Ben-Gurion University of The Negev
Faculty of Engineering
Biomedical Engineering

Analysis calls of mouse pups as a means of predicting autistic behavior

Supervisors:

Prof. Golan Hava

Dr. Lederman Dror

Students:

Shaked Weinberger - 205868607

Date: 15/08/2022

Table of Contents:

1. Abstract

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder which is affected by environmental and genetic factors. Early diagnosis and identification of ASD is particularly important, it enables early intervention in diagnosed children and hence improvement in their behavioral development. Today diagnosis is based on professional behavioral observation, a limited tool since it is subjective, imprecise, and based on analysis of behavior and skills that develop in later stages such as speech and communication.

This research focuses on a mice model for autism, to investigate the biology of the disorder, aiming at developing a new diagnosis tool. Mice are good models for researching autism because they are highly social animals. The mice emit specialized sounds in the ultrasonic range (USV) to communicate in different social contexts. Using analysis of the USV in mice, allows to elucidate and deepen the understanding of the biological mechanisms that cause abnormal social development that are seen in ASD.

To date, preliminary algorithms for segmentation and classification of syllables have been developed and tested. However, they were implemented on different platforms, limiting their practical usage. The goal of this project is to develop a machine learning model that reads audio files containing mice vocalization and identifies whether a mouse is healthy or suffers from ASD-like signs. This project focuses on developing a unified model for identifying autism, from the segmentation stage to the final diagnostic stage. This unified model allows us, for the first time, to test the diagnostic capabilities on data that have not been manually segmented and classified for syllables. The initial results of the unified model showed a final classification accuracy of 79% in identifying whether a mouse is healthy or suffering from ASD-like signs.

2. Introduction

2.1 ASD

ASD refers to a group of neurodevelopmental conditions that are characterized by a wide spectrum of symptoms, skills, and functional levels [1]. Impairments are identified in two domains: communication and social interaction and restricted, repetitive patterns of behavior [2]. ASD is a prevalent condition, with 1 in 68 children identified with symptoms, e.g., developmental delay, communication and social difficulty, that characterize the disease in the first two years of life [1].

Observing abnormal behavior is the basis for a diagnosis, but due to the heterogeneity of symptoms and their severity in ASD, the diagnosis in children may be in late ages [2]-3]. Despite the availability of reliable methods for diagnosis of ASD before the age of two in severe cases, the average age at which a child is diagnosed is currently at 4-5. Early diagnosis of ASD is particularly important because it allows early intervention and treatment in a time when brain circuits are still plastic. Early detection allows for appropriate medical research, counseling, and intervention during child development, which can include treatments, appropriate educational programs, and appropriate medication that can alleviate symptoms. All of these have long-term effects on the child's daily abilities in the areas of language, interpersonal relationships and behaviors, which significantly facilitates their quality of life and family life, as well as their integration into society [4]. Studies found that interventions implemented before age 4, lead to improvements in daily living skills, social behavior, language, and quality of life of children with ASD [1].

The exact cause of ASD is still unknown, although it is believed to have both genetic factors and environmental factors. ASD involve multiple genes and demonstrate great phenotypic variation. One of the genes associated with the increased risk of autism is methylenetetrahydrofolate reductase (MTHFR) whose activity is related to folic acid metabolism. MTHFR deficiency increases the risk to create developmental delay, and autistic symptoms in both humans and mice [4].

2.2 Diagnosis of ASD in children and infants

The importance of early intervention for children with ASD has resulted in attempts to quantify behaviors in infancy that may lead to early detection. Studies show that delayed production of canonical syllables in infants may be an early warning sign of ASD [5]. Another study has shown that siblings of children diagnosed with ASD, considered high risk for ASD, produced significantly fewer speech-like vocalizations, fewer consonant types and fewer canonical syllable shapes compared to children from low-risk families [6].

Studies that have examined speech in children with ASD have revealed that a considerable number of children with ASD exhibit expressive language delays. Of those who do develop speech, many exhibit slower speech rate, short sentences, higher speech intensity, repeating words or phrases for no apparent reason and higher and more variable fundamental frequency (F0) [5].

Other studies examined differences in acoustic characteristics of infant cries in a sample of babies at risk for autism and a low-risk comparison group, it has been shown that at risk infants produced pain-related cries with higher and more variable fundamental frequency (F0) than low-risk infants. These results provide preliminary evidence that disruptions in cry acoustics may be part of an atypical vocal signature of autism in early life [6].

These and other studies indicate that signs of ASD can be detected in babies even before the emergence of verbal speech, and that there are parallel lines between speech and crying signals in children and infants and USV signals in mice in the context of ASD.

2.3 UltraSonic Vocalization

Studies that use rodents, and especially mice, explore specialized sounds in the ultrasonic range called UltraSonic Vocalizations (USV) that are emitted in different social contexts [7]. Using the mouse model of autism helps to elucidate and deepen the understanding of the biological mechanisms that cause abnormal development that are seen in ASD, and develop treatment [4]. In mouse pups, production of USVs is affected by maternal isolation, cold, rotation and genotype and hence the standard test for vocalizations in mice is the ultrasonic distress call of pups separated from the mother or removed from the nest [8-9]. These calls are used for investigate the number of calls and the type of syllables emitted by separated pups in mouse models of autism spectrum disorders [9].

Mice communicate both in the human-audible range with squeaks for long-distance warnings, and in the ultrasound range for short-distance communication, with calls in frequencies ranging from 25–120 kHz. Mice USV frequencies are especially variable and complex. Mice USVs consist of several different syllable types, whose temporal sequencing includes the utterance of repeated phrases [8]. The most common approach is to split the USV of the mice into 10 types of syllables which different from each other in their duration, in their main frequencies and in their spectrograms (e.g., flat, short, upward, downward, etc.) [9-10].

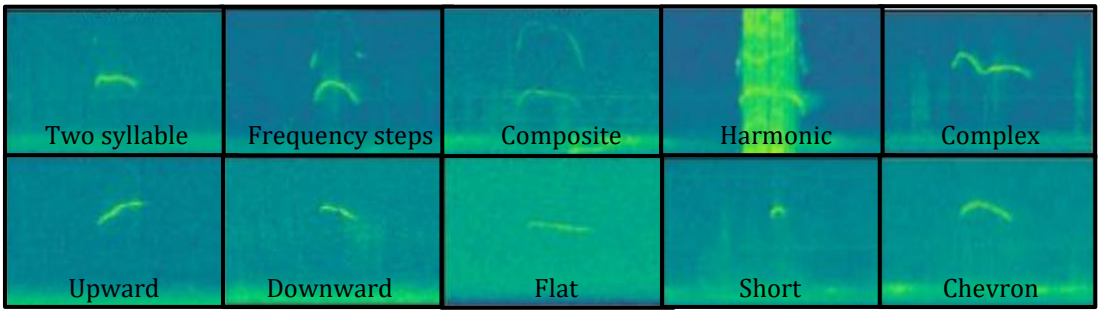

Figure 1: The 10 types of syllables

2.4 Previous works

2.4.1 Manual analysis of syllables

In a previous study conducted at Golan's laboratory, three groups defined by genotype and maternal genotype as follows: Mthfr+/+ (wild-type [WT]) offspring from WT mothers (WT:WT), WT offspring from Mthfr+/− mothers (heterozygote [HT]) (HT:WT) and HT offspring from Het mothers (HT:HT) were created. Analysis of the syllables that the mice from the three groups emitted, they found that pups with maternal Mthfr+/+ (WT:WT) genotype had syllables with higher frequencies (start, end, and mean frequency), compared to pups with maternal Mthfr+/- (HT:WT/HT:HT) genotype. In addition, offspring Mthfr+/- genotype (HT:HT) had lower mean and end frequencies, compared to pups with Mthfr+/+ (WT:WT/HT:WT) genotype. In contrast, the start frequency and bandwidth were lower in the mice with maternal Mthfr+/- genotype and further reduced by offspring Mthfr+/- genotype, such that the Mthfr+/- pups from maternal Mthfr+/- (HT:HT) genotype showed the lowest start frequencies. In the males, the Mthfr+/- genotype at maternal and offspring lowered the frequencies of the start, end and mean frequency of the syllables. In both sexes, the bandwidth was lowered by maternal Mthfr+/- genotype and further lowered by offspring Mthfr+/- genotype.

In summary, it was found that the start frequency, end frequency and bandwidth were the variables that were most significantly affected by the ASD factor in most types of calls and bandwidth and duration showed the largest differences between ASD and control groups.

From an analysis of the proportional production of the different syllables it was revealed that the male offspring of maternal Mthfr+/- genotype (HT:WT and even more in the HT:HT pups), was a lower proportion of syllables with high frequency in syllables of both simple or complex structure and also maternal Mthfr+/- genotype was associated with a marked increase in the proportions of low frequency calls. Altogether, Mthfr genotype decreased the proportion of syllables at the range of 70,000kHz and above, and increased the use of syllables with a range around 50,000kHz. In addition, it was found that the syllables emitted by the experimental groups were shorter, simpler and with lower frequency syllables compared to the experimental groups [11].

2.4.2 Syllable classification

Analysis and understanding of USV signals in mice can give a lot of information about one of the most striking features of autism - communication. Manual analysis of the USV signals includes manual classification of the syllables into the different categories that requires a lot of manpower and time. In addition, the classification is subjective and may vary depending on the classifier.

In recent years, attempts have been made to classify syllables into fewer and more general categories as well as to classify using software, in order to improve the ability to detect alterations in model pups [11].

For instance, in 2017 Segbroeck et al. [12] developed a model for classification of USV into 4 distinct categories. Their approach is based on k-means clustering with a fixed, pre-defined number of clusters. The authors reported around 85% classification accuracy [12]. However they struggled to classify USVs with real-life environmental noise (e.g., mechanical rattling, knocking, and movement by the animals) in some cases [13].

In 2019, Coffey et al. [13] proposed "DeepSqueak", a model for segmentation and classification of USVs. This model uses a Regional Convolutional Neural Network architecture, termed "Faster-RCNN", that allow detection for different USV frequencies made by different rodent strains. This algorithm showed classification accuracy of approximately 95%. However, the classification model was designed for 5 syllables only [13].

The most promising advancement was done by Fonseca et al. [14] who introduced "VocalMat". They used a different set of categories composed of 11 different types. "VocalMat" is a software which provides accurate segmentation and classification based on a CNN (Convolutional neural network). The accuracy of the "VocalMat" classifier was 86% but it struggled classifying some of the syllables type that contain more than one component [14].

In last year's project, the students developed a model for classifying the syllables into the 10 most well-known categories in literature. This model based on CNN that has been trained on a large dataset. The model achieved average accuracy of around 81% [15].

Table 1: USV classification tools comparison

	MUPET [12]	**DeepSqeak** [13]	**VocalMat** [14]	**BGU Project** [15]
Year	2017	2019	2021	2021
Main method	K-mean clustering	Faster RCNN	CNN	CNN
USVs dataset size	~51,000	~56,000	~45,000	6000~
Syllable categories number	60-200 Depend on user-defined	100 Blink to the smallest possible number of categories	11	10+others
Average accuracy result	85%	95%	86%	81%
Limitation	Clustering is applied on the spectral magnitude of the segmented syllables so that amplitude changes can impair classification to the different categories	It may be classified into a small number than wanted of categories	Shows lower accuracy for detecting USVs with multiple components	Small dataset

It should be noted that the above studies aimed at classifying the USV calls into categories, whereas our main interest is in diagnosis, i.e., a tool that can classify the USV calls into “healthy” and “ASD-like”.

2.5 Project Goals

Previous projects focused on developing models of segmentation and classification that have yielded good results. In this project, our goal is to take the research one step forward, and use the models that were developed, to process entire USV signals, characterize them, extract features that may be relevant to autism diagnosis and develop a machine-learning-based model that can use the signals to identify whether a mouse is healthy or suffers from ASD-like features.

The project goals can therefore be stated as follows:

1. Develop an end-to-end model, i.e., a unified model for USV signal processing from the segmentation stage to the final diagnostic stage.
2. Test the diagnostic performance on recorded files that have not been manually analyzed.

3. Methods

3.1 The dataset

In this work we used a dataset that has been collected at Golan's laboratory in recent years. The dataset contains 7923 audio recordings of 70 young mice USVs, males and females, where some of them are healthy and others have with MTHFR gene deficiency. We use recordings of young mice, as the goal is early detection of ASD.

The dataset includes mice from two different breeding schemes. This year we recorded 23 mice pups (3413 audio recordings) from breeding of 2 strains: BALB.C X C57B6, the remaining 47 mice (4510 audio recordings) belong to the BALB.C strain.

For recording the USV signals we used a CM16/CMPA microphone with frequency range of 2 kHz – 200 kHz and Avisoft-RECORDER, a versatile multichannel triggering hard disk recording system that is designed for the special needs of bioacoustics. The highest frequency of USV is around 120 kHz, therefore the sampling frequency was set to 250 kHz. The trigger for recording is based on the appearance of 0.5% of the signal's energy in the frequency range of 10-250 kHz. On each trigger, the software saves 5 secs before the trigger

starts and 5 secs after the trigger ends. All data were saved as audio signals of WAV format and were organized in folders to the mother's name, mice's name, age and session.

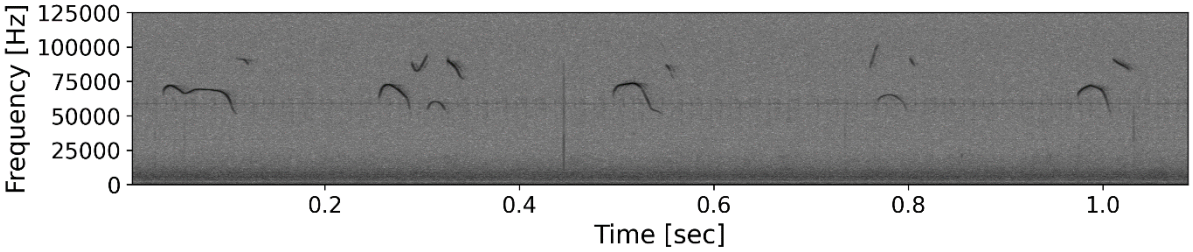

Figure 2: An example of recordings file

The dataset we used contains 7923 audio files with a total of around 60,000 syllables from mice pups differing in strain, gender, genotype and age. The classification model was not trained with a "noise" class. Therefore, to avoid errors such as classifying noise as a syllable, a threshold was set. The threshold requires the probability of a syllable classification to be over 50% certainty, otherwise the syllable is classified as "undefined"[15].

Table 2: Number of syllables for the different groups

Strain	**Age**	**Male**			**Female**		
		WT:WT (N = 8)	**HT:WT (N = 1)**	**HT:HT (N = 1)**	**WT:WT (N = 4)**	**HT:WT (N = 3)**	**HT:HT (N = 6)**
BALB.C X C57B6	**4 days**	2388	392	271	1645	1486	2527
	6 days	5494	381	679	2475	2326	4457
BALB.C		**WT:WT (N = 4)**	**HT:WT (N = 5**	**HT:HT (N = 5)**	**WT:WT (N = 12**	**HT:WT (N = 13)**	**HT:HT (N = 8)**
	4 days	583	983	0	948	892	200
	6 days	1783	509	2717	3058	4198	1348
	8 days	750	1086	2372	833	6233	1991
	10 days	102	149	451	1058	713	266
	12 days	39	181	62	416	790	451

The following figures present a visualization of the data - syllables distribution in males and females with different genotypes and at different ages for each of the strains. At each age, the syllables from all the recorded sessions are included:

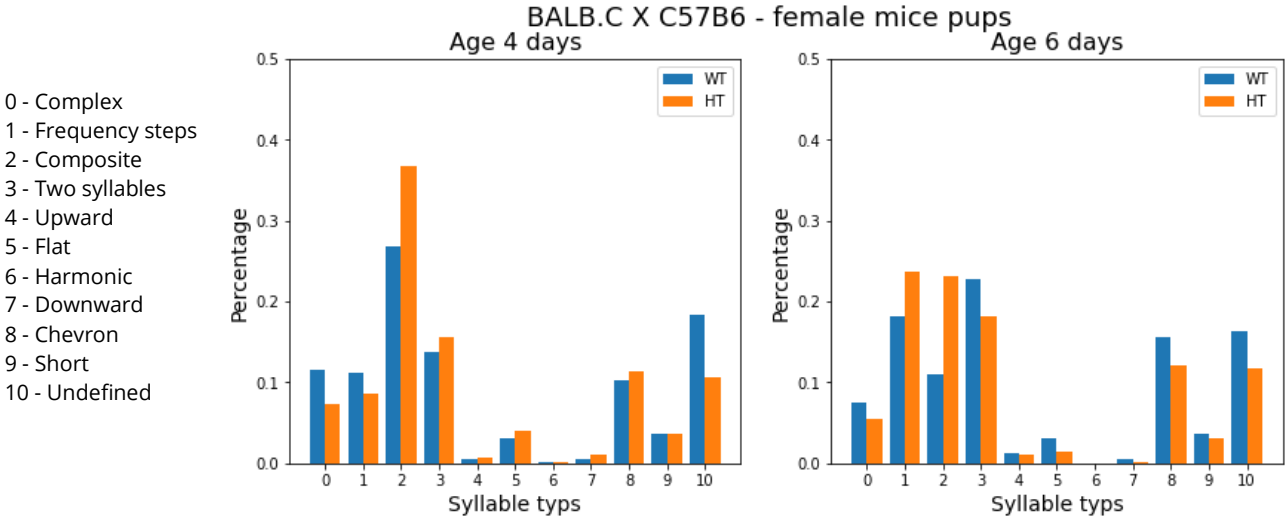

Figure 3: Syllables distribution of BALB.C X C57B6 strain, female comparison by age and genotypes

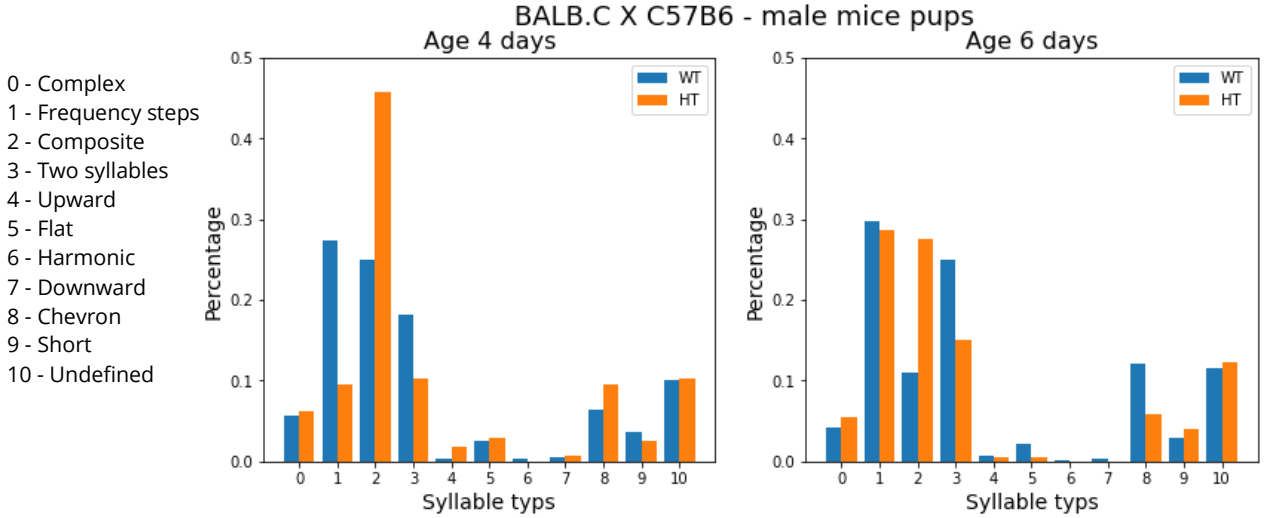

Figure 4: Syllables distribution of BALB.C X C57B6 strain, male comparison by age and genotypes

Figure 3 shows the distribution of syllables for females of the BALB.C X C57B6 strain. The y axis represents the percentage of each syllable out of all syllables recorded for each group. The x axis shows the ages. The blue bars represent the percentage of syllable usage in WT genotype, the orange represent the percentage of syllable usage in HT genotype.

At the age of 4 days, we can see that female pup of WT genotype mainly use "Frequency steps", "Composite" and "Two syllables", all of which are multicomponent syllables. HT genotype mainly uses "Composite" at this age. At the age of 6 days, we can see that both

genotypes mostly use "Frequency steps", "Composite", "Two syllables" and "Chevron" and no major differences are seen by eye.

Figure 4 shows the distribution of syllables for males of the BALB.C X C57B6 strain. Similar for female pups, at the age of 4 days WT genotype mainly uses "Frequency steps", "Composite" and "Two syllables" and HT genotype mainly uses "Composite", and the age of 6 days both genotypes mostly use "Frequency steps", "Composite", "Two syllables" and "Chevron".

It can be seen from figures 3-4 that both in male and female of all ages the syllable "Composite" was used more in HT genotype.

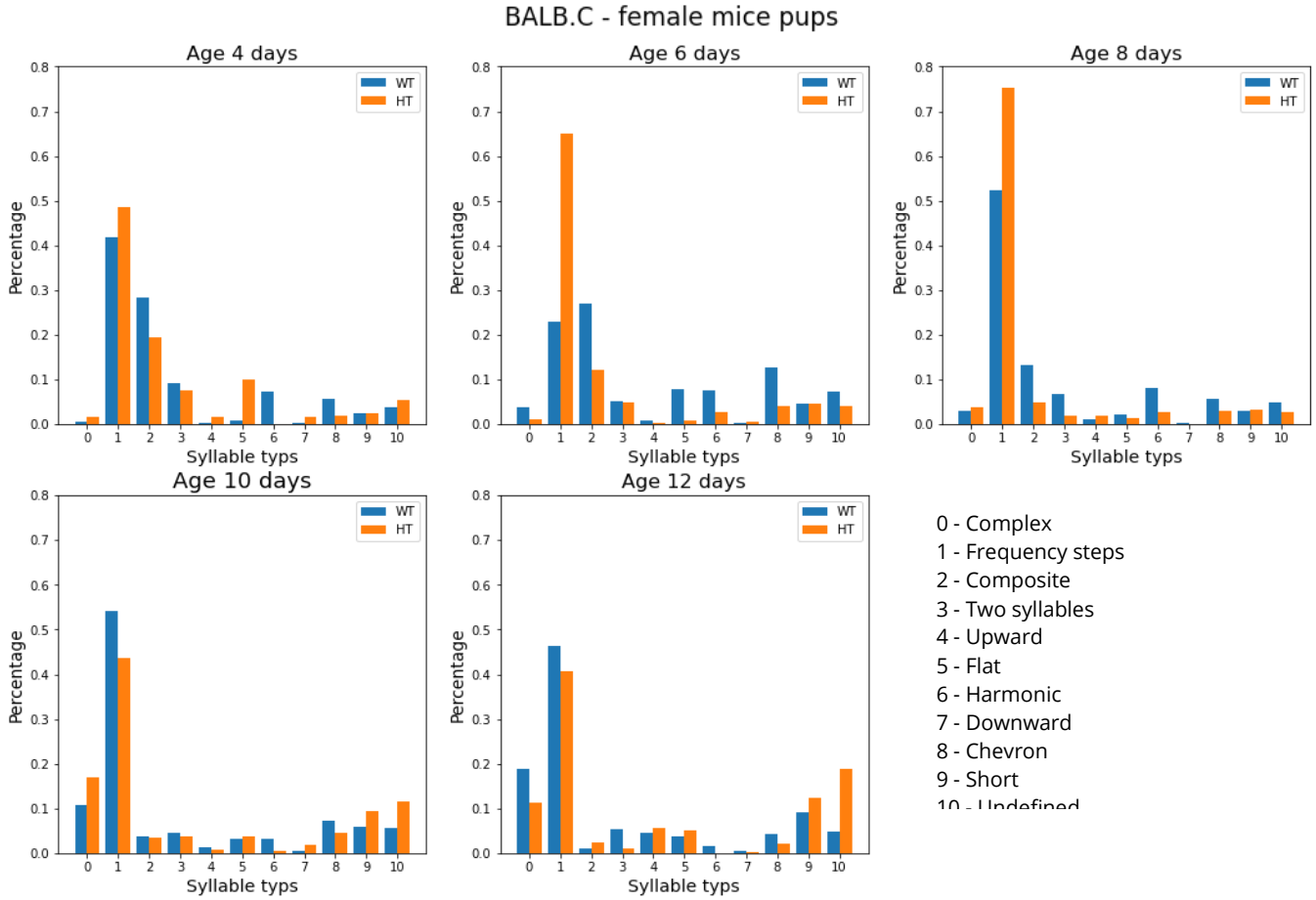

Figure 5: Syllables distribution of BALB.C strain, female comparison by age and genotypes

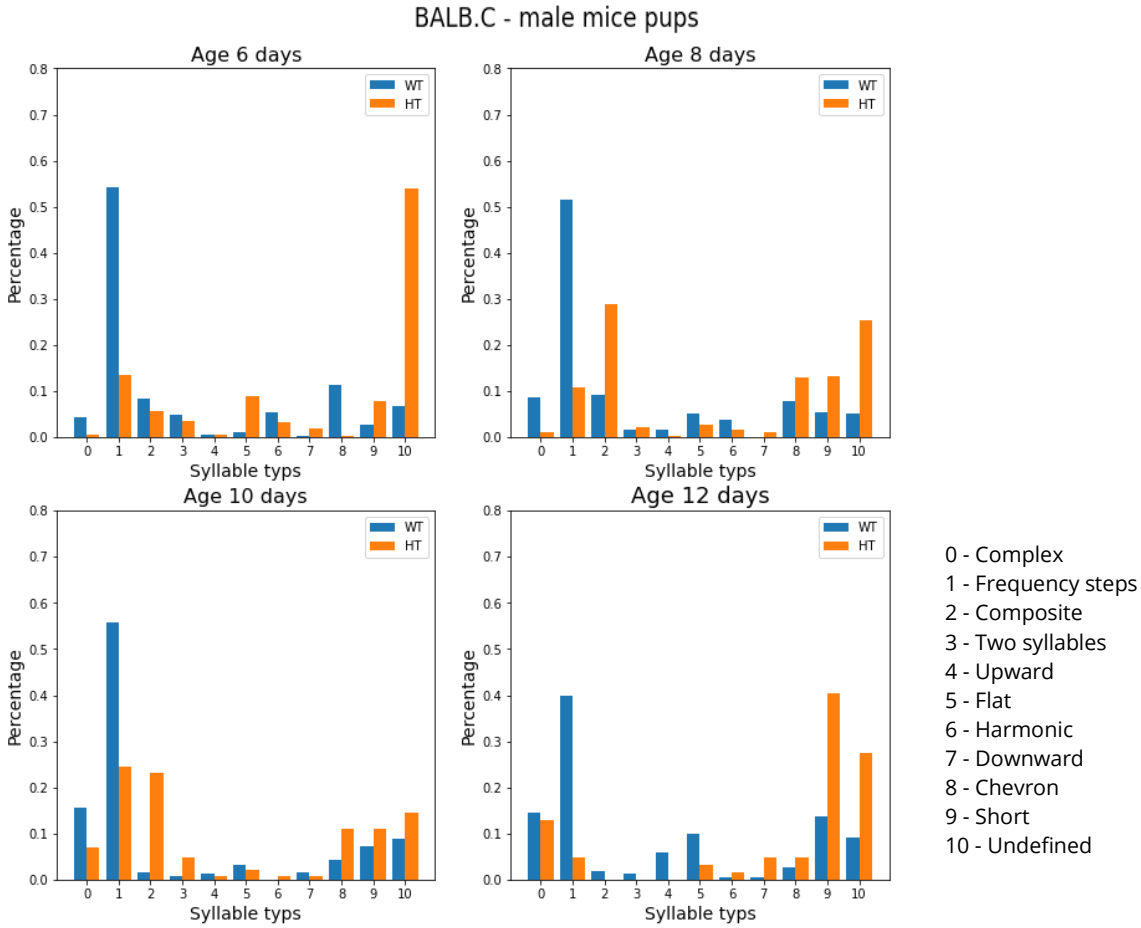

Figure 6: Syllables distribution of BALB.C strain, male comparison by age and genotypes

Figure 5 shows the distribution of syllables for females of the BALB.C strain. It can be seen at all ages for both genotypes that the most usable syllable is "Frequency steps" (except WT genotype at the age of 6 day).

Figure 6 shows the distribution of syllables for males of the BALB.C strain. It can be seen that for WT genotype in all ages the most usable syllable is "Frequency steps", similar to WT genotype females of the same strain. For HT genotype in age of 6 and 10 day the most usable syllable is "Frequency steps", similar to HT genotype females of the same strain. However, for HT genotype in age of 8 day the most usable syllable is "Composite" and for age of 12 day is "Short".

It can be seen from figures 5-6 that except for females at the age of 6 and 8 days the HT genotype has more "Undefined" syllables and it might suggest a developmental delay in communication skills.

The following figures present the mean start frequency, mean end frequency and mean duration for each syllable in males and females with different genotypes of the BALB.C X C57B6 strain.

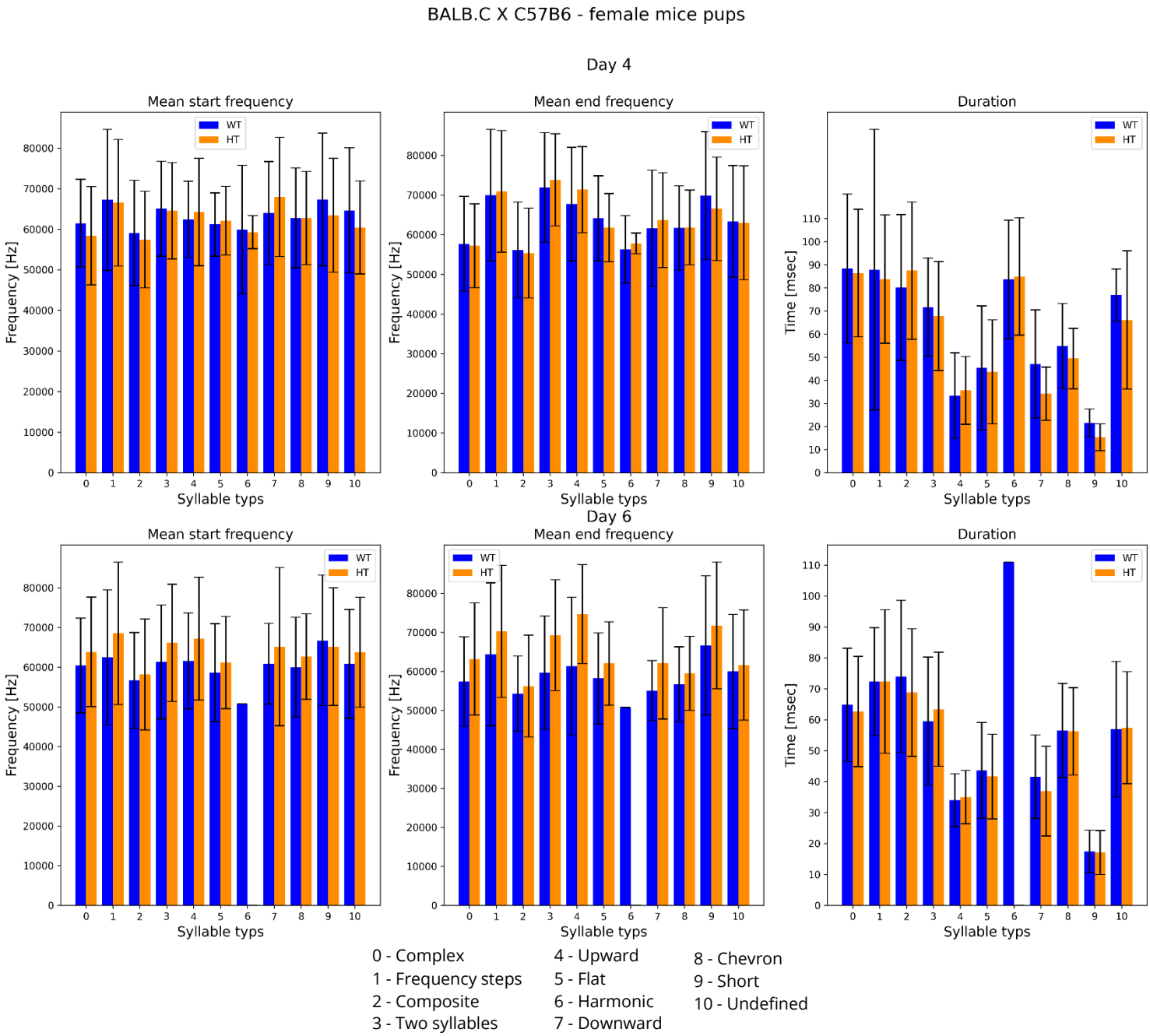

Figure 7: Mean start frequency, mean end frequency and duration of BALB.C X C57B6 strain, female comparison by age and genotypes

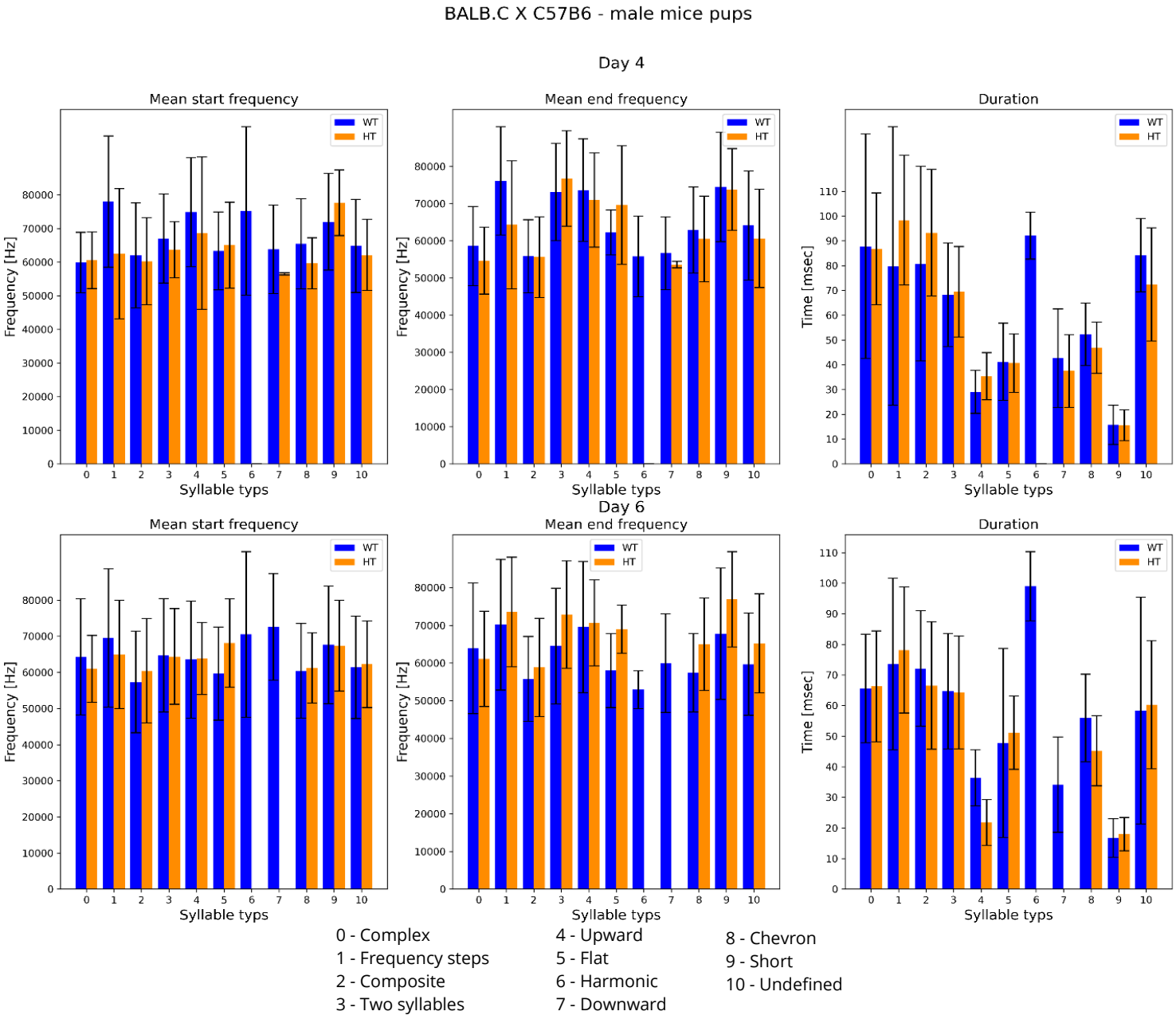

Figure 8: Mean start frequency, mean end frequency and duration of BALB.C X C57B6 strain, male comparison by age and genotypes

Figures 7-8 show that in most cases the syllable frequencies of the HT genotype are higher than the WT genotype, with this difference increasing in age of 6 day. In addition, it can be seen that in most cases the syllable duration of the HT genotype is shorter.

We note that these results consistent with the differences seen in many studies between the speech signals of healthy children and infants and those with ASD. The fundamental frequency of the speech signals of children and infants with ASD is higher. In addition, the speech intensity of children with ASD is higher and they speak in shorter sentences.

3.2 USV signal processing

To complete the task of predicting the genotype of each mouse, we examined several options for action. One of the main was to use models like LSTM (RNN - Recurrent Neural Network) that work on a whole sequence of data such as the direct audio recordings themselves, without the need for processing done earlier by the previous teams (an "end-to-end" approach). We chose not to use this approach, in order to maintain continuity between projects as part of the study and to enable the use of insights that arose in previous studies, and to use simpler models that allow us more control over possible failures along the way.

Figure 9 presents a block diagram of the proposed scheme. The scheme consists of pre-processing, feature extraction and classification.

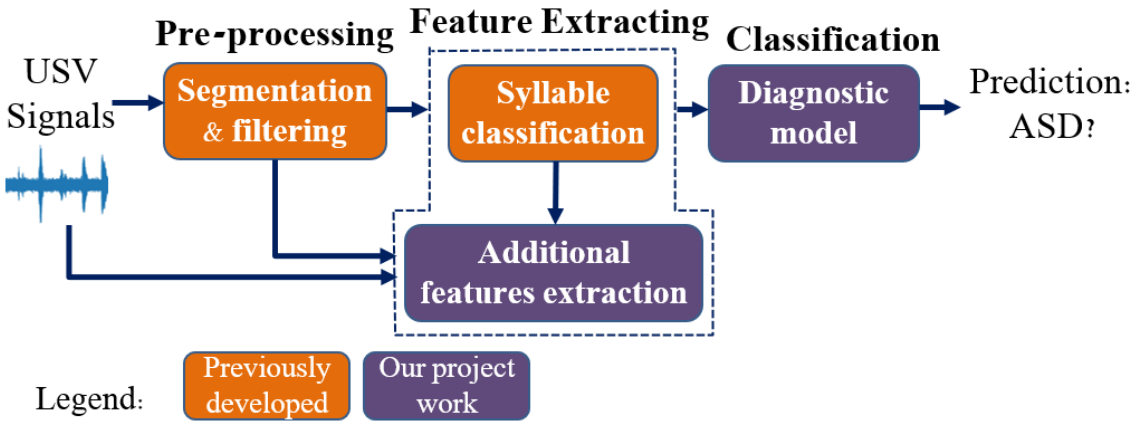

Figure 9: A general scheme of the proposed system

3.2.1 Pre-processing

The pre-processing stage includes a noise reduction filter and a segmentation algorithm which divides the entire USV signals into short segments which each includes one syllable and removes the silence segments.

Inspecting the signals revealed that the syllables characterized by relatively high energy concentrated in a narrow frequency band. Therefore, the segmentation algorithm is based on energy filter bank coefficients. A filter bank is an algorithm that separates the input signal into multiple components, each one carrying a single frequency sub-band of the original signal [16].

The segmentation algorithm developed in previous projects was implemented in MATLAB. In order to develop a uniform model, we converted the segmentation code to Python. When we segmented the new signals that recorded this year, we noticed that the

algorithm was not working properly. After testing a few signals, we found a noise that appears at a specific frequency along the entire signal, so we used BSF (Band stop filter) to filter the noise. We checked the Python segmentation code on about 800 syllables, and we accepted MSE of 0.0001 between algorithm's segmentation and manual segmentation.

Figure 10 shows the spectrum of the signal before and after filtering. In figure (a) the noise can be clearly seen, and figure (b) shows that the noise has been removed.

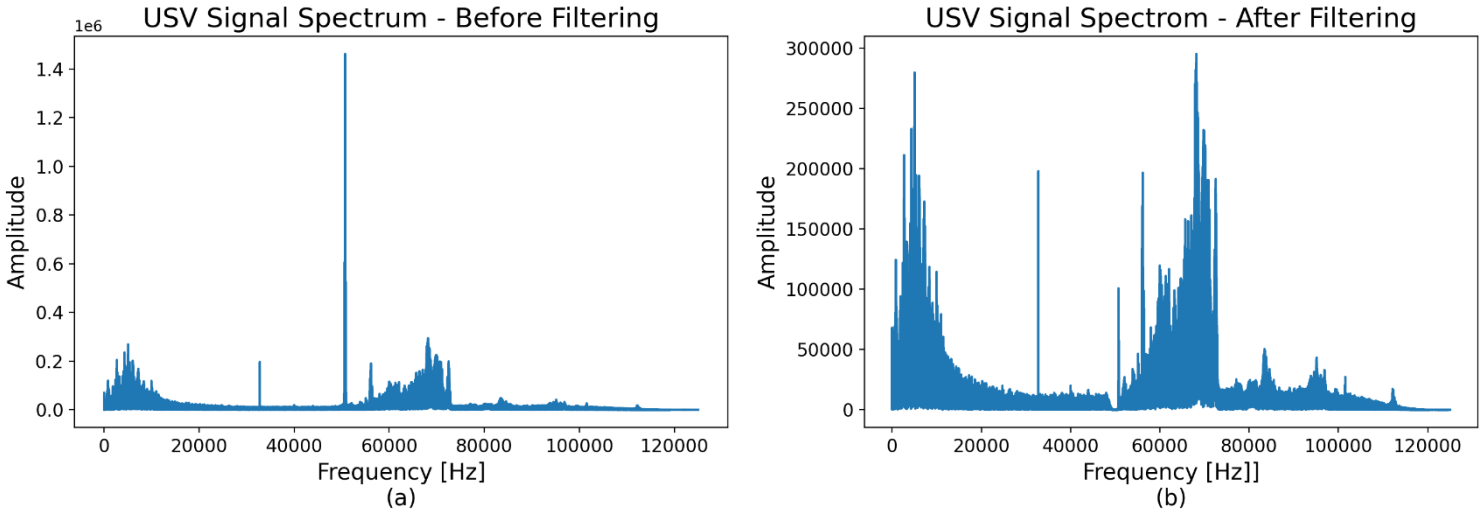

Figure 10: Signal spectrum (a) before and (b) after filtering

3.2.2 Feature Extraction

During the feature extraction stage, diagnostic information is extracted from the USV signal. As part of the transition to a unified and automated model, an automatic calculation of the features was needed. For example, an automatic calculation of the starting and ending frequencies of each syllable was needed, for this we developed a PSD (Power spectral density) based algorithm.

In the initial stage, we based the finding of the features on a manual analysis of the signals and statistical differences that were found in Golan's laboratory. As a second step, we used the outputs of the previous models and identify additional important features and removed irrelevant features. The model uses 10 prominent features that represented by 46 numbers, the features are:

- Syllable distribution - the percentage of appearance of each syllable per audio file (Represented by 10 numbers).

- Syllable mean start frequency - for each syllable type per audio file (Represented by 10 numbers).
- Syllable mean end frequency - for each syllable type per audio file (Represented by 10 numbers).
- Syllable mean duration - for each syllable type per audio file (Represented by 10 numbers).
- Mother genotype - 0 → HT, 1 → WT (Represented by a single number).
- Mean time between syllables - mean of the duration of the silent segments between the syllables per audio file (Represented by a single number).
- Sex – 0 → female, 1→ male (Represented by a single number).
- Strain – 1 → BALB.C X C57B6, 2 → BALB.C (Represented by a single number).
- Age – 4,6,8,10 or 12 days (Represented by a single number).
- Session – recording session 1 or 2 (Represented by a single number).

For the calculation of some of the features, we used the syllable classification algorithm that was developed in previous projects. This model is based primarily on deep learning algorithms for identifying the type of syllables by representing the auditory signals using spectrograms and classifying them using a deep convolution network. The algorithm achieved accuracy of about 81%.

Once the features for each recording are calculated, we generated a matrix of features that we used at the classification stage. The matrix of feature contains a column for each feature and a row for each recording.

3.2.3 Classification

Classification is a systematic arrangement in groups or categories according to established criteria specifically. In this project, in addition to finding the features, we focused on the development of a model that receives a matrix of features and can identify whether a mouse is healthy or suffers from ASD-like features.

Because the information we receive at the end of the process is obtained in the form of a table, we implemented algorithms that allow the properties to be extracted accordingly, and we looked for suitable models for working with tabular information.

We choose to use XGBoost algorithm [17] that has recently been dominating in the world of applied machine learning for structured (tabular) data. In addition, XGBoost algorithm

does not have many hyper-parameters, so it is suitable for the small dataset we have. Moreover, it is fast and easy to use. XGBoost is based on decision trees as each tree learns from the tree in front of it and corrects its errors [17]. Figure 4 show A general architecture of XGBoost algorithm.

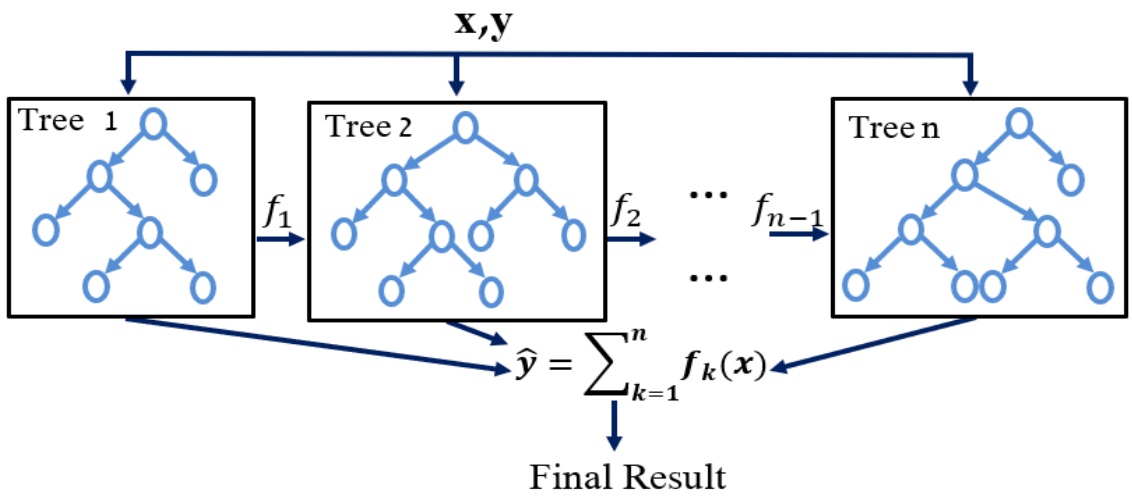

Figure 11: A general architecture of XGBoost [18].

3.2.4 Performance evaluation

For model development and performance assessment, we divided our data into 'training', 'validation' and 'testing' sets. The training set was used to train the models, the validation set was validation and used for hyper-parameter selection, and the test set was used for performance evaluation.

- Training set - contains 4753 recordings files (60%).
- Validation set - contains 1585 recordings files (20%).
- Test set - contains 1585 recordings files (20%).

Table 3 shows the hyper-parameters of the XGBoost model and the values we chose for them. The values are selected after using tools such as grid search on a wide range of values and through trial and error.

Table 3: XGBoost hyper-parameters

Hyper-parameter	Description	Numerical value
n_estimators	The number of decision trees that the model produces in the training process	50
learning_rate	Step size shrinkage used in update to prevents overfitting	0.1
max_depth	Maximum depth of a tree	5
reg_lamda	L2 regularization term on weights	1.5
reg_alpha	L1 regularization term on weights	0.05
min_child_weight	Minimum sum of instance weight (hessian) needed in a child	0.1
scale_pos_weight	Control the balance of positive and negative weights, useful for unbalanced classes	0.8
colsample_bytree	The subsample ratio of columns when constructing each tree	0.6

4. Results

It is important to note that since we do not have enough mice in the dataset:

1. We treat each recording as independent. It is clear that this assumption is not valid, but it is inevitable.
2. There is no separation between the recordings of the mice in the different sets, which means that there can be recordings of the same mice in both the training set and the validation/test set. Separating the mice causes instability of the model, probably due to the small dataset at hand.

4.1 Preliminary results

First, we ran the unified model on a small dataset that included 3752 recordings files from 66 mice. The training set contained 2250 recordings and the validation set contained 751 recordings.

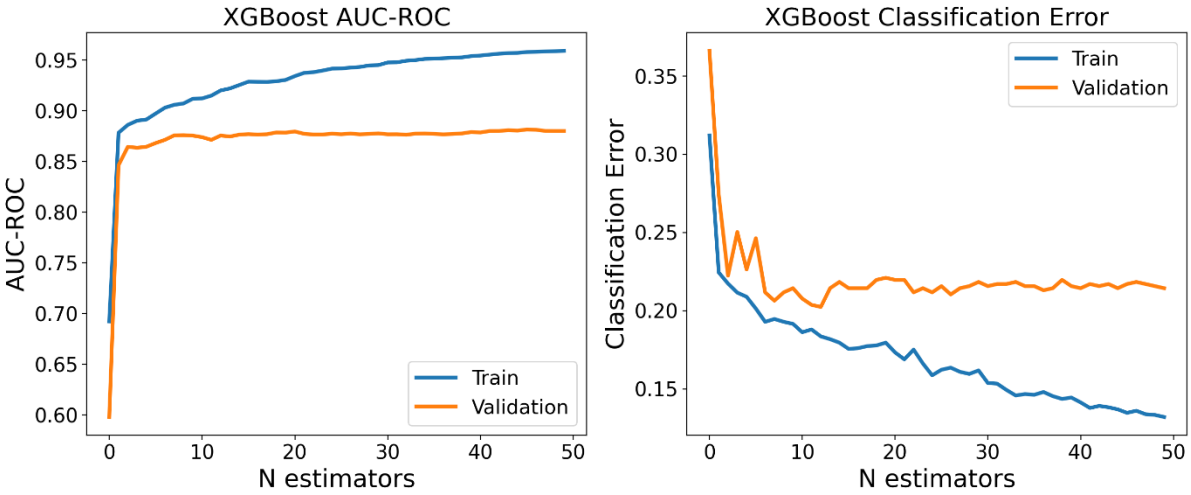

Figure 12: Accuracy and error graphs – small dataset

The right graph shows the classification error as a function of the number of estimators, i.e., the number of trees the model produces. It can be seen that for the training set the model converges and the error decreases as a function of N, on the other hand for the validation set the error remains more or less constant starting from a certain point. In addition, a small overfitting can be noticed, which is probably because the dataset is not large enough and that there is no separation between the recordings of the mice in the different sets. Which means that there can be recordings of the same mice in both the training set and the validation/test set.

The left graph shows the accuracy as a function of the number of estimators, while similarly for the training set the accuracy increases as a function of N and for the validation set it remains constant starting from a certain point.

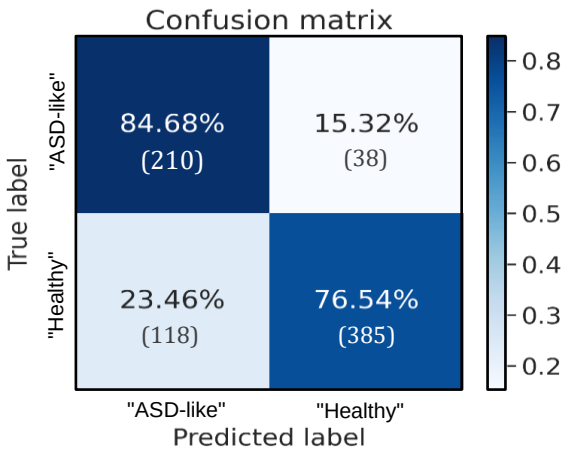

Figure 13: Confusion matrix for test set

Figure 13 shows the confusion matrix for the test set that included 751 recordings. The results indicate the percentages of correct and incorrect classification by category, where the columns of the matrix represent the classification results of the model and the rows represent the real data. It can be seen that 76% of the recordings of the healthy mice were correctly classified and 84% of the recordings of the sick mice were correctly classified. The average accuracy for the test set is 79%.

4.2 Increasing the dataset size

In the second step we increased the dataset size by using recordings that were not manually analyzed. Now the dataset contains 7923 recordings files of 70 mice. The training set contained 4753 recordings and the validation set contained 1585 recordings.

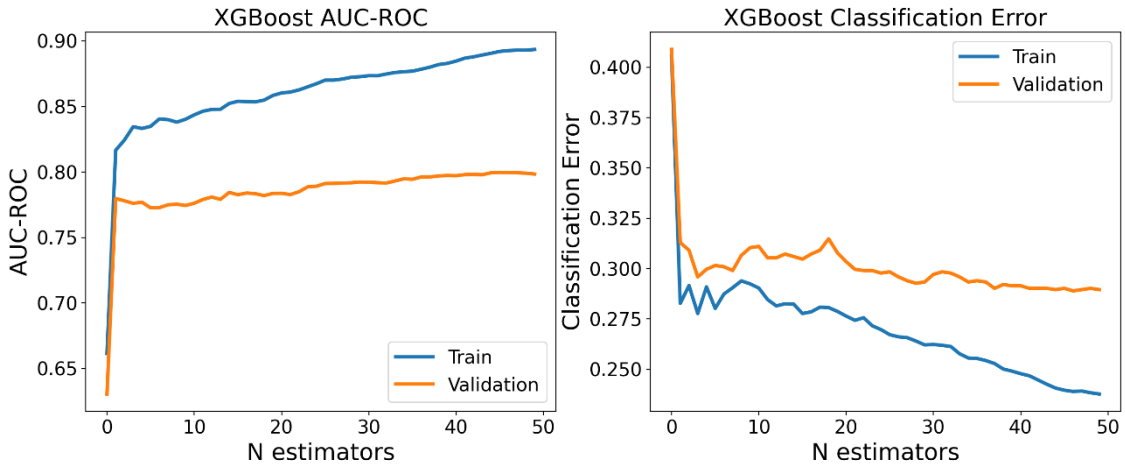

Figure 14: Accuracy and error graphs

It can be seen that as a compared with the use of small data, the classification error sighnificantly increased, maybe because there is an increase in the number of samples of mice from a group HT:WT (WT offspring from Mthfr+/− mothers [HT]). These mice are defined as "healthy", but the mother's genotype appears to have an effect, which is expressed, among other things, in the USV signals.

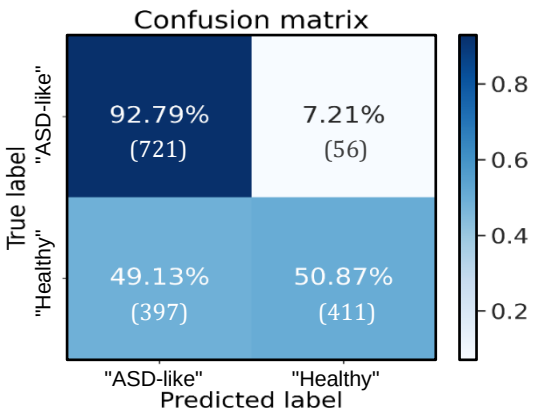

Figure 15: Confusion matrix for test set

The confusion matrix shows that there is a decrease in the percentage of recordings of the healthy mice that were correctly identified, 50% of the recordings were correctly identified, while there is an increase in the percentage of recordings of the sick mice that were correctly identified, 92% were correctly identified. The average accuracy is 70%.

4.3 Additional improvements

Finally, we added three new features: the age of the mouse, the session in which the mouse was recorded and the strain of the mouse (we use recordings of two different strains). In addition, the training set was rebalanced using sample weight.

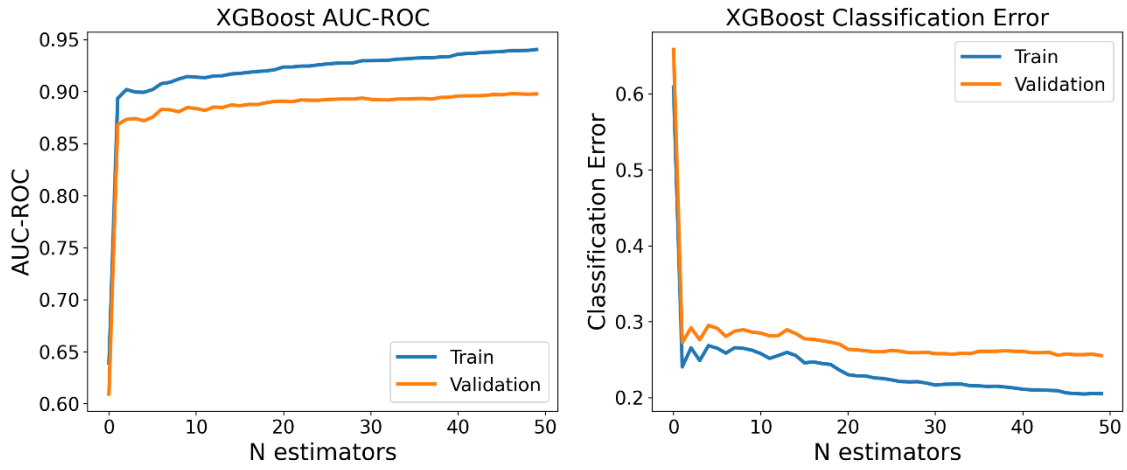

Figure 16: Accuracy and error graphs – new features

In figure 16 it can be noticed that the error decreases and the accuracy increase as a result of the addition of the new features, and that there is a decrease in overfitting that we saw in the previous results.

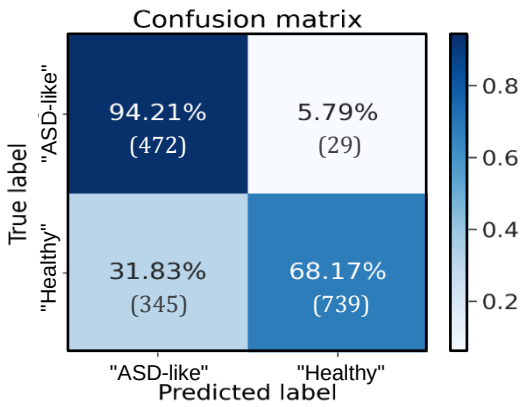

Figure 17: Confusion matrix for test set

This result shows that there is an increase in the percentage of recordings of the healthy mice that were correctly identified, 68% of the recordings were correctly identified, while the percentage of recordings of the sick mice that were correctly identified remained similar, 94% of the recordings were correctly identified. The average accuracy is 76%.

4.4 Feature importance

One of the most significant features that the model uses is the mother's genotype. For comparison with the results we obtained, we trained an identical model to predict the mouse genotype based on mother's genotype alone. The results obtained were 65% accuracy for the test set. The result is worst as compared with our prediction results. We use of features extracted from the analysis of the recordings, hence their added value. Another illustration of this fact can be provided by identifying the most significant features in the classification process.

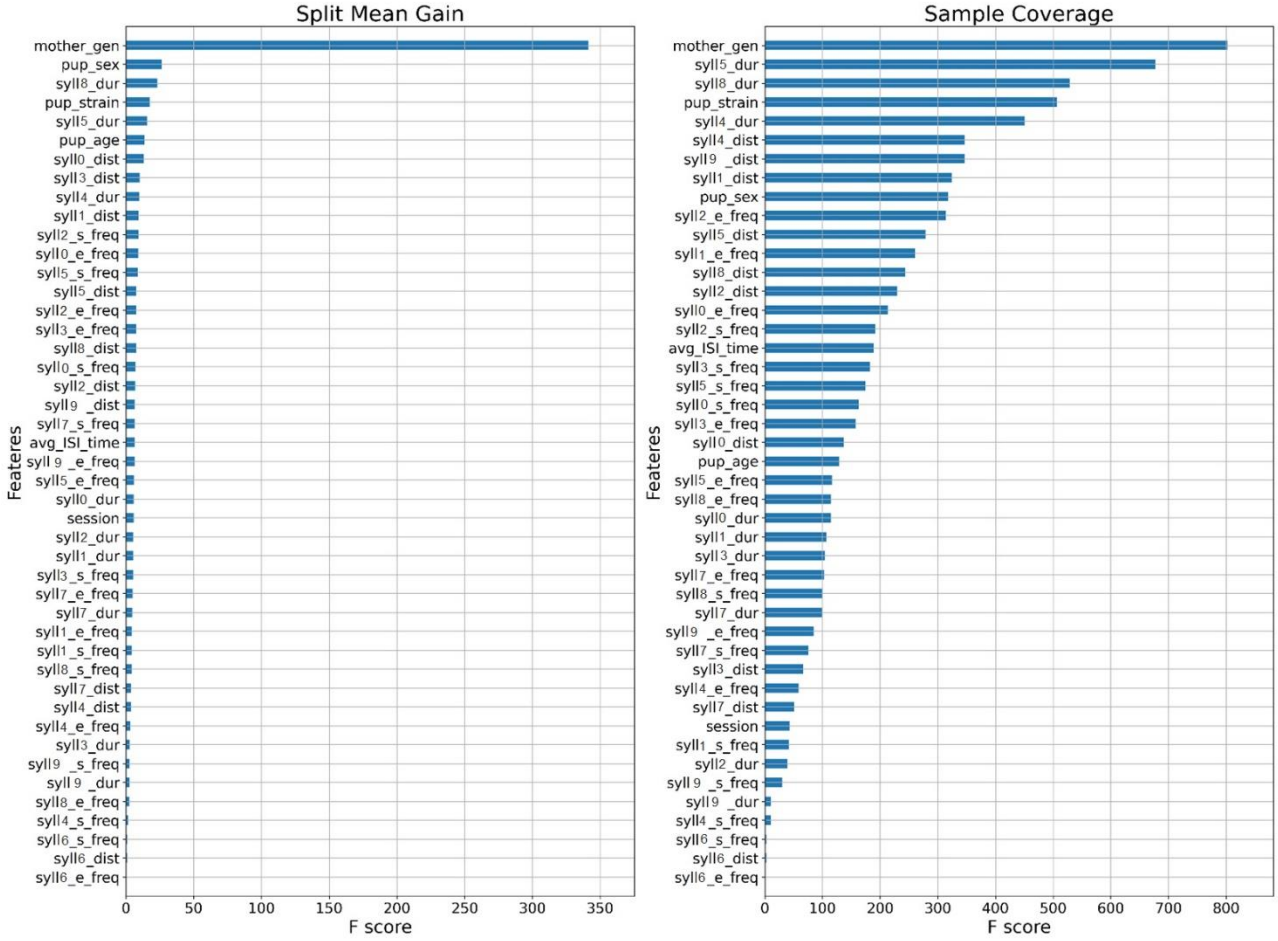

Figure 18: Feature importance

Feature importance tools refers to techniques that assign a score to input features based on how useful they are for generating a prediction. The 'Gain' implies the relative contribution of the feature to the model, calculated by taking each feature's contribution for each tree in the model. A higher value of this metric when compared to another feature implies it is more important for generating a prediction. The 'Coverage' metric means the relative number of observations related to this feature [19].

It can be clearly seen from figure 18 that the most significant characteristic is the mother's genotype. In addition, there are characteristics that were not important at all - either because they do not contain useful information for classification, or because not enough samples containing them have been collected (For example syllable 6 - "Harmonic", that can be seen in figures 3-6 which is not much of it).

5. Discussion

Early detection of ASD is a critical parameter in the context of treating ASD and improving the patient's life. Therefore, many efforts are invested in this area in several research directions. Our research focuses on a mice model for autism and one of the most prominent features of autism - communication.

This research is carried out with the understanding that there are differences between the speech signals of children with ASD and healthy children, as well as differences in the crying signals of infants. These differences are expressed in the context of frequencies, rhythm, intensity and repetition similar to the features on which we based our model.

In this project we developed a unified and automated model for identification of ASD in mice through USV signals processing was developed. This model allows, for the first time, to use raw data that have not been manually processed before. It means there is no need to manually segment the signals and classify the syllables into categories, as has been done to this day.

Currently, the model is mainly used to improve and efficient the pre-clinical study of diagnosis and treatment of ASD. To date, USV analysis required tedious and time-consuming manual labor, and our model provides a fast and efficient automatic tool for USV analysis.

With this tool, it is possible to gather a massive amount of data from the recorded dataset, about the differences in USV emission between the mice genotypes.

It is important to note that the main problem in the project is that there is a small number of mice in the dataset, which causes the instability of the model when the identification is performed per mouse and not per recording as we present.

The results obtained in the project show that there is a future for this research. We hope that this project will be an important step in the further advancement of research and the development of tools to diagnose autism at an early age.

5.1 Future work

As part of the ideas raised throughout the work process and as a result of our conclusions, we offer several ways in an attempt to improve the model:

- Extract additional features and focus on most relevant. For example: the total number of syllables per recording, the time it takes for the mice to start speaking from the moment of separation from the mother, etc.
- Examining the model parameters and finding the best parameters for the model.
- Convert the model from subject-dependent to subject-independent, that means making sure all the recordings of each of the mice will be in only one set.
- Increasing the dataset size to make it possible to use deep learning methods and neural networks to implement the model.

6. Summary & conclusions

The goal of this project is to develop a machine-learning-based model that analyzes the USVs signals and identifies whether a mouse is healthy or suffers from ASD-like signs. The process of analyzing the USV signals is done in several steps which were described in the report.

We chose to use the XBGOOST algorithm for feature classification. XGBoost is the most successful model for tabular data. It is based on decision trees and makes it relatively easy to detect errors in the learning process. The model uses features that were calculated from the dataset and were selected after reviewing literature and reading articles on the subject.

After we unified all the algorithms into one model, increased the dataset, extracted features, and trained the model, we obtained a relatively high classification rate for ASD - average of 79%.

7. References

[1] J. Elder, C. Kreider, S. Brasher, and M. Ansell, "Clinical impact of early diagnosis of autism on the prognosis and parent-child relationships," *PRBM*, vol. 10, pp. 283–292, Aug. 2017, doi: 10.2147/PRBM.S117499.

[2] S. L. Hyman, S. E. Levy, and S. M. Myers, "Identification, Evaluation, and Management of Children With Autism Spectrum Disorder," *Pediatrics*, vol. 145, no. 1, p. e20193447, Jan. 2020, doi: 10.1542/peds.2019-3447.

[3] A. Masi, M. M. DeMayo, N. Glozier, and A. J. Guastella, "An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options," *Neurosci. Bull.*, vol. 33, no. 2, pp. 183–193, Apr. 2017, doi: 10.1007/s12264-017-0100-y.

[4] N. Sadigurschi and H. M. Golan, "Maternal and offspring MTHFR genotypes interact in a mouse model to induce ASD-like behavior," *Genes, Brain and Behavior*, p. e12547, Dec. 2018, doi: 10.1111/gbb.12547.

[5] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network," *IEEE Access*, vol. 8, pp. 139489–139500, 2020, doi: 10.1109/ACCESS.2020.3012532.

[6] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical Cry Acoustics in 6-Month-Old Infants at Risk for Autism Spectrum Disorder: Infant cry and risk for autism," *Autism Res*, vol. 5, no. 5, pp. 331–339, Oct. 2012, doi: 10.1002/aur.1244.

[7] S. Agarwalla, N. S. Arroyo, N. E. Long, W. T. O'Brien, T. Abel, and S. Bandyopadhyay, "Male-specific alterations in structure of isolation call sequences of mouse pups with 16p11.2 deletion," *Genes, Brain and Behavior*, vol. 19, no. 7, Sep. 2020, doi: 10.1111/gbb.12681.

[8] H. Wang, S. Liang, J. Burgdorf, J. Wess, and J. Yeomans, "Ultrasonic Vocalizations Induced by Sex and Amphetamine in M2, M4, M5 Muscarinic and D2 Dopamine Receptor Knockout Mice," *PLoS ONE*, vol. 3, no. 4, p. e1893, Apr. 2008, doi: 10.1371/journal.pone.0001893.

[9] M. L. Scattoni, S. U. Gandhy, L. Ricceri, and J. N. Crawley, "Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism," *PLoS ONE*, vol. 3, no. 8, p. e3067, Aug. 2008, doi: 10.1371/journal.pone.0003067.

[10] J. B. Panksepp *et al.*, "Affiliative Behavior, Ultrasonic Communication and Social Reward Are Influenced by Genetic Variation in Adolescent Mice," *PLoS ONE*, vol. 2, no. 4, p. e351, Apr. 2007, doi: 10.1371/journal.pone.0000351.

[11] I. Shekel *et al.*, “Isolation-induced ultrasonic vocalization in environmental and genetic mice models of autism,” *Frontiers*, 2021.

[12] M. Van Segbroeck, A. T. Knoll, P. Levitt, and S. Narayanan, “MUPET—Mouse Ultrasonic Profile ExTraction: A Signal Processing Tool for Rapid and Unsupervised Analysis of Ultrasonic Vocalizations,” *Neuron*, vol. 94, no. 3, pp. 465-485.e5, May 2017, doi: 10.1016/j.neuron.2017.04.005.

[13] K. R. Coffey, R. G. Marx, and J. F. Neumaier, “DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations,” *Neuropsychopharmacol.*, vol. 44, no. 5, pp. 859–868, Apr. 2019, doi: 10.1038/s41386-018-0303-6.

[14] A. H. Fonseca, G. M. Santana, G. M. Bosque Ortiz, S. Bampi, and M. O. Dietrich, “Analysis of ultrasonic vocalizations from mice using computer vision and machine learning,” *eLife*, vol. 10, p. e59161, Mar. 2021, doi: 10.7554/eLife.59161.

[15] G. Lavy and V. Wolfman, “Segmentation and classification of ultrasonic calls as a tool for evaluation of social communication in a mouse model of autism,” *Ben-Gurion University of the Negev*, p. 28, Aug. 2021.

[16] S. R. M. Penedo, M. L. Netto, and J. F. Justo, “Designing Digital Filter Banks Using Wavelets,” *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, p. 33, Dec. 2019, doi: 10.1186/s13634-019-0632-6.

[17] 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*, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[18] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, “A hybrid ensemble method for pulsar candidate classification,” *Astrophys Space Sci*, vol. 364, no. 8, p. 139, Aug. 2019, doi: 10.1007/s10509-019-3602-4.

[19] H. Zheng, J. Yuan, and L. Chen, “Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation,” *Energies*, vol. 10, no. 8, p. 1168, Aug. 2017, doi: 10.3390/en10081168.