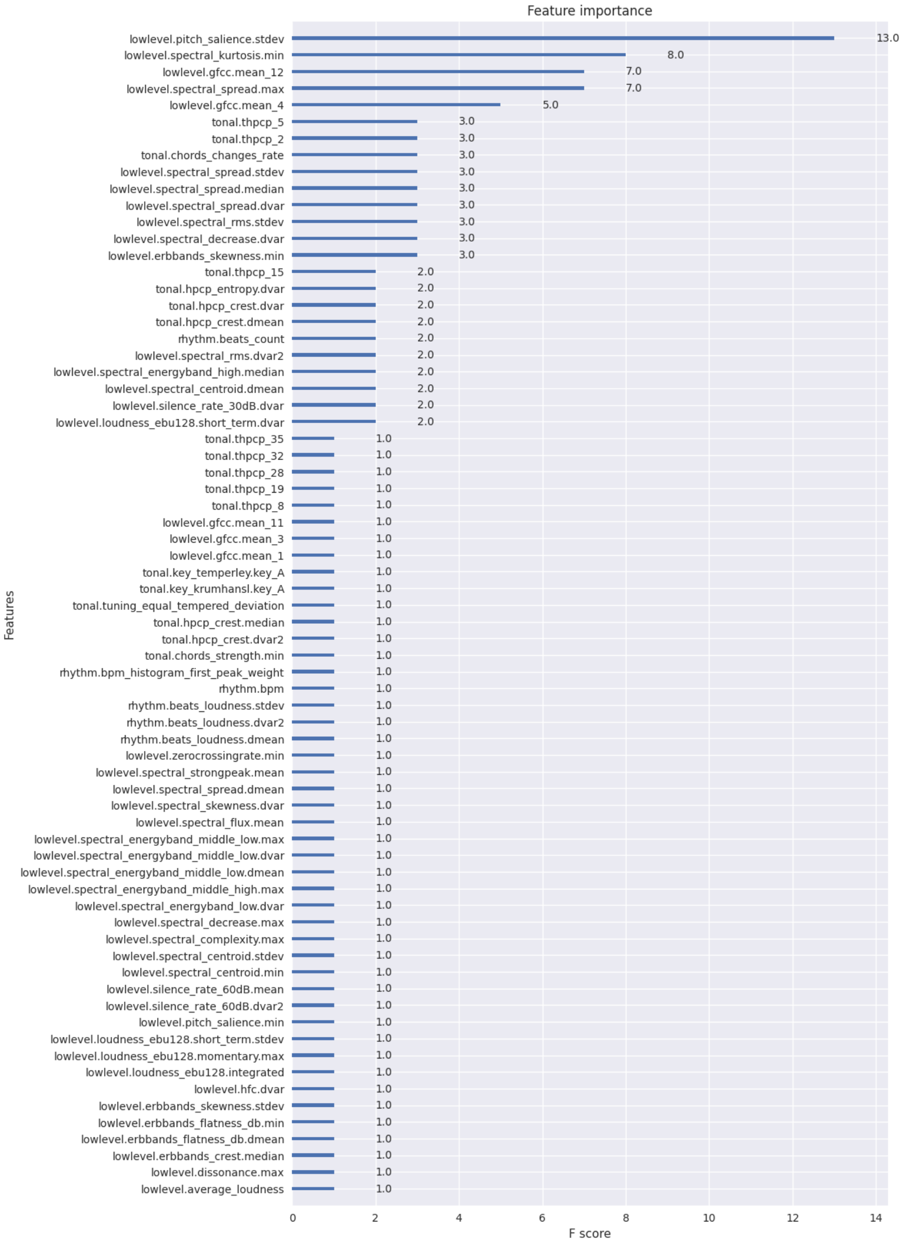


Figure A.33: Importance plot yearly ESC-MF classifier 2015.

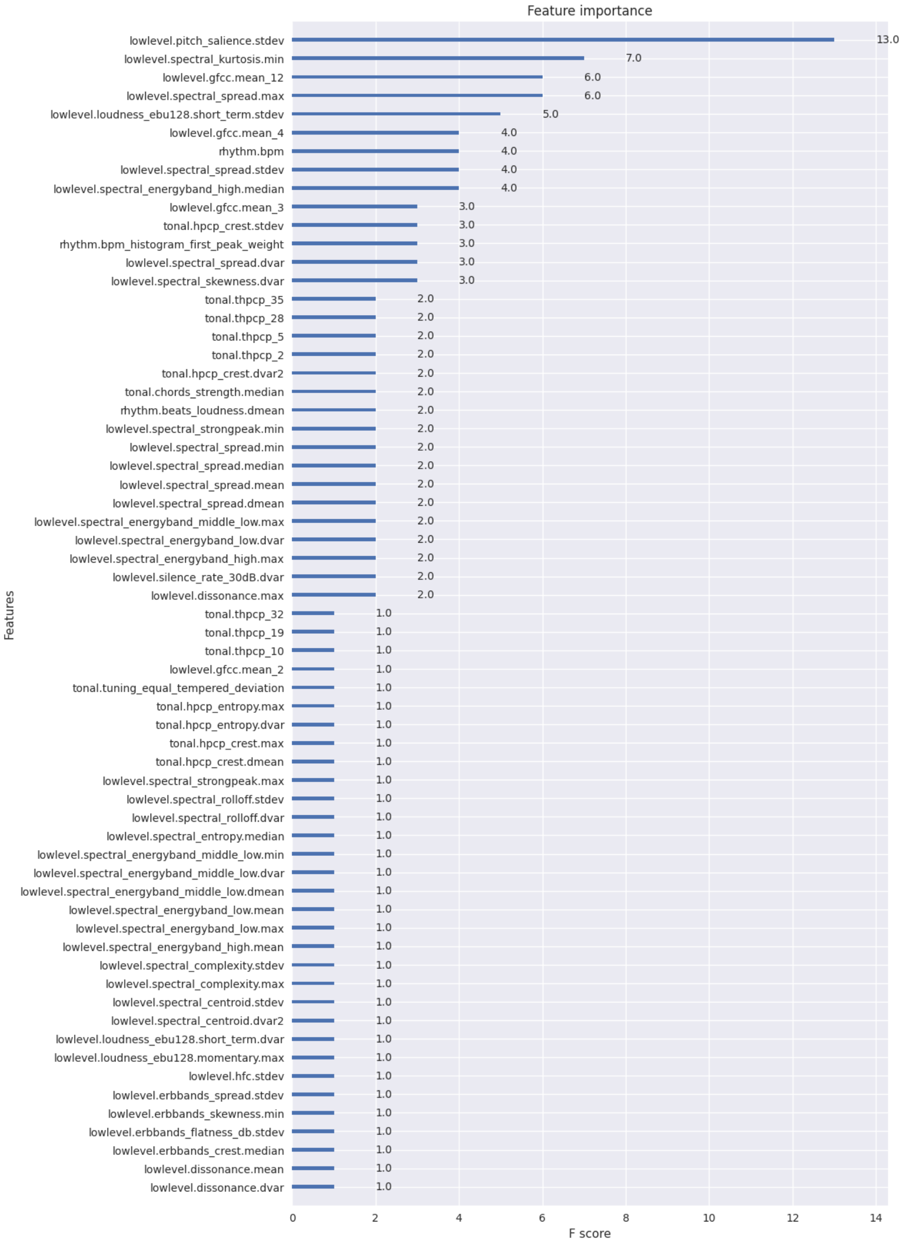


Figure A.34: Importance plot selection ESC-MF classifier 2015.

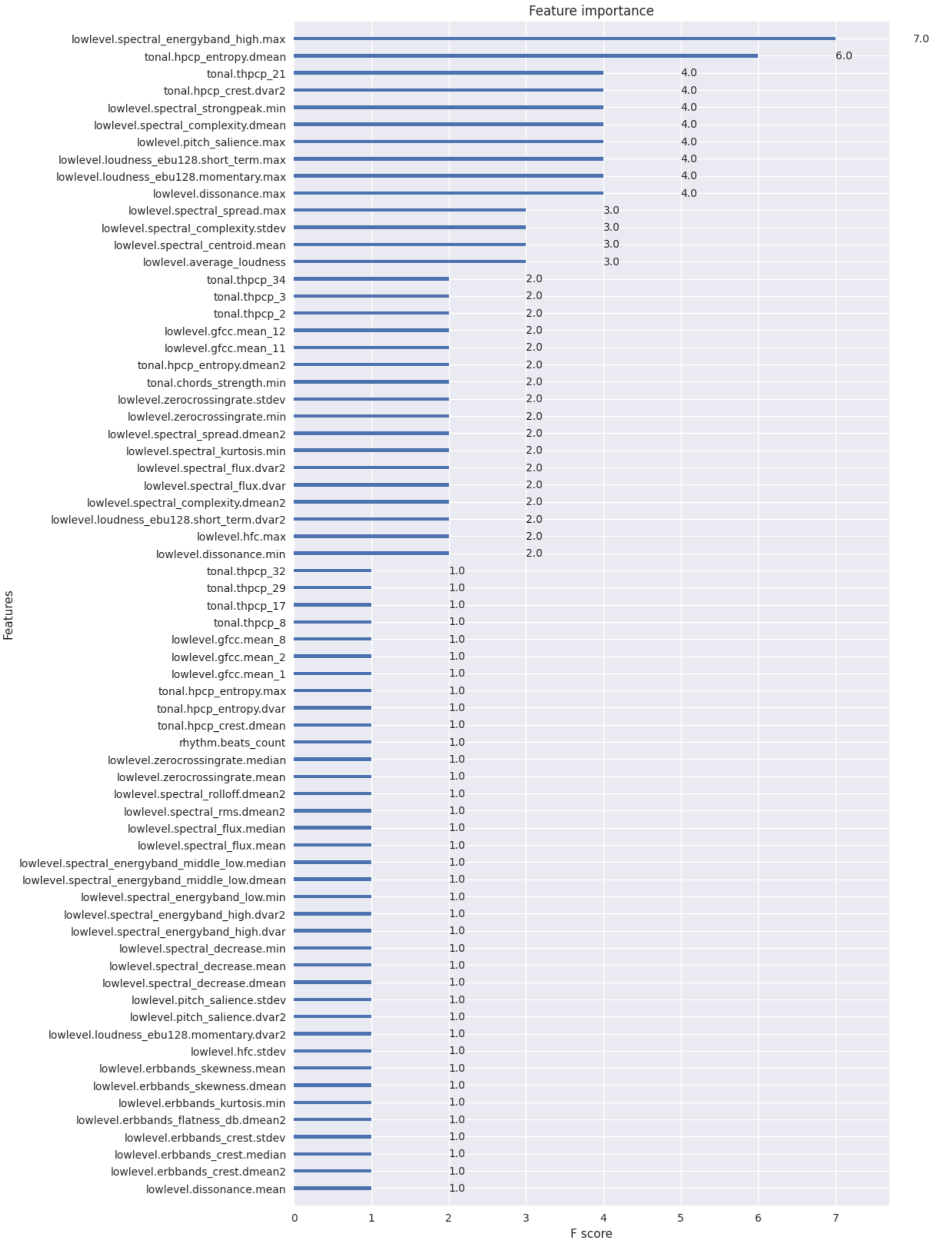


Figure A.35: Importance plot yearly ESC-MF classifier 2016.

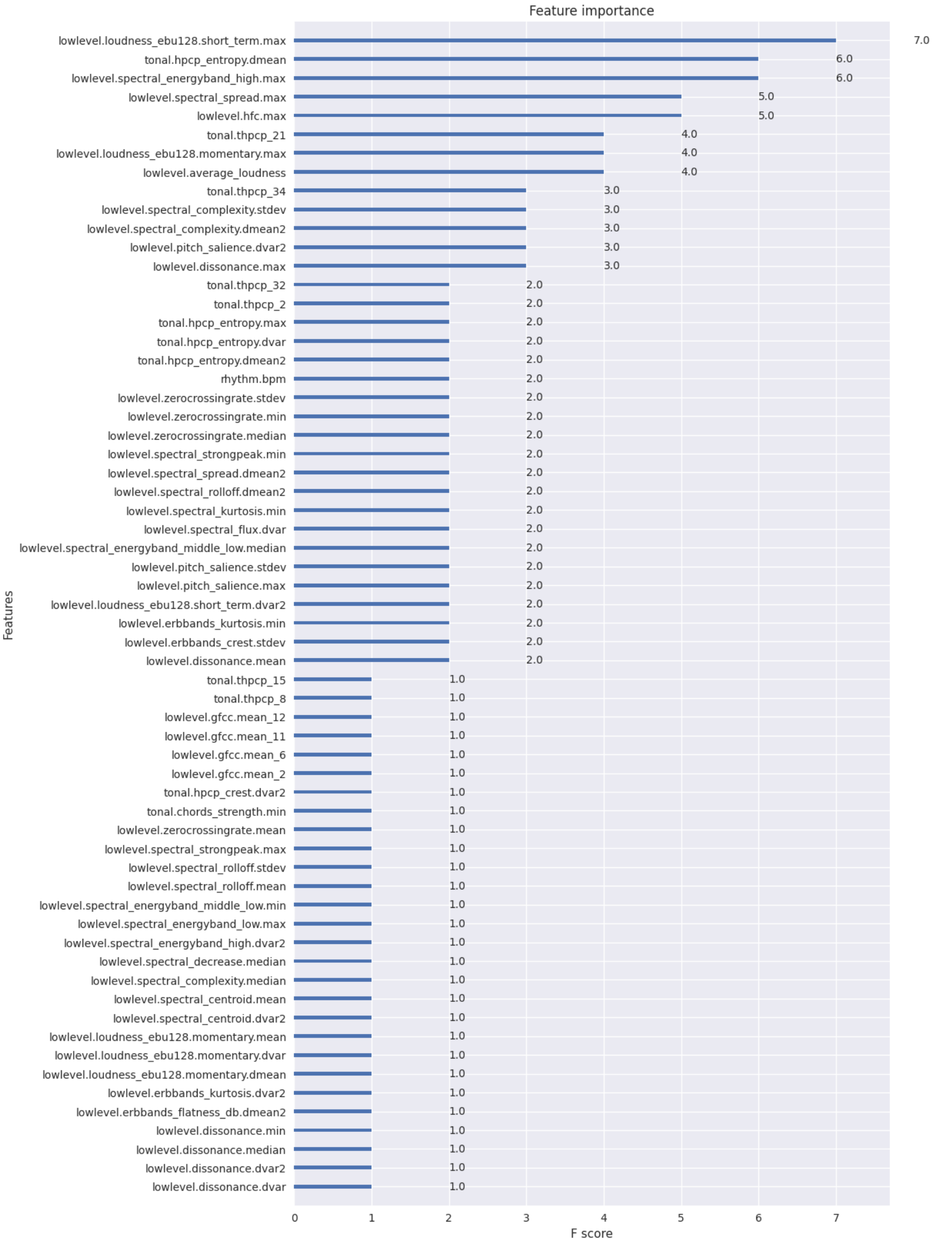


Figure A.36: Importance plot selection ESC-MF classifier 2016.

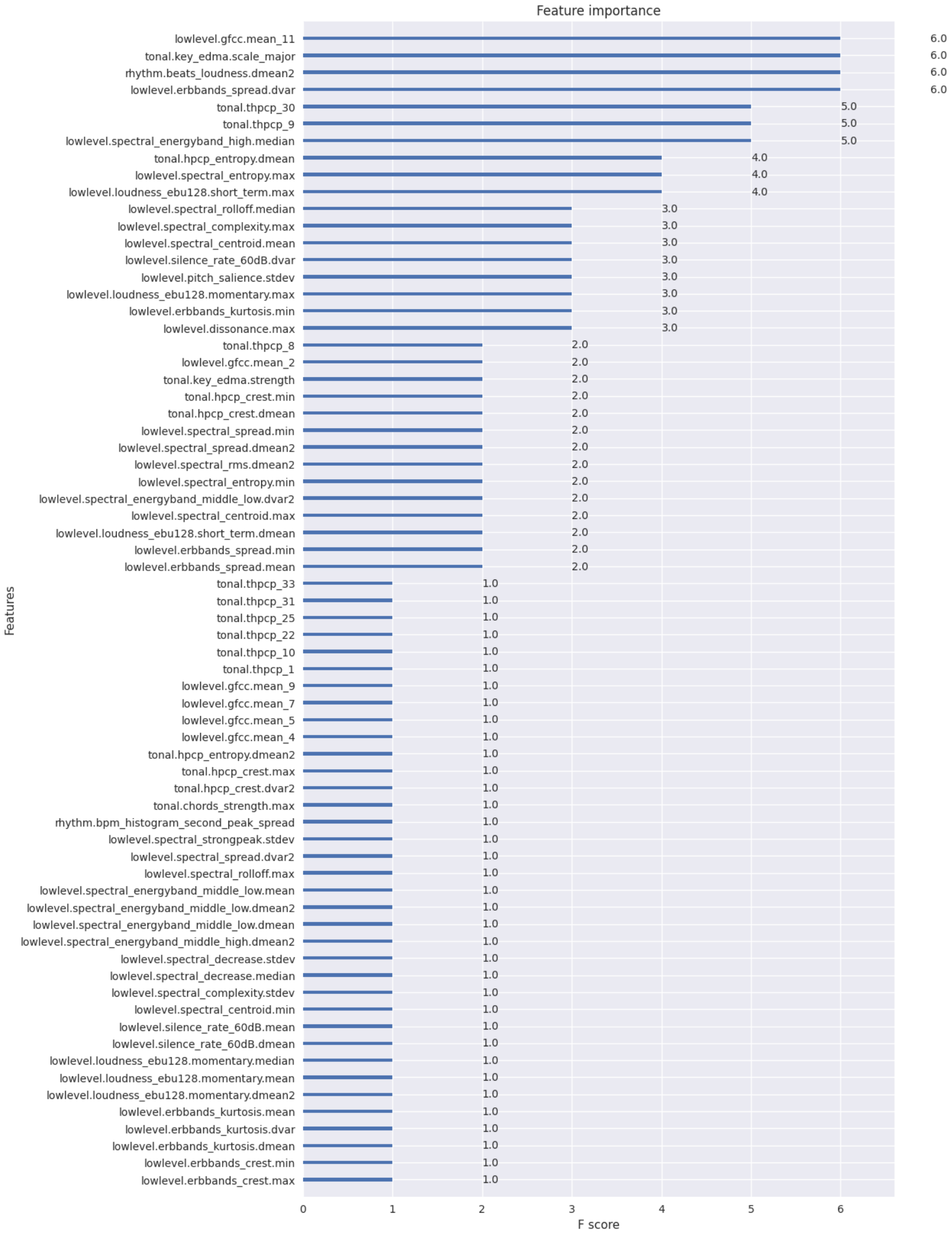


Figure A.37: Importance plot yearly ESC-MF classifier 2017.

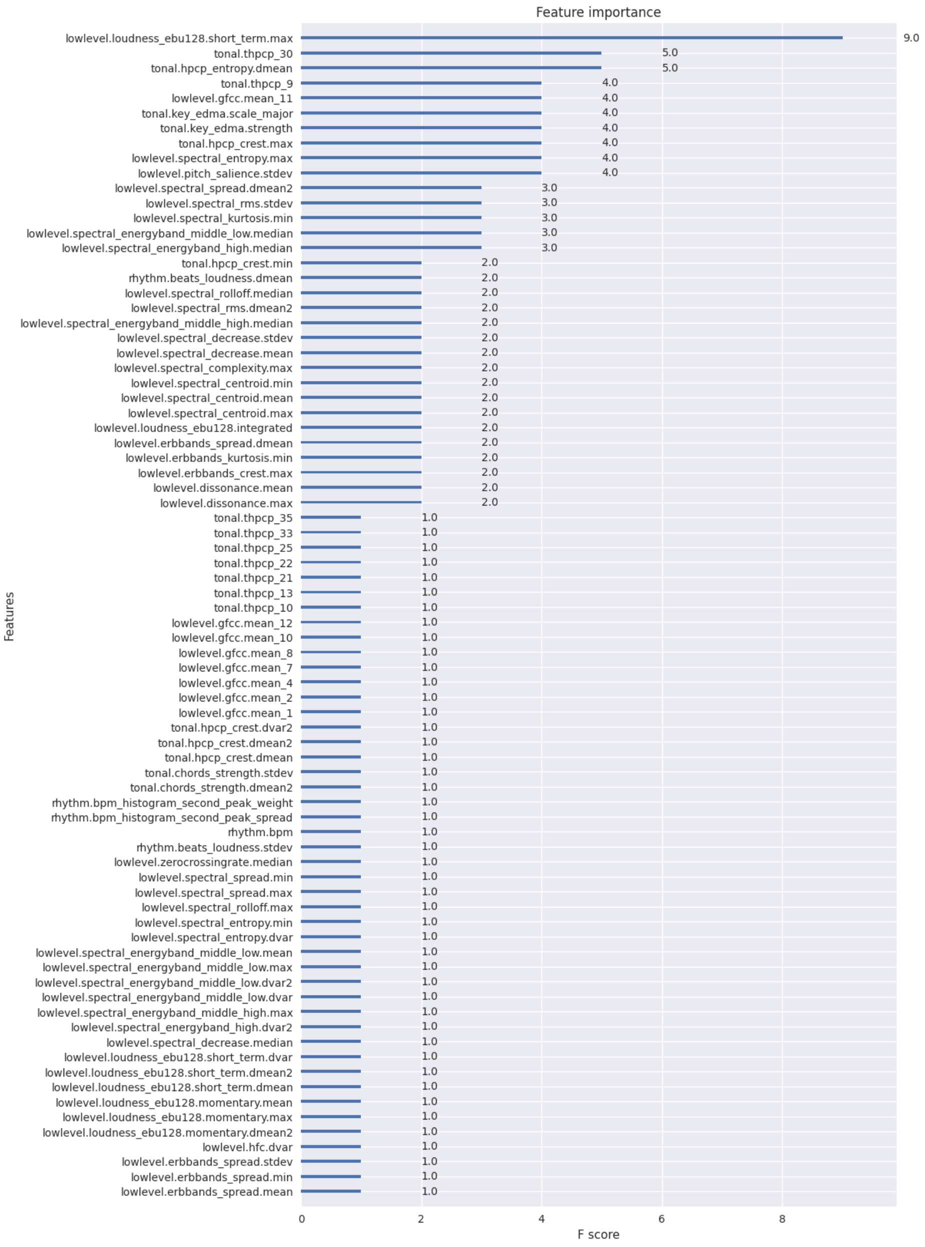


Figure A.38: Importance plot selection ESC-MF classifier 2017.

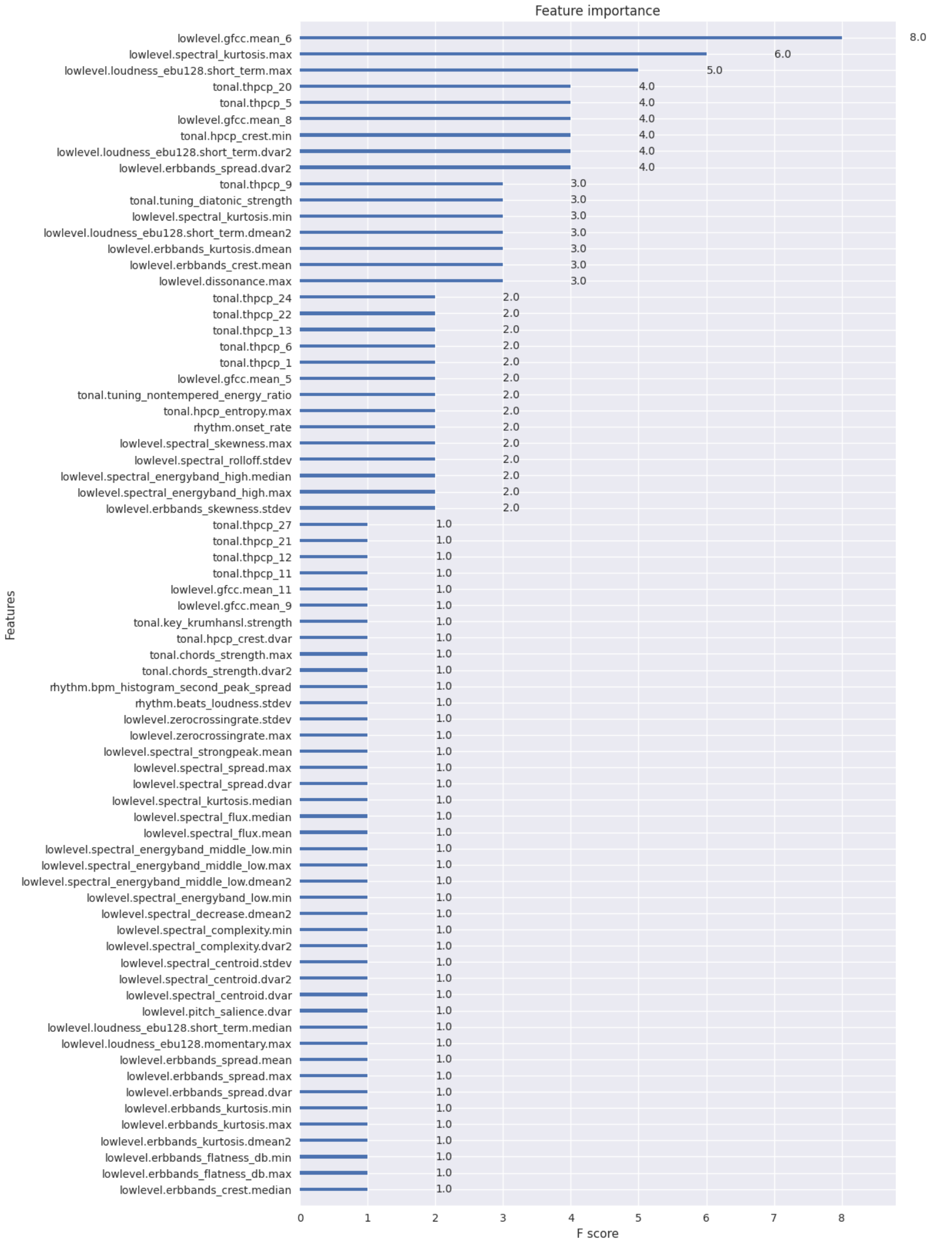


Figure A.39: Importance plot yearly ESC-MF classifier 2018.

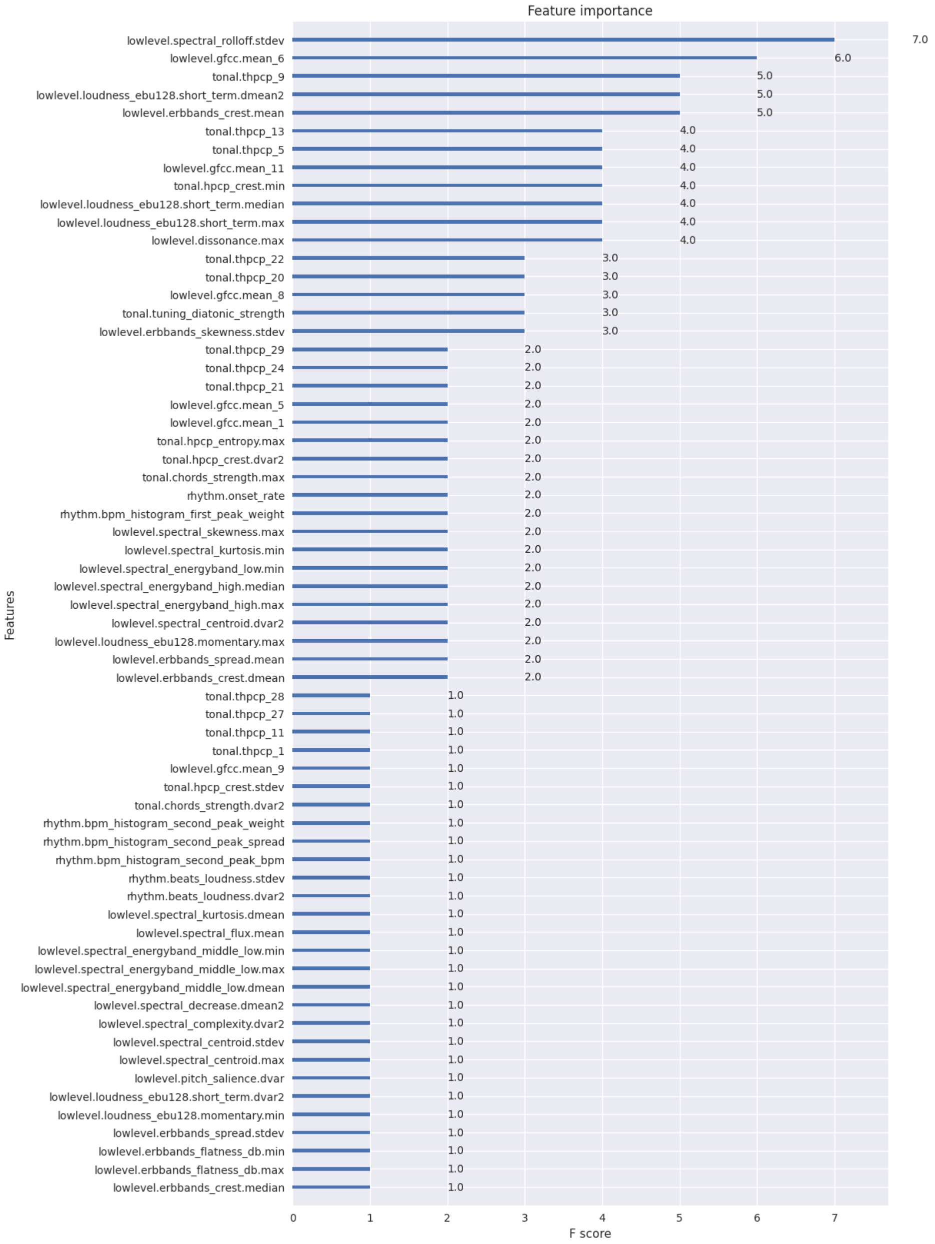


Figure A.40: Importance plot selection ESC-MF classifier 2018.

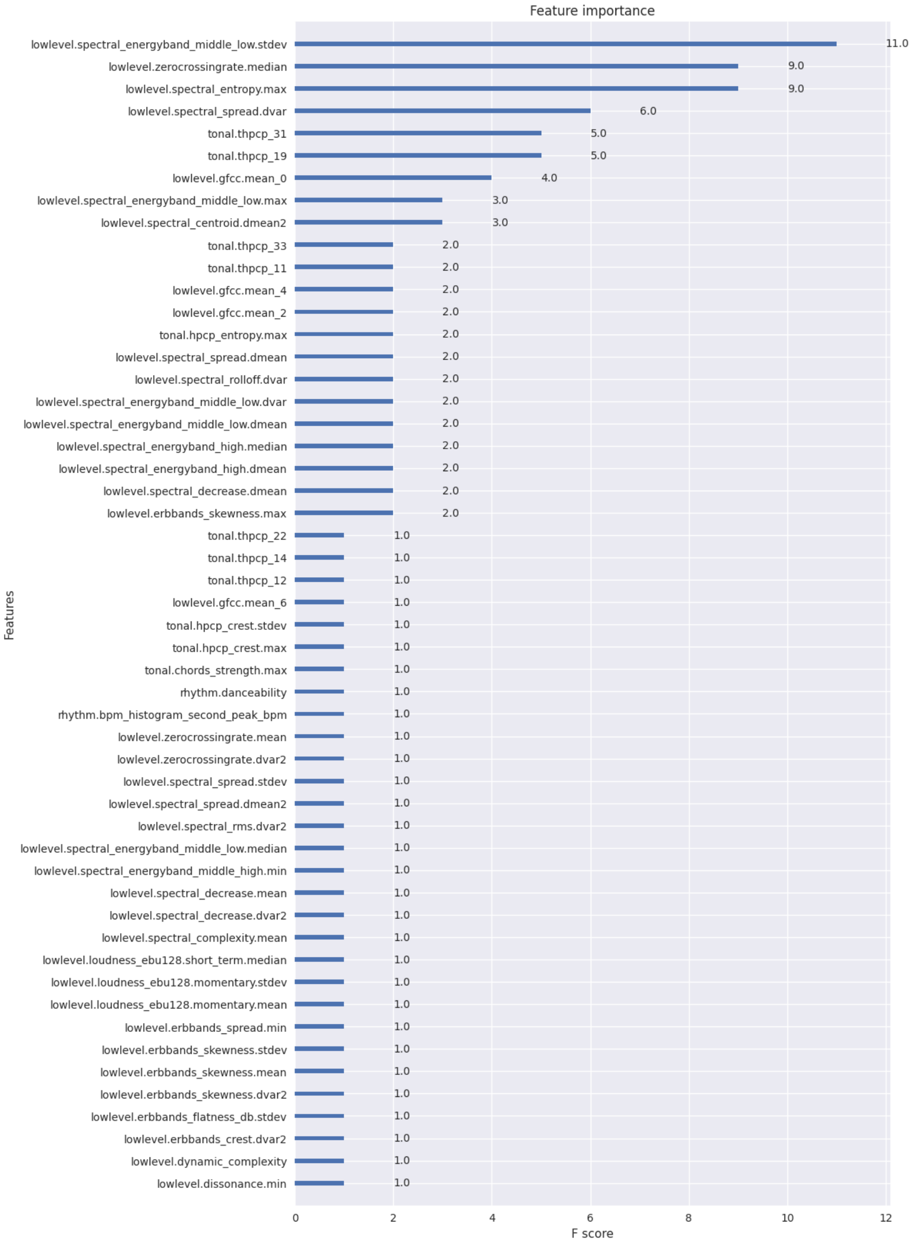


Figure A.41: Importance plot yearly ESC-MF classifier 2019.

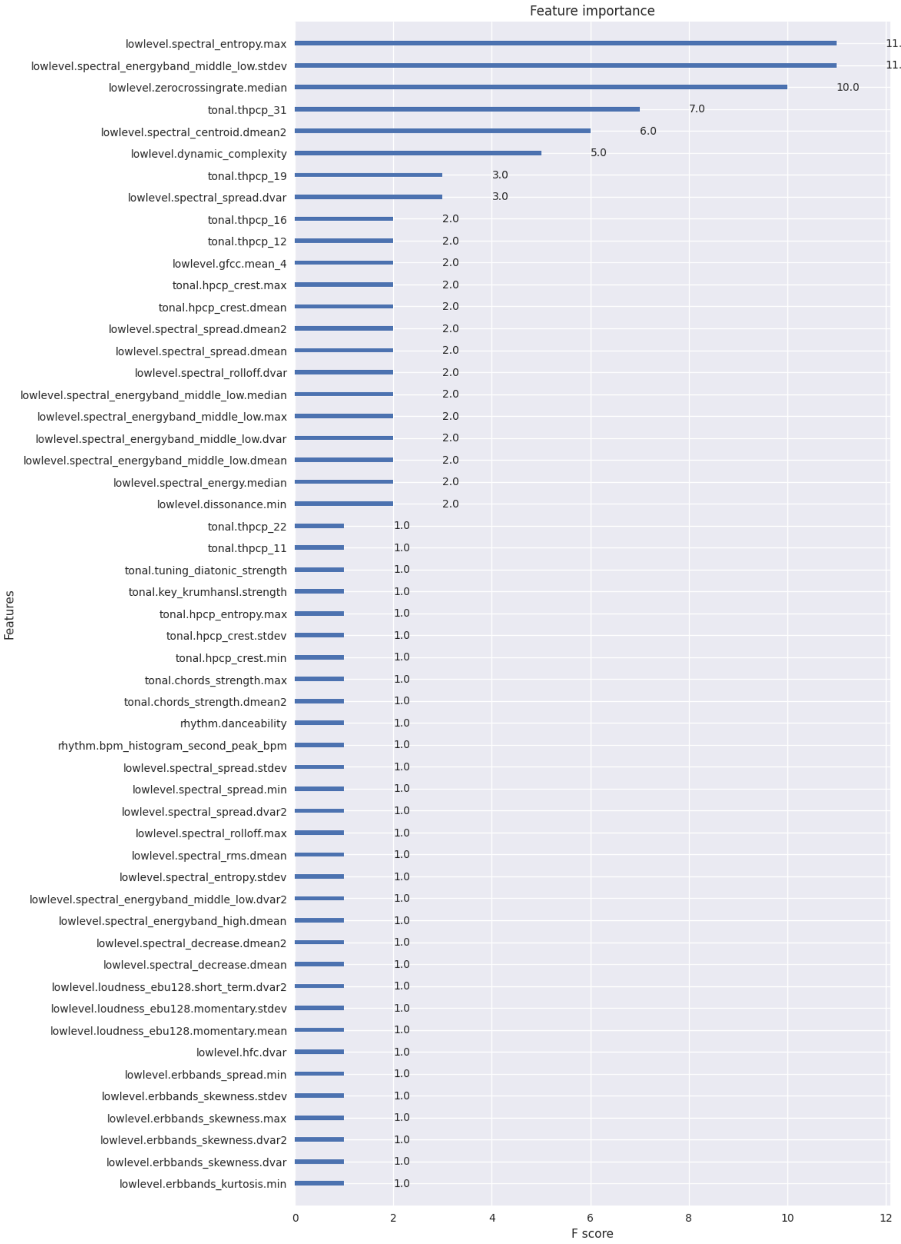


Figure A.42: Importance plot selection ESC-MF classifier 2019.

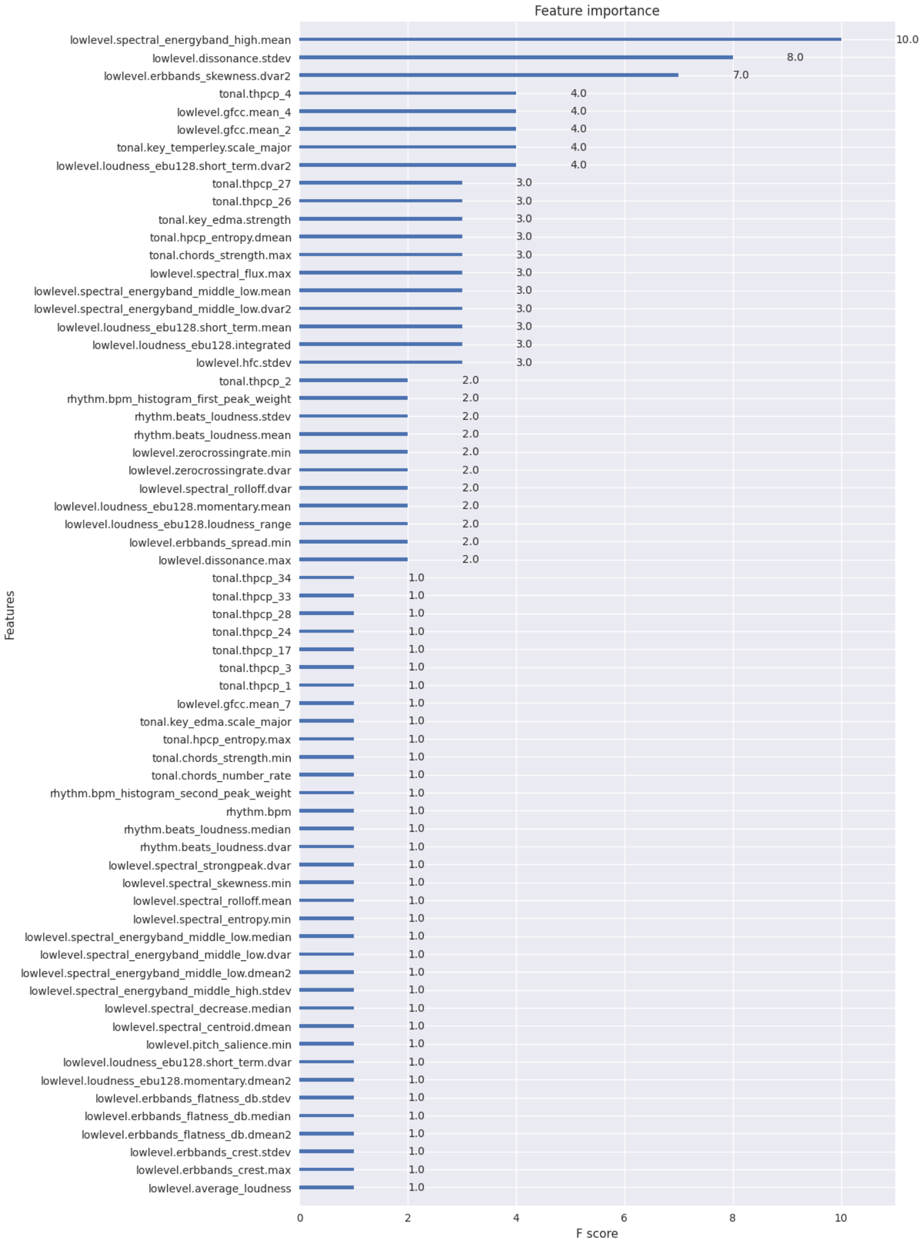


Figure A.43: Importance plot yearly ESC-MF classifier 2021.

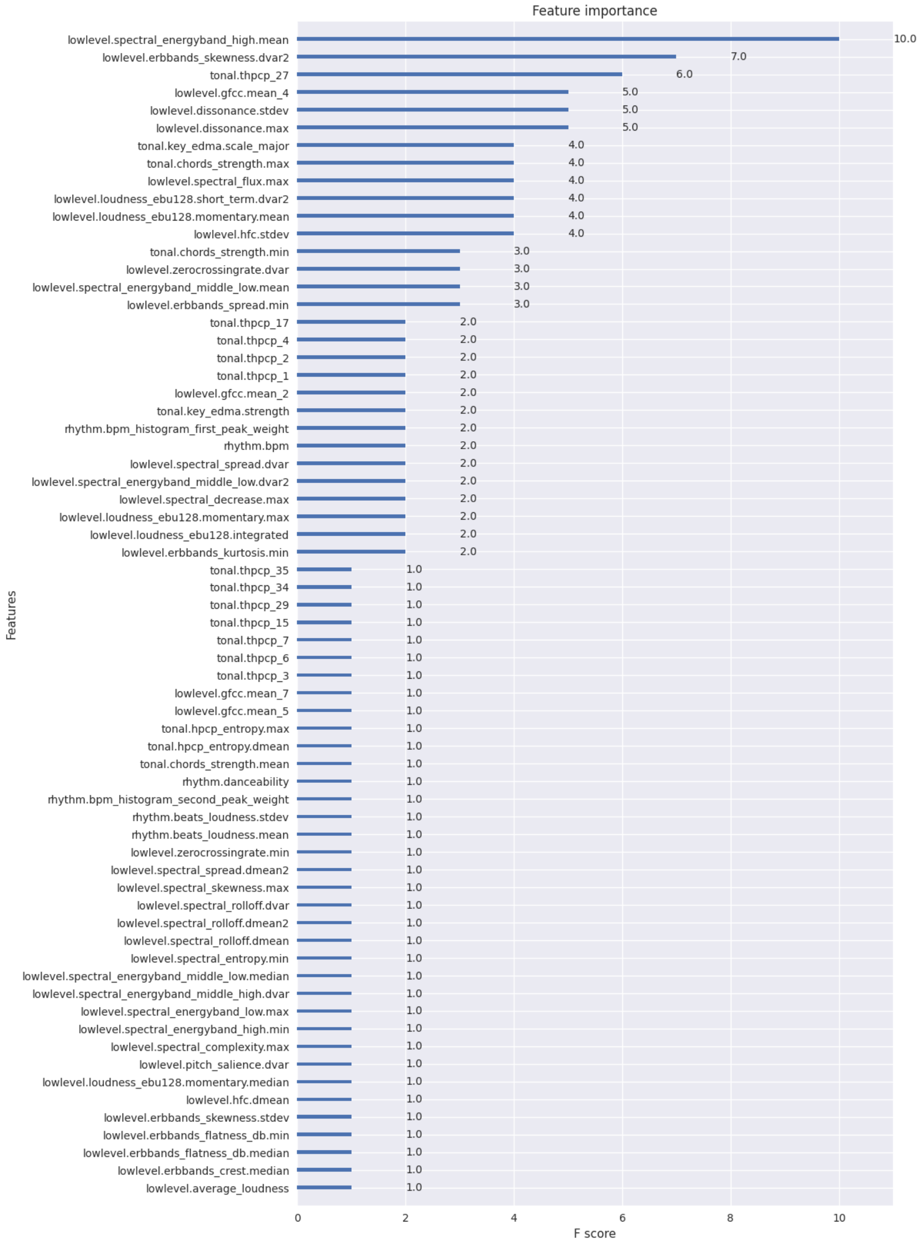
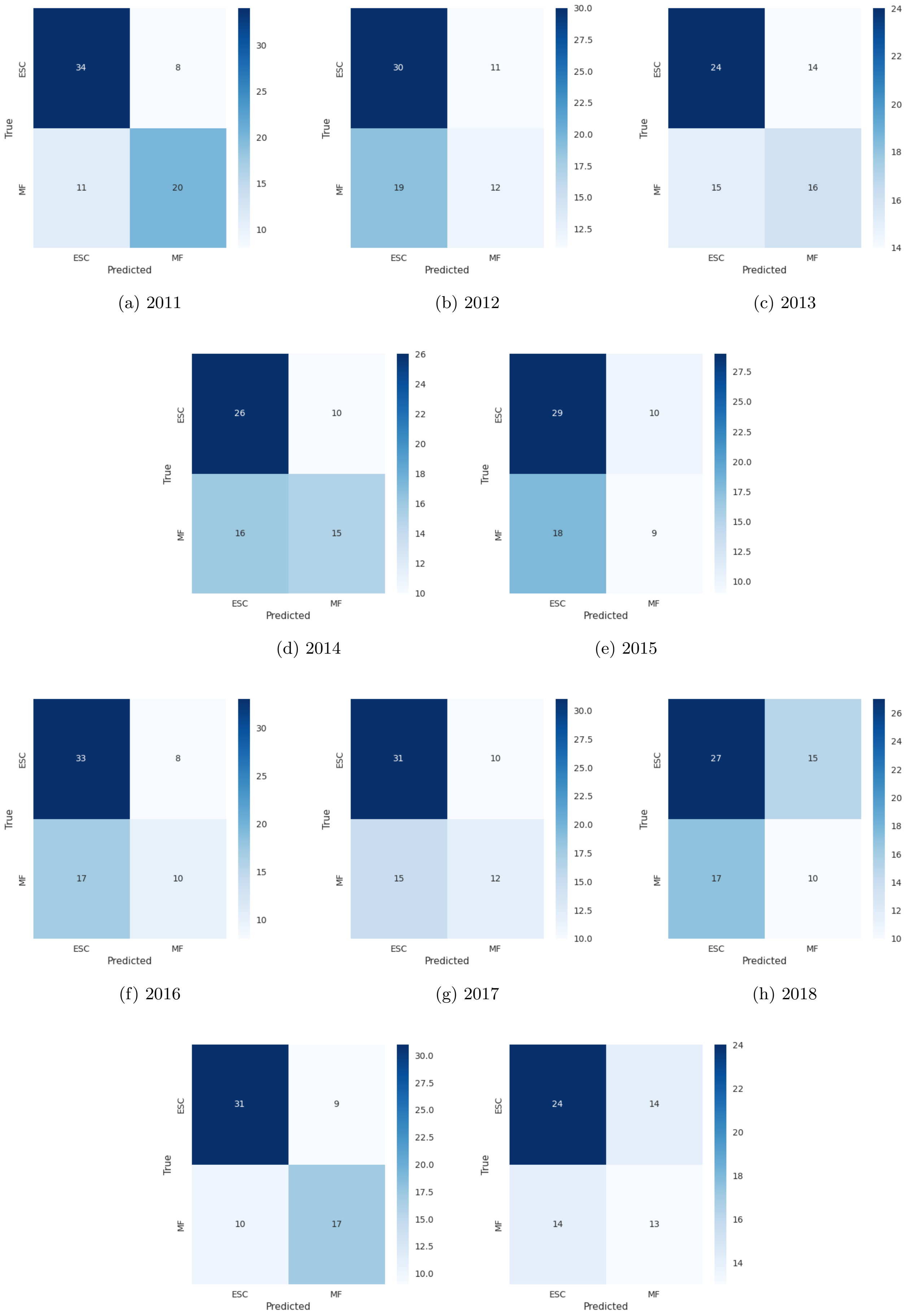


Figure A.44: Importance plot selection ESC-MF classifier 2021.

A.3.2 Confusion Matrices

Here we include the confusion matrices generated by the yearly and selection classifiers.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

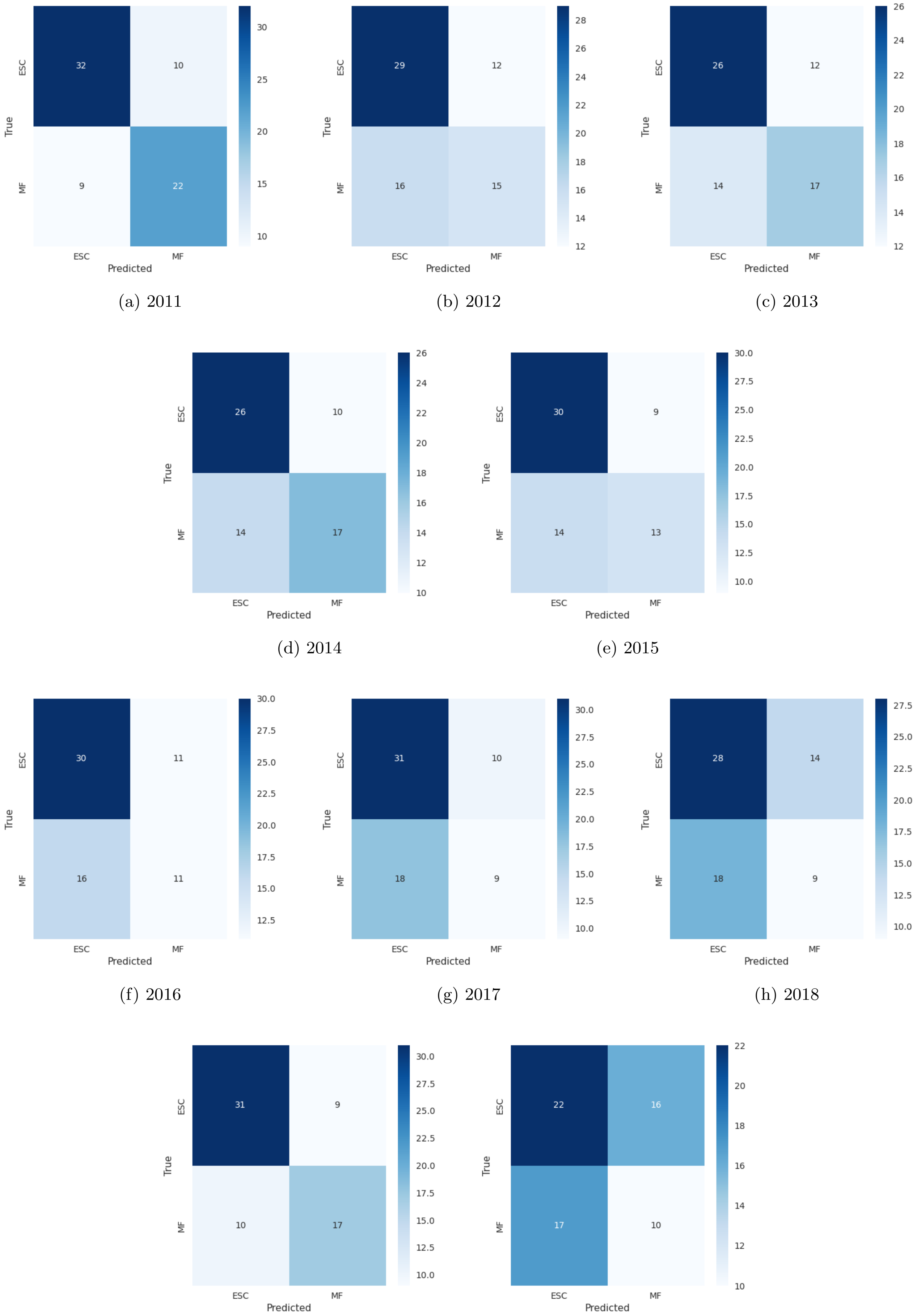
(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure A.45: Confusion matrices per year.
Constructed from evaluating the yearly ESC-MF classifiers.



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

(g) 2017

(h) 2018

(i) 2019

(j) 2021

Figure A.46: Confusion matrices per year.
Constructed from evaluating the selection ESC-MF classifiers.

A.4 Three-Class Classification

We include the importance plot generated by the general three-class Eurovision-Sanremo-Melodifestivalen classifier.

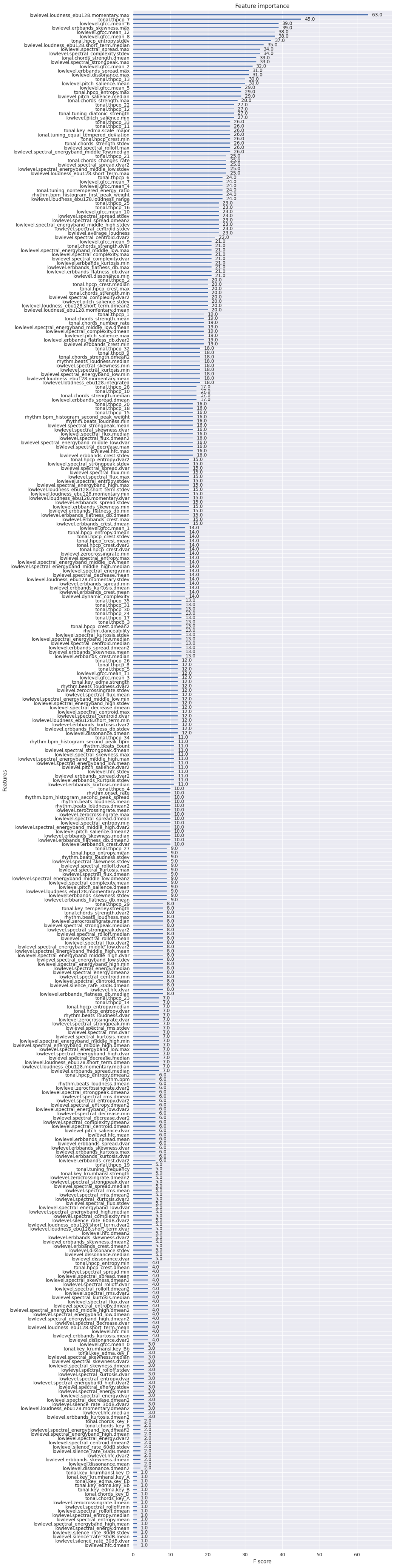


Figure A.47: Importance plot general ESC-SR-MF classifier.