scientific reports

OPEN

Predicting the gender of individuals with tinnitus based on daily life data of the TrackYourTinnitus mHealth platform

Johannes Allgaier[1]✉, Winfried Schlee[2], Berthold Langguth[2], Thomas Probst[3] & Rüdiger Pryss[1]

Tinnitus is an auditory phantom perception in the absence of an external sound stimulation. People with tinnitus often report severe constraints in their daily life. Interestingly, indications exist on gender differences between women and men both in the symptom profile as well as in the response to specific tinnitus treatments. In this paper, data of the TrackYourTinnitus platform (TYT) were analyzed to investigate whether the gender of users can be predicted. In general, the TYT mobile Health crowdsensing platform was developed to demystify the daily and momentary variations of tinnitus symptoms over time. The goal of the presented investigation is a better understanding of gender-related differences in the symptom profiles of users from TYT. Based on two questionnaires of TYT, four machine learning based classifiers were trained and analyzed. With respect to the provided daily answers, the gender of TYT users can be predicted with an accuracy of 81.7%. In this context, worries, difficulties in concentration, and irritability towards the family are the three most important characteristics for predicting the gender. Note that in contrast to existing studies on TYT, daily answers to the worst symptom question were firstly investigated in more detail. It was found that results of this question significantly contribute to the prediction of the gender of TYT users. Overall, our findings indicate gender-related differences in tinnitus and tinnitus-related symptoms. Based on evidence that gender impacts the development of tinnitus, the gathered insights can be considered relevant and justify further investigations in this direction.

Many people experience a long-term noise in their ears, which is widely known as tinnitus, also described as a whistling or ringing sound[1] in the ears. About 10–15% of the worldwide population report this kind of symptoms[2,3]. Although many people perceiving tinnitus do not experience a considerable burden, about 2.4% of the worldwide population severely suffers from tinnitus on a daily basis[4]. In most of these cases, tinnitus is a subjective perception that can only be perceived by the affected person. Inversely, rare forms of tinnitus exist, for which the perceived sound is caused by a source in the body that can be objectively measured (e.g., blood flow or muscle contractions). As an important consequence of the discussed aspects, no general treatment, which is able to effectively reduce tinnitus symptoms like loudness and its related fluctuation, exists yet. On the individual basis, tinnitus can be reduced, for example, by the use of cognitive behavioral therapies[5]. To characterize the general status of available treatments with respect to the well-known heterogeneity of tinnitus patients[6,7], they are rare and their development is difficult.

To better and more effectively deal with this heterogeneity, researchers often focus on the identification of subgroups of tinnitus patients. Identified subgroups might be used for investigations on treatments for an identified subgroup instead of a general treatment for all tinnitus patients. However, the clustering of tinnitus patients through the identification of subgroups is not an entirely new research question. Hitherto, several approaches aimed at the clustering of tinnitus patients depending on their symptom profiles[8,9], or depending on neuroimaging data[10]. Furthermore, the authors of[11] developed the `Tinnitus Primary Function Questionnaire` to examine the effect of tinnitus on thoughts and emotions, hearing, sleep, and concentration. The

[1]Institute of Clinical Epidemiology and Biometry, University of Wuerzburg, Wuerzburg, Germany. [2]Department for Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany. [3]Department for Psychotherapy and Biopsychosocial Health, Danube University Krems, Krems an der Donau, Austria. ✉email: johannes.allgaier@uni-wuerzburg.de

authors established correlations between these four effects and derived secondary limitations for the individuals in their daily life. The consideration of potential differences in gender are another approach on subgroup research. A recent special issue shows the latter kind of interest in research[12]. In the already published articles of this special issue, for example, one work deals with gender differences of chronic tinnitus patients[13]. All of the presented works show that gender differences are a valuable research direction in particular and with respect to research on subgroups of tinnitus patients in general. In addition, research evidence exists that the gender impacts the development of tinnitus and the response to treatments. For example, in this recent work[14], the authors investigated treatments of 316 patients and found significant treatment differences between males and females. For instance, females improved better in orofacial therapies. Or, in the work of[15], it was found, among other findings, that stress was positively correlated with tinnitus severity only in males. These and other findings clearly show that gender-related differences are relevant for investigations of tinnitus patients and their symptom profiles.

In the discussed context, the use of mobile applications to monitor health symptoms is becoming more and more popular, also denoted by mobile and digital health (mHealth). With respective mHealth solutions, the collection of data becomes easily possible, especially on a daily basis. Furthermore, data can be collected close to the user's daily life with the goal to foster self-monitoring and eventually may support health care in clinical practice[16]. For example, the authors of[17] monitored and investigated mental health conditions by using a mHealth solution, while the authors of[18] showed the general potential and impact of mHealth applications. For TrackYourTinnitus (TYT), the daily use, among other reasons, enables individuals to be better deal with the variations of the tinnitus over time. On the flip side, mHealth solutions also revealed drawbacks, which are discussed by many recent works. For example, potential discrepancies of app developers and patients of mHealth apps are investigated more in-depth by[19], while general challenges are discussed by[20]. In the discussed setting, it should always be kept in mind that a daily smartphone usage might also worsen the individual tinnitus situation as users are reminded about their problems on a frequent basis. However, research works exist that have shown that the daily use of mobile technology does not aggravate the overall health condition, see for example[21]. Despite such findings, the daily focus on a disease when using mHealth solutions should always be considered carefully.

For the identification of tinnitus subgroups, the collection of longitudinal ecologically valid data sets based on mHealth solutions has been recognized by several researchers. Technically, mobile crowdsensing techniques[22] or Ecological Momentary Assessments[23] are mainly utilized to gather the required data sets. For tinnitus research, these technologies have already shown that they can collect valuable data[24,25]. To identify subgroups of tinnitus patients, data sources established by the use of mHealth solutions have also revealed to be appropriate[26]. Several of these works have presented their findings on data of the TrackYourTinnitus platform (TYT), which was developed to evaluate daily symptom fluctuations of tinnitus patients. TYT comprises two mobile native (developed without using frameworks) applications (an Android and an iOS app), a website (http://www.trackyourtinnitus.org), and a server application that stores the data generated by the apps. The platform was developed by an interdisciplinary team of computer scientists, medical doctors, and psychologists. It can be freely used by interested users, the apps can be downloaded through the official app stores from Apple and Google. In essence, the following complete the following procedure: First, they have to fill out three registration questionnaires after downloading the app. After that, they decide on the number of daily notifications. Each notifications reminds the user to fill out a daily questionnaire, comprising so-called EMA questions, which aim at the momentary tinnitus situation of a user. In addition, the environmental sound level is collected through the microphone of the used smartphone when filling out the daily questionnaire. In terms of feedback, the app visualizes the gathered data and through the website, interested users can download their collected data. TYT does not offer further features. Although the platform aims at data for research and it could be assumed that this is of less interest, so far, the platform has gathered more than 100,000 daily questionnaires by more than 3000 users from all over the world. We learned that despite the fact that TYT is an open research project in the sense of a long-running observational study, two aspects are of importance for users to participate. First, the project is without any commercial interest. Second, data is collected anonymously except one reason. If users want to reset their password, they have to provide their mail address. In general, the secure handling of data collected by the use of a smartphone is an important aspect since smartphones provide a lot of opportunities to gather data that indirectly might reveal the user. For example, when GPS data is collected and the location of a user is sent to a central server. In general, works exist that have developed complex configurations with which users can control the provision of mHealth-related, see for example[27]. Interestingly, such works show that users are less interested to control much themselves, therefore it is important that a mHealth solutions tries to secure data and privacy in the best possible way by design. In the case of TYT, only questionnaire data and the environmental sound level are gathered, which might be also one reason to use it frequently by many users. To conclude, the TYT project is running since 2014 and revealed various investigation opportunities, including those, which were initially not planned[26,28]. Beyond TYT, other mHealth solutions have been developed and presented to support diagnosis and therapy of tinnitus patients[5,29,30], which emphasizes the potential of mHealth in this context.

Moreover, the combination of mHealth and machine learning has become very popular recently. The directions followed in this context are manifold. On the one hand, considerations on sparse mHealth data are subject to research when using machine learning methods in the given context[31,32]. On the other hand, large mHealth data sets exist that are investigated by the use of machine learning methods[33]. Moreover, the development of new machine learning methods and the evaluation of existing ones is also considered presently[34,35].

In this work, gender-related differences of TYT users are investigated, hereby based on the following thoughts: Existing insights on TYT, existing works on machine learning methods to identify subgroups of TYT users, and the amount of existing data of TYT users distributed between females and males. Further note that TYT is technically based on mobile crowdsensing techniques[36] and utilizes Ecological Momentary Assessments (EMA) to capture ecologically valid data sets of tinnitus patients. Since 2014, the TYT mHealth platform has gathered more than 100,000 completed questionnaires from its users. With respect to the identification of subgroups, machine

No.	Research question	Machine learning algorithm				Results
		SVM	Tree	RF	NN	
i	Is it generally possible to learn a mapping function from X to y where X are questions that the user answered daily and y is a binary target representing the gender of a user?	✓	✓	✓	✓	Precision on average: male: 81.5% female: 84.3%
ii	Which machine learning model is most suitable for this task and a high prediction power?	✓	✓	✓	✓	Mean accuracy on a fivefold cross validation set: Random Forest classifier (81.7%)
iii	Which are the features with the highest importance to predict the gender?			✓		Most important features are: q8_4: Worries about the tinnitus q8_5: difficulties in following a conversation

Table 1. Overview of the three Research Questions *i-iii*, the used classifiers and the results. SVM = Support Vector Machine, Tree = Decision Tree, RF = Random Forest, NN = Multilayer Perceptron Neural Network. A checkmark means that this classifier has been used to answer the research question.

learning based investigations on the TYT source already exist. For example, in[37], the differences of TYT Android and iOS users were investigated, while in[38], entity (i.e., individual TYT users) similarity was investigated to label the future observations referring to an entity.

For the investigation at hand, two prerequisites are important: First, it must be defined which type of gender differences are addressed in this work. The authors of[12] define the following important differences: the (1) biological classification encoded in the DNA and the (2) understanding of the respective social roles, behavior, and expressions. In this work, we refer our considerations to the latter type of difference. Second, it must be defined which gender-related aspects of TYT users shall be investigated. The answer to this question is that our goal is to predict the gender of the user of a provided daily assessment. A daily TYT assessment, in turn, is based on the filled-out daily questionnaire, which comprises 8 EMA questions (users can opt which questions they actually want to fill out; in addition, 1 question varies among users based on an answer given to the perceived worst symptom provided through one baseline questionnaire) that capture the current situation of a TYT user (see this work for a detailed explanation[39]). Note that TYT users have two options to fill out this questionnaire. The first option entails receiving up to 12 random notifications per day, which then remind users to fill out the questionnaire, while the second option allows users to determine fixed points in time to receive the notifications. Furthermore, baseline questionnaires, which must be answered when using the smartphone app for the first time, provide the information on the gender of a TYT user. Based on this information, 15 features were identified—out of the 8 daily questions—for the gender prediction task, covering aspects like stress, worries, arousal, depression, mood, or the loudness of the momentarily perceived tinnitus. A detailed explanation of the features is provided in Table 3.

Given these two prerequisites, the overall goal of the work at hand is the prediction of the gender of the user of a given daily TYT assessment based on machine learning methods. A binary classification is therefore accomplished that deals with the following detailed questions (note that for the classification task, technically, Sklearn[40] has been used):

(i) Is it possible to learn a mapping function from X to y of TYT individuals, for which X are questions that the user answered daily and y is a binary target representing the gender of the respective TYT user?
(ii) Which machine learning model is mostly suitable for this task and has a high prediction power?
(iii) Which are the features with the highest importance to predict the gender?

It is briefly discussed whether other approaches have trained binary classifiers on mHealth related data with respect to research questions on gender-related differences. In general, works exist that have trained a binary classifier on mHealth data. For example, the authors of[41] used such a classifier for respiration disorders of mHealth applications. Furthermore, approaches exist that investigated gender differences in the general context of mHealth solutions. However, their focus is different to the one that is investigated in this work. More specifically, other works[42,43] investigate differences when using mHealth technologies from a general point of view. That means that they investigate whether there is a difference between men and women when addressing medical issues while using mHealth solutions. Yet, the focus of these works is different to the presented work: they start with the gender and try to establish which bias this might generate on the use of a solution. In contrast, this work starts from the data source and tries to predict the gender. Although these two perspectives address the same overall research context and are therefore intertwined, the research questions they are addressing are different. Still, to the best of the authors‘ knowledge, similar works that present a binary classifier on mHealth data with respect to results on gender-related differences do not exist yet.

Results

In this section, the three research questions are discussed subsequently. First, it is discussed whether it is generally possible to solve the gender prediction task by using machine learning with relevant results. Next, the hyper-parameters of the chosen classifiers must be fine-tuned. Finally, by using the knowledge from Research Questions *i* and *ii*, the question must be answered, which of the features are mostly suitable to classify the gender. A summary of this section is provided in Table 1.

Classifier	Precision male	Precision female	F1-score
Support Vector Machine	0.80	0.86	0.83
Decision Tree	0.81	0.80	0.81
Neural Network	0.82	0.83	0.83
Random Forest	0.83	0.88	0.85

Table 2. Comparison of the four used classifiers in terms of precision per gender and F1-score. Number of examples is denoted by $m = 1702$. Used features: `{q1, q2, ..., q7, q8_5}`, test size 20%. Note that the feature labels `qx` are further explained in Table 3.

Research question i. In this study, gender is considered to be binary as there is no data for diverse tinnitus patients. Given that the target classes are uniformly distributed, random guessing for a binary classification task leads to an accuracy of 50% on average. Consequently, a mapping from X to y is adding information if the accuracy of a classifier is higher than 50%. If it is significantly higher than 50%, it must be decided based on the achieved accuracy whether it is actually relevant or useful. X was used as the (sub)set of features and y as the target for `gender`, with `{male, female}` as possible classes.

The classification task was accomplished using Python, as this is one of the most used languages for Machine Learning[40], which enables comparisons to many other research results. Four classifiers from the scikit learn library were used for the investigations: A Support Vector Machine, a Multilayer Perceptron Neural Network, a Decision Tree, and a Random Forest. All of them were able to guess the gender with a significantly higher accuracy than 50%. These classifiers were selected as they are well known to get high accuracy scores for high dimensional classification tasks on small to middle-sized datasets[44–47].

Note that the more features were added to the classifiers, the higher was the accuracy. For the testing set, a fivefold cross validation was used to avoid overfitting. As can be seen from Table 2, the random forest classifier had the highest prediction power in this distribution.

Research question ii. As there is no other satisfying metric such as training time or minimal false positives rates, it was decided to further investigate the classifiers accuracy.

To do so, a fine-tuning of the hyper-parameters of the Random Forest classifier was performed. This tuning is also known as a grid search[48,49]. Therefore, the hyper-parameters of interest were selected, which can be seen in Fig. 1. Then, one of the hyper-parameters was varied while keeping all others constant. The resulting parameters-dictionary was passed to the Random Forest classifier into the same training and testing set of the approaches of Research Question *iii*, again with a fivefold cross validation[50–52]. Here, a fivefold split was used instead of a tenfold split for the purpose of having a sufficient testing size. Additionally, this allows to speed up training and testing time as well as to vary more hyper-parameters within the grid search. The cross validation further prevents the Random Forest from overfitting of the training set[53]. For each possible combination of the parameters dictionary, the accuracy was saved. After trying all variations, the variation with the highest accuracy determined the final parameters set up of the Random Forest classifier in the testing set.

The number of decision trees in the random forest was increased up to 1,000 for a slight improvement of the overall accuracy. However, a further increase of `n_estimators` did not improve the score in the testing set. If the `max_depth` parameter was lowered to 10, the lowest standard deviation of 2% within the fivefold cross validation was attained. The best ranked Random Forest classifier received an accuracy of 87% in the first cross validation set. The average cross-validated test score is **81.65%**, with a standard deviation of 4%.

Research question iii. There exist several techniques to determine feature importance, such as random, heuristic, or complete approaches[54]. In order to answer the third Research Question iii, three strategies were pursued. Before the strategies were accomplished, a sub-dataframe was created that contains the feature of interest and the target gender. This sub-dataframe was then filtered, so that it equally contains 50% men and 50% women.

As the first strategy, a closer look was put on the random forest approach. Importantly, it has no bias in terms of the underlying distribution of the mapping function. The forest simply measures the impact in accuracy. The higher the accuracy score for a mapping from a feature to the target is, the higher its impact on the target is. The second approach tried to measure the impact of single features using correlations with the target `gender`. The correlation matrix also helps the authors to get a more detailed insight into the cross-correlation between the features and a single-viewed impact of a feature on the target. The higher the correlation is, the higher the impact to the target is. Note that the correlation method varied with the scaling (binary, discrete, continuous) of a feature. For a univariate classification on gender, a rise in accuracy was expected if the correlation rises. Third, the permutation importance for a univariate Random Forest classification per feature was calculated[55] as follows: First, the classifier was trained on a training set. Then, using cross-validation, a baseline metric was evaluated on a testing set. The permutation importance was then defined as the difference of the baseline metric with the trained feature and the baseline metric with a completely random, artificial feature.

All approaches have different units to measure the impact (Accuracy, r-value, and percentage-improvement). In order to make these three approaches comparable, a ranking of the results of the three approaches was created (see Fig. 2), and statistics for the two gender groups added, respectively. The dynamic questions `q_i`, with i = 0, 1, ..., 8 have on average a better ranking than the questions `q_1, q_2, ..., q_7`. Throughout all three approaches, *strong worries* (ranked first) and *difficulties in following a conversation* (ranked second) are the two most important features in order to predict the gender. The p-value column shows that these gender differences are all significant. From a statistical point of view, the mean difference between the two groups *male* and *female* generally supports the hypothesis that male individuals experience tinnitus differently than female individuals.

Discussion

The authors are aware of the fact that by including the dynamic question `q8` (The follow-up questions about the worst tinnitus symptom), only a smaller subset of TYT users could be investigated (out of all individuals), which is predestined to have a higher bias. Instead of 80,966 examples, the subsets had sizes between 3400 (4%) and 14,000 (17%) user examples. The different sizes of male and female individuals by gender can also be seen in Fig. 4. That means., if `q8_5` (Difficulties in following a conversation) is chosen, it means that 10.9% of the women are included in the dataset. These subsets decrease in size again once an equal split for the target (50% men and 50% women) is performed. As a conceivable result, these subsets could not be representative anymore for the underlying distribution that has a size of $m = 80{,}966$. Consequently, the distribution of the chosen subset was compared with and without feature `q8_5` (Difficulties in following a conversation). Note that the features `q1, q2, ..., q7` were always included. For both female and male individuals, the null-hypothesis cannot be rejected, namely that these samples are drawn from the same distribution, as can be seen in Fig. 3. Grouped by gender, the distribution of the whole dataset and the sub-dataset for the features *handedness* and *family history of tinnitus complaints* was also compared. For these gender-grouped features, no significant differences between the samples could be revealed. We further compared the baseline characteristics of those individuals that only filled out the baseline characteristics and those that filled out both, baseline and follow-up questionnaires (see Table 4). These two groups also show no significant differences in distribution. In addition, the completion for the daily questionnaire differs at a gender-based level and a user-based level. More specifically, most users fill out the daily questionnaire between 1 and 10 times, while others fill it out 100 times or more. The filling-out behavior can be seen in Fig. 5. This means that some users are more represented in the training and testing set than others. However, this does not lead to a different distribution of the baseline characteristics.

Less notably, the gender classification accuracy increases if `q_8` (worst symptom) is added. That is due to the fact that there are gender differences in the worst symptom of a tinnitus patient. If a closer look is taken at Fig. 4, striking differences can be seen in the distribution of the worst symptom. Women tend to have more difficulties in falling asleep, whereas men tend to suffer relatively more by having difficulties in following a conversation. The authors of[56] revealed similar symptoms of individuals in their work on tinnitus problems. Understanding speech and sleep problems were ranked as the most challenging ones without grouping by gender. The symptom `sensitive to environmental noises` could be biased by hyperacusis. Individuals with sensitive noise perception would tend to report higher scores here. Since hyperacusis is not assessed in the baseline questionnaire, we cannot consider it. In addition, more factors might bias the discussed symptom (e.g., if one of the parents worked in a noisy factory for a longer period of time, which is not captured by TYT) (Fig. 5).

When taking a closer look to the correlations of features `q4` (Mood of user) and `q8_7` (Depressed because of tinnitus), which is depicted in Fig. 6, a negative value can be seen. It is evident why these features should be negatively correlated. An observation with a strong positive correlation appears for the features stressfulness and loudness of the perceived tinnitus: The louder the tinnitus is, the more stressful it is.

The authors are aware of the trade-off between the depth of a tree within the forest and the standard deviation of the accuracy for a cross-validation set. A higher accuracy could be achieved for a single cross-validation set by increasing the depth of a tree. However, by increasing the depth, a higher variance must be expected between the cross-validation sets, which is an indicator for overfitting of the training set.

For Research Question iii (Which is the most important feature?), the result in the lower-ranked features is ambiguous. For the top three most important features, all three methods rank *strong worries* and *difficulties in following a conversation* firstly and secondly, respectively. For the non-changing questions `q1, q2, ..., q7`, however, it is not clear which one could be ranked in the middle or lower for a univariate feature importance. In summary, it can be said that the dynamic question `q_8` is rated more important than the non-changing ones.

The results of the presented investigation are both clinically relevant as well as helpful for users of the TYT platform. Regarding clinical relevance, as profound indicators exist that gender differences exist for tinnitus patients[57], TYT can be a valuable alley to learn more about daily fluctuations of tinnitus patients with respect to their gender. As our result show that the answers of the daily questionnaires can predict the gender of TYT users, inversely, the daily answers can be indicators for the symptom differences of men and women. As we further found out that the worst symptom is an important feature, we are in line with other research works beyond the scope of mHealth data[13,58–60]. Furthermore, studies that have found gender-related differences in tinnitus patients without using mHealth solutions might particularly benefit from the use of mHealth. For example, in the work presented by[61], it is shown that gender-related differences exist for insomnia. As built-in sensors of smartphones can be used in the context of insomnia[62], mHealth solutions might leverage findings like shown in[61]. Due to the gender-related differences we have found in TYT, it is likely that for other research questions like insomnia mHealth solutions can be helpful as well or even leverage already revealed results. We therefore conclude that in the context of gender-related differences of tinnitus patients, data that were collected with the use of mHealth solutions like TYT are relevant for medical research and clinical practice. Regarding the aspect of helping users with the findings shown here, consider, for example, the work of[58]. One outcome of the latter work describes that anxiety is only associated with bothersome tinnitus in men. Anxiety, in turn, can be easily monitored using a solution like TYT. In this particular case, the gender-related differences can be used to help, for example, men in coping with their anxiety syndrome by learning more about their daily fluctuations (if such fluctuations exist) when using TYT on a daily basis. Inversely, TYT can be used to figure out more variables that are associated with the gender and tinnitus, which might lead to the development of focused measures that may help to mitigate the tinnitus of men or woman more effectively. To conclude from a tinnitus perspective, TYT has gathered a lot of data and with this data source we were able to reveal that the question on the worst symptom (answered daily) has a high prediction power of the gender of TYT users. Since TYT asks about several worst symptoms, we consider this type of daily questions important. On the other hand, the combination with the other daily questions lead to the final result to predict the gender of TYT users, which we consider as a new outcome of TYT data and research on mHealth in this context.

Overall, the question was investigated whether the answers of male and female tinnitus patients are useful to gain a gender-based differentiation. Therefore, three research questions were investigated: (i) Is it possible to learn a mapping from X to y for the daily tinnitus questionnaire?, (ii) which is the most suitable classifier for this task, and (iii) which are the most important features? Four different classifiers of the sklearn[40] library from Python were trained to classify the gender of a patient. The most important feature cannot be clearly determined. This result is ambiguous for different feature importance approaches. However, increasing the number of features resulted in a higher classification accuracy. Although the utilization of the possible features showed different results, the gender of the user from a provided daily questionnaire could be revealed with a relevant accuracy. The findings thus might be a valuable basis for the development of more individualized tinnitus treatments, even beyond the scope of TYT.

Materials and methods

The study was approved by the Ethics Committee of the University Clinic of Regensburg (ethical approval No. 15-101-0204). All users read and approved the informed consent before participating in the study. The study was carried out in accordance with relevant guidelines and regulations.

The features. For the gender prediction task, two linked data sets were used. The first one, named *Tinnitus Sample Case History Questionnaire (TSCHQ)*, is only provided to a individual *once*, and asks questions like *date of birth*, *handedness*, *family history of tinnitus complaints*, the target variable *gender*, and the worst symptom that is related with tinnitus. Baseline characteristics from this questionnaire can be seen in Table 4. Note that this table only contains individuals that filled out both, the baseline and the daily questionnaire. The worst symptom thereby can be one of the following:

- I am feeling depressed because of the tinnitus.
- I find it harder to relax because of the tinnitus.
- I have strong worries because of the tinnitus.
- Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film.
- Because of the tinnitus it is hard for me to get to sleep.
- Because of the tinnitus it is difficult to concentrate.
- Because of the tinnitus I am more irritable with my family, friends and colleagues.
- Because of the tinnitus I am more sensitive to environmental noises.
- I don't have any of these symptoms.

The second data set, named *daily questionnaire*, contains daily given answers of a registered individual. This daily questionnaire includes eight questions about the current tinnitus state, i.e., the tinnitus situation and the feelings of the individual *right now*. However, the eighth *dynamic* question depends on the worst symptom of the individual from the TSCHQ questionnaire and asks whether the individual has this specific worst symptom right now or not. If an individual user answered *I don't have any of these symptoms* in the beginning, no question appears in the daily questionnaires. As a consequence, the number of answers for question 8 depends on the number of individuals that have selected this worst symptom in the questionnaire TSCHQ. On the other hand, the number of answers for questions one to seven equals each other. These questions are seen by every individual and are as follows:

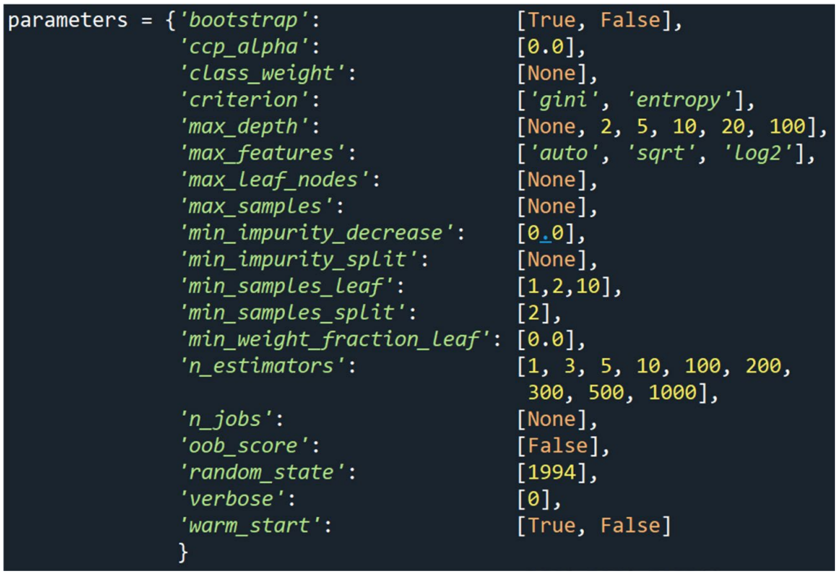
1. Did you perceive the tinnitus right now?
2. How loud is the tinnitus right now?
3. How stressful is the tinnitus right now?
4. How is your mood right now?
5. How is your arousal right now?
6. Do you feel stressed right now?
7. How much did you concentrate on the things you are doing right now?
8. *This question depends on the worst symptom selected in the questionnaire TSCHQ.*

Depending on the features that are selected for the classification task, the number of examples m depends on the eighth dynamic question.

Data preparation. The raw data set with the daily answers had the size ($m = 83349$, $n = 19$), where m denotes the number of samples, and n the number of columns. The columns of interest are `individual_id, q1, q2, ..., q7, q8_1, q8_2, ..., q8_8`. In total, the preparation of the data set needed much efforts, namely the following considerations and steps:

	Meaning	Scaling	Implementation	Count	Mean	Std
Question1	Did you perceive the tinnitus right now?	Binary	YesNoSwitch	80,969	0.76	0.43
Question2	How loud is the tinnitus right now?	Continuous	Slider in range (0,1)	80,969	0.46	0.3
Question3	How stressful is the tinnitus right now?	Continuous	Slider in range (0,1)	80,969	0.36	0.28
Question4	How is your mood right now?	Discrete	SAM from 0 to 1 with step size 0.125	80,969	0.56	0.21
Question5	How is your arousal right now?	Discrete	SAM from 0 to 1 with step size 0.125	80,969	0.26	0.22
Question6	Do you feel stressed right now?	Continuous	Slider in range (0,1)	80,969	0.28	0.24
Question7	How much did you concentrate on the things you are doing right now?	Continuous	Slider in range (0,1)	80,969	0.58	0.31
Question8_0	Because of the tinnitus it is hard for me to get to sleep	Binary	YesNoSwitch	7919	0.35	0.48
Question8_1	I am feeling depressed because of the tinnitus	Binary	YesNoSwitch	10,361	0.23	0.42
Question8_2	I find it harder to relax because of the tinnitus	Binary	YesNoSwitch	13,904	0.45	0.5
Question8_3	I don't have any of these symptoms	NULL	NULL	NULL	NULL	NULL
Question8_4	I have strong worries because of the tinnitus	Binary	YesNoSwitch	10,839	0.27	0.45
Question8_5	Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film	Binary	YesNoSwitch	11,877	0.33	0.47
Question8_6	Because of the tinnitus it is difficult to concentrate	Binary	YesNoSwitch	8220	0.42	0.49
Question8_7	Because of the tinnitus I am more irritable with my family, friends and colleagues	Binary	YesNoSwitch	3391	0.32	0.47
Question8_8	Because of the tinnitus I am more sensitive to environmental noises	Binary	YesNoSwitch	9179	0.09	0.29
Qender	0 = Male, 1 = female	Binary	Single Choice	80,969	0.26	0.44

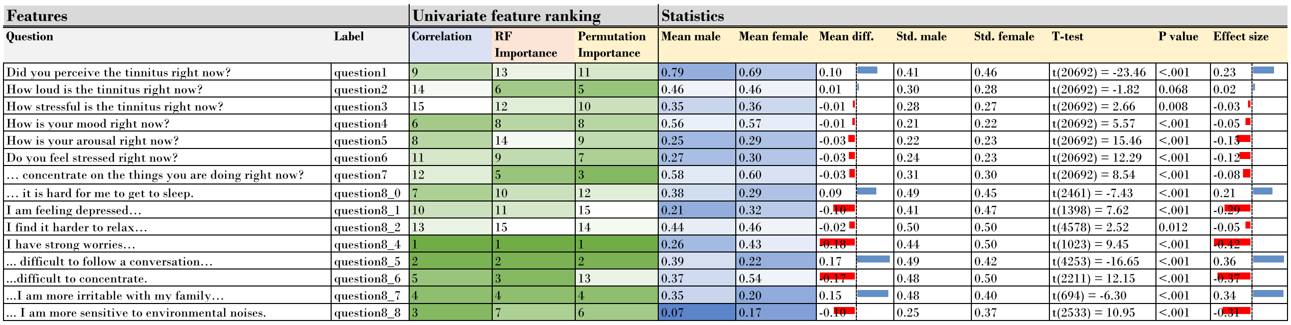
Table 3. Description of the dataframe used for the machine learning approaches. Note that the count for the questions 8_0, 8_1, ..., q_8 is dependent on the number of individuals that selected this answer in the baseline questionnaire. If an individual selected *I don't have any of these symptoms*, no follow-up question appeared, so that these values are NULL. SAM = Self Assessment Manikin[65].



```
parameters = {'bootstrap':                [True, False],
              'ccp_alpha':                [0.0],
              'class_weight':             [None],
              'criterion':                ['gini', 'entropy'],
              'max_depth':                [None, 2, 5, 10, 20, 100],
              'max_features':             ['auto', 'sqrt', 'log2'],
              'max_leaf_nodes':           [None],
              'max_samples':              [None],
              'min_impurity_decrease':    [0.0],
              'min_impurity_split':       [None],
              'min_samples_leaf':         [1,2,10],
              'min_samples_split':        [2],
              'min_weight_fraction_leaf': [0.0],
              'n_estimators':             [1, 3, 5, 10, 100, 200,
                                           300, 500, 1000],
              'n_jobs':                   [None],
              'oob_score':                [False],
              'random_state':             [1994],
              'verbose':                  [0],
              'warm_start':               [True, False]
              }
```

Figure 1. Set of hyper-parameters for a grid search in order to improve the forest's accuracy. Note that not all hyper-parameters have be varied, such as n_jobs, oob_score or verbose. Only hyper-parameters were varied that have a higher impact on the accuracy score. However, static parameters are listed for the purpose of integrity.

The individual_id is crucial to merge *TSCHQ* with the daily questionnaire in order to get the gender for a sample of answers. As a consequence, all rows where individual_id is NULL were dropped. This affected 1.2% of the samples, i.e., 82,351 samples remained. In the next step, values for q4(mood right now) and q5(arousal right now) were replaced that have been reported incorrectly from Android devices. For these questions, an individual user can select a position in a self-assessment manikin individual interface feature to represent his or her mood with 9 different steps (i.e., the granularity). However, the Android implementation rounds the values to tenths, which leads to incorrect values. For example, 0.13 has to become 0.125, or 0.88 has to become 0.875.



Features		Univariate feature ranking			Statistics							
Question	Label	Correlation	RF Importance	Permutation Importance	Mean male	Mean female	Mean diff.	Std. male	Std. female	T-test	P value	Effect size
Did you perceive the tinnitus right now?	question1	9	13	11	0.79	0.69	0.10	0.41	0.46	t(20692) = -23.46	<.001	0.23
How loud is the tinnitus right now?	question2	14	6	5	0.46	0.46	0.01	0.30	0.28	t(20692) = -1.82	0.068	0.02
How stressful is the tinnitus right now?	question3	15	12	10	0.35	0.36	-0.01	0.28	0.27	t(20692) = 2.66	0.008	-0.03
How is your mood right now?	question4	6	8	8	0.56	0.57	-0.01	0.21	0.22	t(20692) = 5.57	<.001	-0.05
How is your arousal right now?	question5	8	14	9	0.25	0.29	-0.03	0.22	0.23	t(20692) = 15.46	<.001	-0.15
Do you feel stressed right now?	question6	11	9	7	0.27	0.30	-0.03	0.24	0.23	t(20692) = 12.29	<.001	-0.12
… concentrate on the things you are doing right now?	question7	12	5	3	0.58	0.60	-0.03	0.31	0.30	t(20692) = 8.54	<.001	-0.08
… it is hard for me to get to sleep.	question8_0	7	10	12	0.38	0.29	0.09	0.49	0.45	t(2461) = -7.43	<.001	0.21
I am feeling depressed…	question8_1	10	11	15	0.21	0.32	-0.10	0.41	0.47	t(1398) = 7.62	<.001	-0.29
I find it harder to relax…	question8_2	13	15	14	0.44	0.46	-0.02	0.50	0.50	t(4578) = 2.52	0.012	-0.05
I have strong worries…	question8_4	1	1	1	0.26	0.43	-0.18	0.44	0.50	t(1023) = 9.45	<.001	-0.42
... difficult to follow a conversation…	question8_5	2	2	2	0.39	0.22	0.17	0.49	0.42	t(4253) = -16.65	<.001	0.36
...difficult to concentrate.	question8_6	5	3	13	0.37	0.54	-0.17	0.48	0.50	t(2211) = 12.15	<.001	-0.37
...I am more irritable with my family…	question8_7	4	4	4	0.35	0.20	0.15	0.48	0.40	t(694) = -6.30	<.001	0.34
… I am more sensitive to environmental noises.	question8_8	3	7	6	0.07	0.17	-0.10	0.25	0.37	t(2533) = 10.95	<.001	-0.31

Figure 2. Comparison of three approaches to determine the most important feature for gender prediction. A ranking value of 1 means that this feature is most important to predict the gender.

	Characteristic					
	n	Age (std)	Right-handed	Left-handed	Both sides	Existing family history of tinnitus complaints
Male	1871	49.23 (14.40)	1345 (71%)	282 (13%)	244 (16%)	426 (23%)
Female	875	46.04 (14.72)	650 (75%)	126 (11%)	99 (14%)	235 (27%)
Total	2746	48.71 (14.89)	1994 (73%)	408 (12%)	343 (15%)	661 (24%)

Table 4. Baseline characteristics of the Tinnitus Sample Case History Questionnaire (TSCHQ) for all individuals that filled out at least one follow-up questionnaire. Individual users that registered for this study, but did not fill out at least one follow-up questionnaire, are not considered in this table.

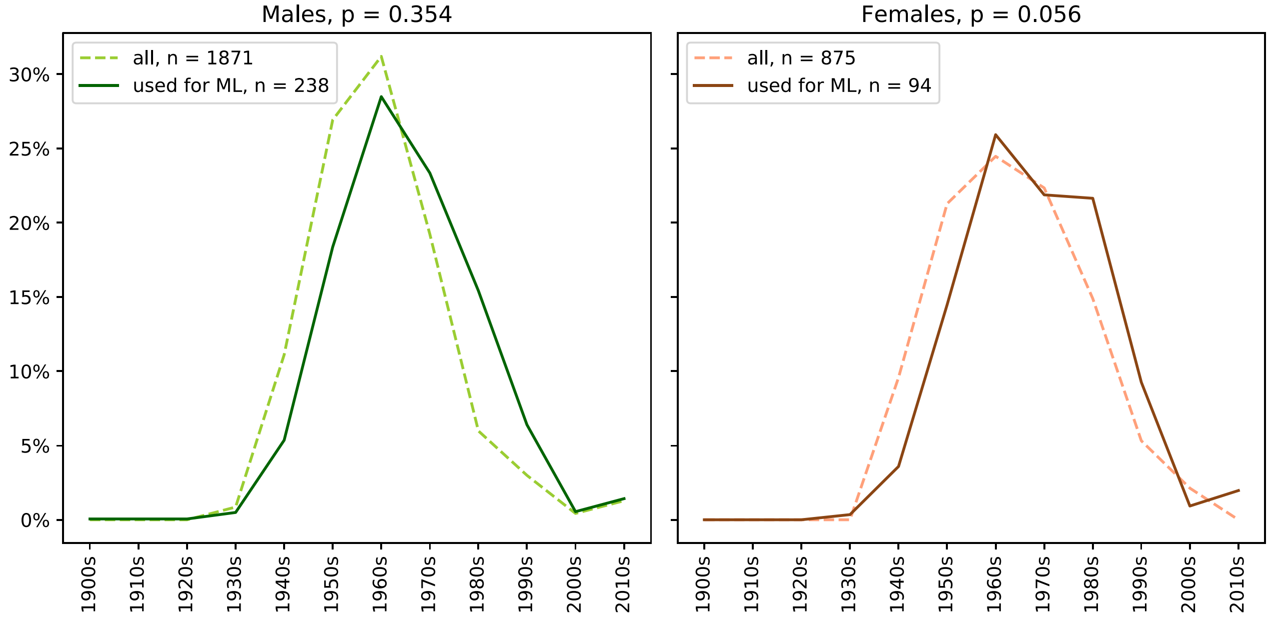


Figure 3. The dashed lines denote the age distribution for the all individuals, whereas the solid lines indicate the subset of individuals used for the machine learning calculations. This subset has a size of $m = 11{,}877$, and contains 238 + 94 individual users. For all users, m equals to 80,969. Note that the high p-values for both groups indicate equality of the age distribution.

Missing value treatment. As every question is optional, sometimes individuals skipped questions. Therefore, the imputation module from the Sklearn library was used to fill in missing values. In order not to change the data distribution, the data set per individual was calculated. If any of the values for questions 1, 2, ..., 7 was `NULL`, the missing value treatment was performed. Therefore, the non-null values per column were counted. If there are two or more non-null values, an individual-specific KNN imputation for slider questions with `range(0,1)` and Boolean questions[63] was performed. In case an individual user always skipped a specific question, there is no reference how this individual user usually would have answered this question. In such cases, a simple imputation was performed with a median value of the whole data set for slider questions and a most frequent replace

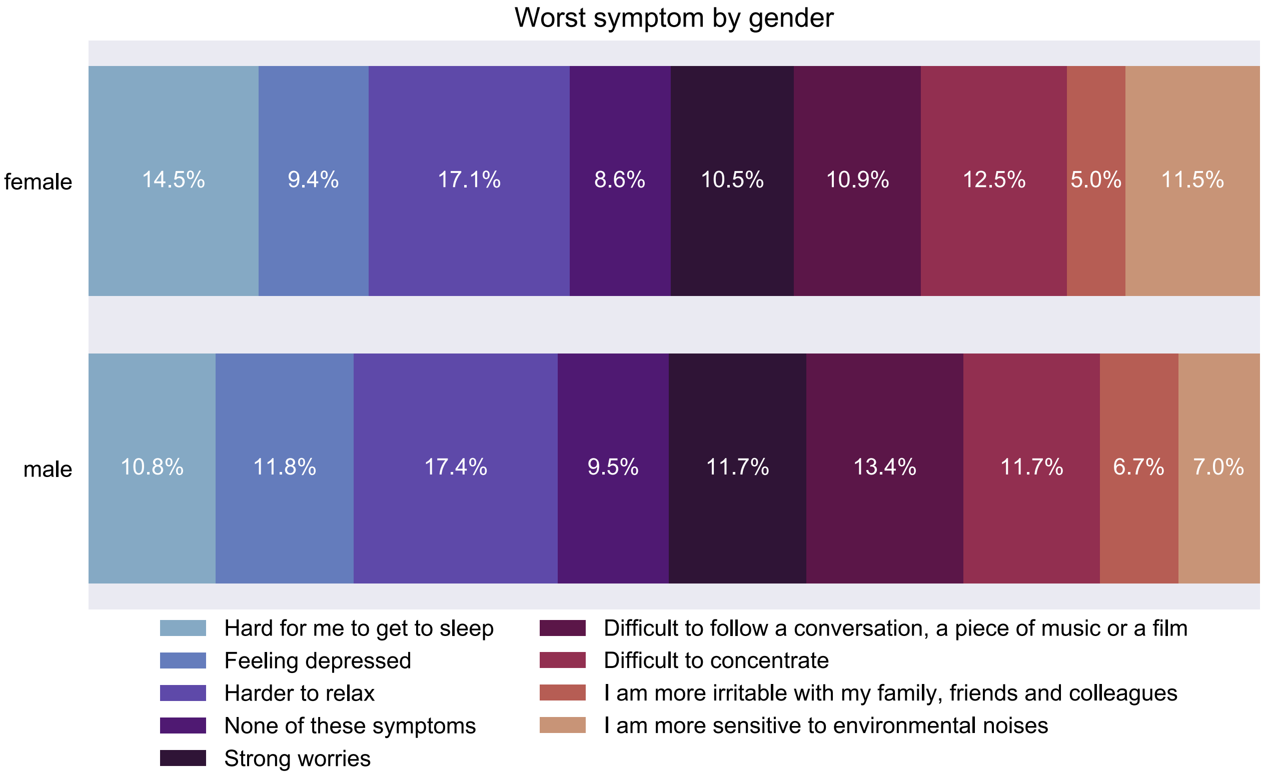


Figure 4. Distribution of the worst symptom grouped by gender in a horizontal stacked plot. Each row of the figure adds up to 100%.

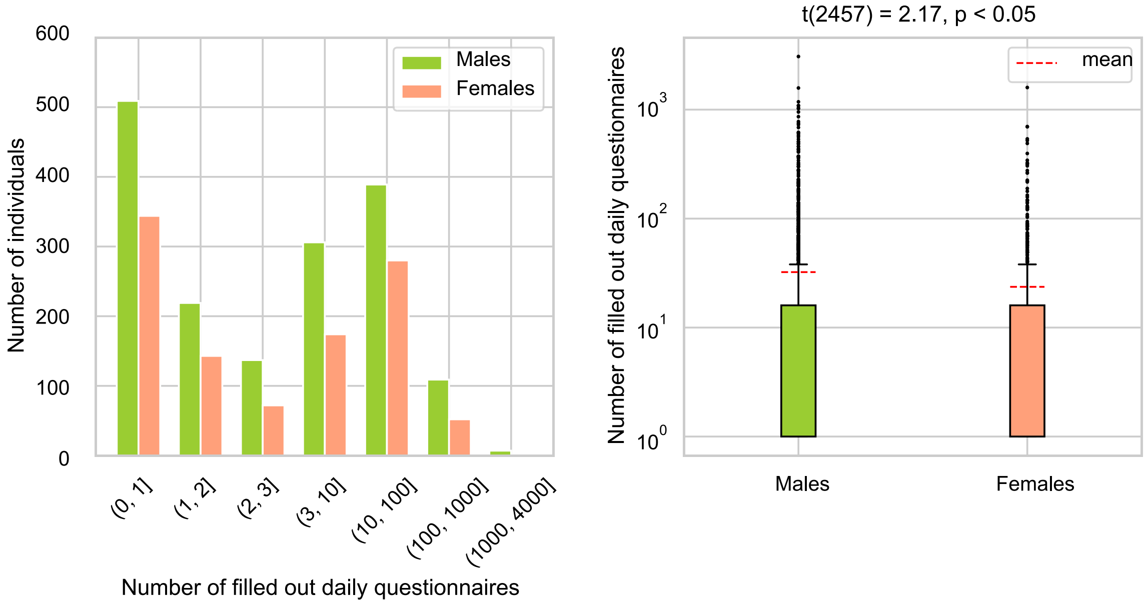
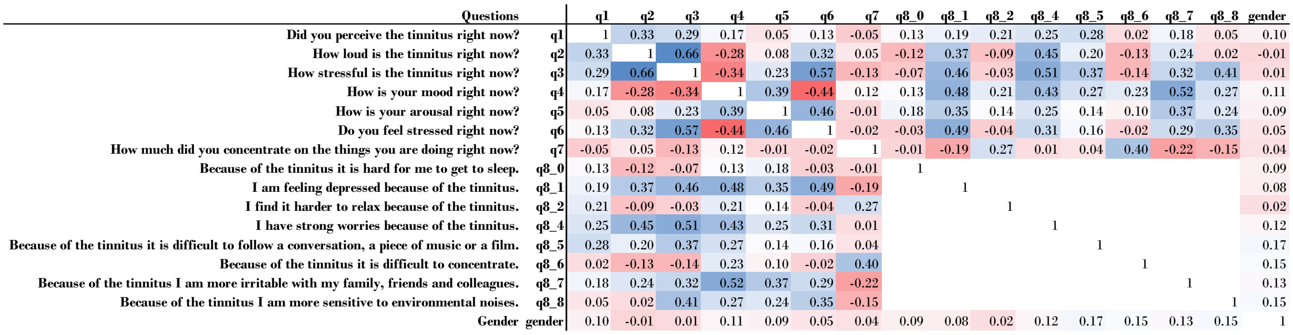


Figure 5. Number of filled out daily questionnaires per group (left) and per gender (right). The red-dashed line in the right plot indicates the mean value. Most of the individual users filled out the questionnaire only once. On average, men answered the questionnaire 32 times (± 124 std), and women 24 times (± 82 std) with t(2757) = 2.17 and $p < 0.05$. Notably, there is one male user that filled out the daily questionnaire 3073 times.

for Boolean questions, respectively. An iterative imputation approach was not used as suggested by the authors of[64], because then it would be required to round the estimation of Boolean questions to integer values and fit respective answers to a valid value in $\{0, 0.125, ..., 1\}$. For the dynamic variable `question8`, missing value treatment does not make sense, as the questions are different. For example, if an individual user has selected *feeling depressed* as a worst symptom, his or her question eight is *"Are you feeling depressed right now?"*. For all the other linked questions, the individual has never seen another dynamic question like *"Are you sensitive to environmental*



Questions		q1	q2	q3	q4	q5	q6	q7	q8_0	q8_1	q8_2	q8_4	q8_5	q8_6	q8_7	q8_8	gender
Did you perceive the tinnitus right now?	q1	1	0.33	0.29	0.17	0.05	0.13	-0.05	0.13	0.19	0.21	0.25	0.28	0.02	0.18	0.05	0.10
How loud is the tinnitus right now?	q2	0.33	1	0.66	-0.28	0.08	0.32	0.05	-0.12	0.37	-0.09	0.45	0.20	-0.13	0.24	0.02	-0.01
How stressful is the tinnitus right now?	q3	0.29	0.66	1	-0.34	0.23	0.57	-0.13	-0.07	0.46	-0.03	0.51	0.37	-0.14	0.32	0.41	0.01
How is your mood right now?	q4	0.17	-0.28	-0.34	1	0.39	-0.44	0.12	0.13	0.48	0.21	0.43	0.27	0.23	0.52	0.27	0.11
How is your arousal right now?	q5	0.05	0.08	0.23	0.39	1	0.46	-0.01	0.18	0.35	0.14	0.25	0.14	0.10	0.37	0.24	0.09
Do you feel stressed right now?	q6	0.13	0.32	0.57	-0.44	0.46	1	-0.02	-0.03	0.49	-0.04	0.31	0.16	-0.02	0.29	0.35	0.05
How much did you concentrate on the things you are doing right now?	q7	-0.05	0.05	-0.13	0.12	-0.01	-0.02	1	-0.01	-0.19	0.27	0.01	0.04	0.40	-0.22	-0.15	0.04
Because of the tinnitus it is hard for me to get to sleep.	q8_0	0.13	-0.12	-0.07	0.13	0.18	-0.03	-0.01	1								0.09
I am feeling depressed because of the tinnitus.	q8_1	0.19	0.37	0.46	0.48	0.35	0.49	-0.19		1							0.08
I find it harder to relax because of the tinnitus.	q8_2	0.21	-0.09	-0.03	0.21	0.14	-0.04	0.27			1						0.02
I have strong worries because of the tinnitus.	q8_4	0.25	0.45	0.51	0.43	0.25	0.31	0.01				1					0.12
Because of the tinnitus it is difficult to follow a conversation, a piece of music or a film.	q8_5	0.28	0.20	0.37	0.27	0.14	0.16	0.04					1				0.17
Because of the tinnitus it is difficult to concentrate.	q8_6	0.02	-0.13	-0.14	0.23	0.10	-0.02	0.40						1			0.15
Because of the tinnitus I am more irritable with my family, friends and colleagues.	q8_7	0.18	0.24	0.32	0.52	0.37	0.29	-0.22							1		0.13
Because of the tinnitus I am more sensitive to environmental noises.	q8_8	0.05	0.02	0.41	0.27	0.24	0.35	-0.15								1	0.15
Gender	gender	0.10	-0.01	0.01	0.11	0.09	0.05	0.04	0.09	0.08	0.02	0.12	0.17	0.15	0.13	0.15	1

Figure 6. Heatmap for feature-gender cross-correlations. The last column (resp. the last row) shows the correlation of the whole data set (without equal splits for male and female individuals) with the target gender. Depending on the feature scaling, different correlation approaches (Cramer's V, Pointbiserial and Pearson) have been used. The matrix reveals strong positive correlations between stressfulness and loudness of the tinnitus or negative correlations between mood and stressfulness of an individual user. The heatmap was formatted using MS Excel 365. Correlation metrics were calculated using SciPy 1.5.0 within a Python 3.7 environment.

noises right now?", as the individual did not report this as the worst symptom. Consequently, these NULL values were left untreated.

Calculation of the correlation matrix. The values of Fig. 6 were calculated using three different methods depending on the scaling of the features. Note that it is not possible to calculate the correlations of the q8 questions to each other as they are pairwise disjoint. If both features are continuous, the Pearson correlation has been used[66]. If one feature is either discrete or binary and the other is continuous, the Pointbiserial correlation was calculated[67]. Finally, if both features are discrete or binary, the Corrected Cramer's V correlation has been calculated[68]. Further note that Cramer's V correlation is defined for a range of (0,1), whereas Pearson and Pointbiserial for a range of (− 1,1).

Univariate feature classification. For this classification task, a random forest classifier was used as proposed by the authors of[69]. In order not to get a biased estimation of the feature importance, a grouped data set per feature was calculated. As can be seen in Table 3, the number of examples n varies per feature. Therefore, the feature was taken with the smallest training examples (q8_7), and randomly 50% men and 50% women from the target gender were selected. In the next step, X was defined as the feature space of shape (m, n), with m = number of examples, and n = 1, as only one feature was used. Then, a Random Forest classifier from Sklearn was instantiated, including 80% of randomly chosen examples, which denotes the training set. Next, the accuracy on the remaining 20% of the examples was calculated, which denotes the testing set. Note that there is no development set for this subtask, as hyper-parameter tuning is not performed initially. For each feature, this procedure was repeated 10 times and the mean of those 10 accuracies were determined. The features q8_4, q8_5 (worries, difficulties in following a conversation) and q8_6 (difficulties in concentration) reach accuracy values greater than 0.58, which is significantly better than random guessing. Consequently, these features are ranked top three.

Comparison. Comparing the results of the three feature importance approaches, the result for the top two features is unambiguous. However, the correlation approach ranks *sensitivity on environmental noises* on a third place, whereas the permutation and random forest approach *difficulties in concentration* have different results on this rank place.

Supervised machine learning application.

Feature selection. After determining which variables were more and which less important for a univariate approach, the best set of features (multivariate approach) had to be identified in order to find a mapping from X to y, where X is a subset of all features and y is a binary gender prediction with male and female individuals. However, an arbitrary combination of features is only possible within the feature set of $\{q1, q2, ..., q7\}$. Only one out of the features from question 8 can be added optionally. This constraint leads to 1143 valid subsets of the data set. In order to get the best feature list, every single combination of valid subsets to a 80-20 training-testing-split of the data set was applied, before storing its accuracy and the corresponding feature list to a Python dictionary. Given a Random Forest classifier, it can be simply said that a feature list is superior to another if its accuracy on average in the testing set is higher. Without any of the dynamic questions from {q8_0, q8_1, ..., q8_8}, the best set contains the features {q2, q3, ..., q7}. Note that q1 is not included. This set leads to an accuracy of 72.7%, with a testing size of $n = 8276$. If one of the q8-questions is added to the feature set, the most promising combination contains {q1, q2, ..., q7, q8_5}, with an accuracy of 81.7% on average, and a test size of $n = 1702$.

Classifier comparison. This section covers aspects to address Research Question *ii*: Which machine learning model is most suitable for predicting the gender of a individual user and has a high prediction power? More specifically, four supervised machine learning classifiers were investigated: A Support Vector Machine[70], a Mul-

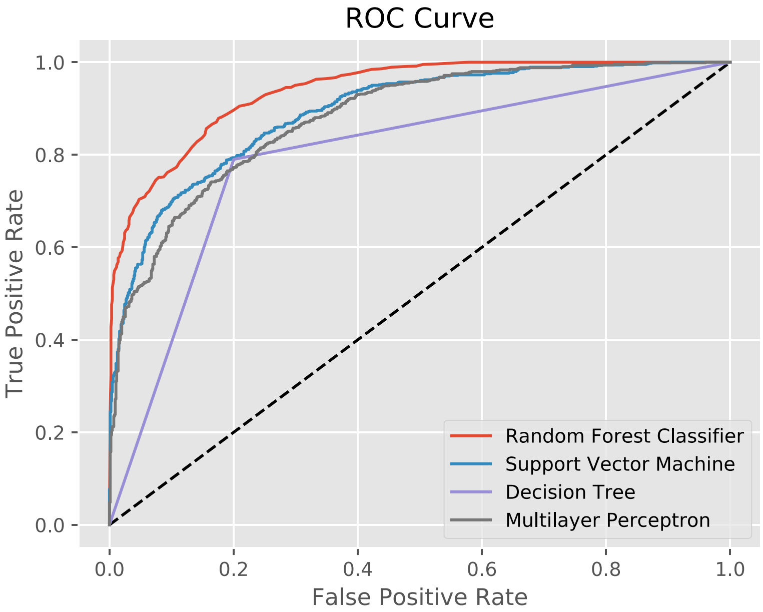


Figure 7. ROC curve for compared classifiers. As the decision tree contains only pure subsets, the class probabilities are either 0 or 1. This leads to a triangled ROC curve.

```
SVC(C=1.0, break_ties=False, cache_size=200, class_weight=None, coef0=0.0,
            decision_function_shape='ovr', degree=3, gamma='scale', kernel='rbf',
            max_iter=-1, probability=False, random_state=1994, shrinking=True,
            tol=0.001, verbose=False)

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
            max_depth=None, max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort='deprecated',
            random_state=1994, splitter='best')

MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9,
            beta_2=0.999, early_stopping=False, epsilon=1e-08,
            hidden_layer_sizes=(8, 16, 32, 2), learning_rate='adaptive',
            learning_rate_init=0.001, max_fun=15000, max_iter=500,
            momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
            power_t=0.5, random_state=1994, shuffle=True, solver='adam',
            tol=0.0001, validation_fraction=0.1, verbose=False,
            warm_start=False)

RandomForestClassifier(bootstrap= False, ccp_alpha= 0.0, class_weight= None,
            criterion= 'entropy', max_depth = 50, max_features= 'auto',
            max_leaf_nodes= None, max_samples= None, min_impurity_decrease= 0.0,
            min_impurity_split= None, min_samples_leaf= 1, min_samples_split= 2,
            min_weight_fraction_leaf= 0.0, n_estimators= 1000, n_jobs= None,
            oob_score = False, random_state= 1994, verbose= 0, warm_start= True)
```

Listing 1. Hyperparameter set-up for the used classifiers.

tilayer Perceptron Neural Network (MLP)[71], a Decision Tree[72] and a Random Forest[69]. With the same testing size from the previous section of $n = 1702$, the following results were obtained. The Decision Tree reached the lowest accuracy with 79%, followed by the Support Vector Machine with 80%, and the Multilayer Perceptron with 81%. The Random Forest classifier reached 86% in accuracy in the best cross validation set. The ROC curve in Fig. 7 affirms the superiority of the Random Forest classifier for this specific classification. The Support Vector Machine and the Multilayer Perceptron have a very similar performance. The Decision Tree contains only pure subsets in its final leaves, which leads to a triangled ROC curve and in this case, eventually meaning the lowest performance.

Hyper-parameter set-up. In a first approach, the four classifiers have been used mainly with a default set from the Python scikit-learn library[40]. Then, several hyper-parameters were slightly adjusted, i.e., the number of neurons per layer for the Multilayer Perceptron Regressor, and the splitter criterion for the Decision Tree classifier. The details of the hyper-parameters (Supplementary Information) can be seen in Listing 1.

According to the classifiers accuracy, the Random Forest classifier seems to be most suitable for this task, which was used to answer Research Question *ii*.

Data availability

The Python code to replicate the Machine Learning classifiers, figures and tables is available on github: https://github.com/joa24jm/tyt_gender_prediction.

Received: 23 December 2020; Accepted: 12 August 2021
Published online: 15 September 2021

References

1. Kiang, N., Moxon, E. & Levine, R. Auditory-nerve activity in cats with normal and abnormal cochleas. In *Ciba Foundation Symposium-Sensorineural Hearing Loss* 241–273 (1970).
2. Davis, A. & Rafaie, E. A. Epidemiology of tinnitus. In *Tinnitus Handbook* Vol. 1, 23 (2000).
3. Langguth, B. A review of tinnitus symptoms beyond—'Ringing in the ears': A call to action. *Curr. Med. Res. Opin.* **27**, 1635–1643 (2011).
4. Halford, J. B. & Anderson, S. D. Anxiety and depression in tinnitus sufferers. *J. Psychosom. Res.* **35**, 383–390 (1991).
5. Mehdi, M. *et al.* Contemporary and systematic review of smartphone apps for tinnitus management and treatment. (2020).
6. Cederroth, C. R. *et al.* Towards an understanding of tinnitus heterogeneity. *Front. Aging Neurosci.* **11**, 53 (2019).
7. Cederroth, C. R. *et al.* Medicine in the fourth dimension. *Cell Metab.* **30**, 238–250 (2019).
8. Tyler, R. *et al.* Identifying tinnitus subgroups with cluster analysis. *Am. J. Audiol.* **17**, 176–184 (2008).
9. Langguth, B. *et al.* Different patterns of hearing loss among tinnitus patients: A latent class analysis of a large sample. *Front. Neurol.* **8**, 46 (2017).
10. Schecklmann, M. *et al.* Cluster analysis for identifying sub-types of tinnitus: A positron emission tomography and voxel-based morphometry study. *Brain Res.* **1485**, 3–9 (2012).
11. Tyler, R. *et al.* Development and validation of the tinnitus primary function questionnaire. *Am. J. Audiol.* **23**, 260–272 (2014).
12. Pryss, R. *et al.* Smart Mobile Data Collection in the Context of Neuroscience. *Front. Neurosci.* **15**, (2021).
13. Niemann, U., Boecking, B., Brueggemann, P., Mazurek, B. & Spiliopoulou, M. Gender-specific differences in patients with chronic tinnitus-baseline characteristics and treatment effects. *Front. Neurosci.* **14**, 487 (2020).
14. Van der Wal, A. *et al.* Sex differences in the response to different tinnitus treatment. *Front. Neurosci.* **14**, 422 (2020).
15. Han, T. S., Jeong, J.-E., Park, S.-N. & Kim, J. J. Gender differences affecting psychiatric distress and tinnitus severity. *Clin. Psychopharmacol. Neurosci.* **17**, 113 (2019).
16. van Os, J. *et al.* The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. *Depression Anxiety* **34**, 481–493 (2017).
17. Torous, J., Friedman, R. & Keshavan, M. Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions. *JMIR mHealth uHealth* **2**, e2 (2014).
18. Martínez-Pérez, B., De La Torre-Díez, I. & López-Coronado, M. Mobile health applications for the most prevalent conditions by the world health organization: Review and analysis. *J. Med. Internet Res.* **15**, e120 (2013).
19. Rowland, S. P., Fitzgerald, J. E., Holme, T., Powell, J. & McGregor, A. What is the clinical value of mHealth for patients?. *NPJ Digit. Med.* **3**, 1–6 (2020).
20. Seifert, A., Hofer, M. & Allemand, M. Mobile data collection: Smart, but not (yet) smart enough. *Front. Neurosci.* **12**, 971 (2018).
21. Pryss, R. *et al.* Exploring the time trend of stress levels while using the crowdsensing mobile health platform, trackyourstress, and the influence of perceived stress reactivity: ecological momentary assessment pilot study. *JMIR mHealth uHealth* **7**, e13978 (2019).
22. Pryss, R. Mobile crowdsensing in healthcare scenarios: taxonomy, conceptual pillars, smart mobile crowdsensing services. In *Digital Phenotyping and Mobile Sensing*, 221–234 (Springer, 2019).
23. Kraft, R. *et al.* Combining mobile crowdsensing and ecological momentary assessments in the healthcare domain. *Front. Neurosci.* **14**, 164 (2020).
24. Schlee, W. *et al.* Measuring the moment-to-moment variability of tinnitus: The trackyourtinnitus smart phone app. *Front. Aging Neurosci.* **8**, 294 (2016).
25. Probst, T., Pryss, R., Langguth, B. & Schlee, W. Emotional states as mediators between tinnitus loudness and tinnitus distress in daily life: Results from the "trackyourtinnitus" application. *Sci. Rep.* **6**, 1–8 (2016).
26. Schlee, W. *et al.* Momentary assessment of tinnitus–how smart mobile applications advance our understanding of tinnitus. In *Digital Phenotyping and Mobile Sensing*, 209–220 (Springer, 2019).
27. Beierle, F. *et al.* What data are smartphone users willing to share with researchers? *Journal of Ambient Intelligence and Humanized Computing* **11**(6), 2277–2289 (2020).
28. Kraft, R. *et al.* Comprehensive insights into the trackyourtinnitus database. (2020).
29. Sereda, M., Smith, S., Newton, K. & Stockdale, D. Mobile apps for management of tinnitus: Users' survey, quality assessment, and content analysis. *JMIR mHealth uHealth* **7**, e10353 (2019).
30. Mehdi, M. *et al.* Smartphone apps in the context of tinnitus: Systematic review. *Sensors* **20**, 1725 (2020).
31. Cheung, Y. K. *et al.* Are nomothetic or ideographic approaches superior in predicting daily exercise behaviors? Analyzing n-of-1 mhealth data. *Methods Inf. Med.* **56**, 452 (2017).
32. Unnikrishnan, V. *et al.* Predicting the health condition of mhealth app users with large differences in the number of recorded observations-where to learn from? In *International Conference on Discovery Science*, 659–673 (Springer, 2020).
33. Aguilera, A. *et al.* mHealth app using machine learning to increase physical activity in diabetes and depression: Clinical trial protocol for the diamante study. *BMJ Open* **10**, e034723 (2020).
34. Said, A. B., Mohamed, A., Elfouly, T., Abualsaud, K. & Harras, K. Deep learning and low rank dictionary model for mhealth data classification. In *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*, 358–363 (IEEE, 2018).
35. Qureshi, K. N., Din, S., Jeon, G. & Piccialli, F. An accurate and dynamic predictive model for a smart m-health system using machine learning. *Inf. Sci.* **538**, 486–502 (2020).
36. Pryss, R., Reichert, M., Langguth, B. & Schlee, W. Mobile crowd sensing services for tinnitus assessment, therapy, and research. In *2015 IEEE International Conference on Mobile Services*, 352–359 (IEEE, 2015).
37. Pryss, R. *et al.* Applying machine learning to daily-life data from the trackyourtinnitus mobile health crowdsensing platform to predict the mobile operating system used with high accuracy: Longitudinal observational study. *J. Med. Internet Res.* **22**, e15547 (2020).
38. Unnikrishnan, V. *et al.* Entity-level stream classification: Exploiting entity similarity to label the future observations referring to an entity. *Int. J. Data Sci. Anal.* **9**, 1–15 (2020).

39. Pryss, R. *et al.* Prospective crowdsensing versus retrospective ratings of tinnitus variability and tinnitus-stress associations based on the trackyourtinnitus mobile platform. *Int. J. Data Sci. Anal.* **8**, 327–338 (2019).
40. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Fekr, A. R., Janidarmian, M., Radecka, K. & Zilic, Z. Respiration disorders classification with informative features for m-health applications. *IEEE J. Biomed. Health Inform.* **20**, 733–747 (2015).
42. Khatun, F. *et al.* Gender differentials in readiness and use of mHealth services in a rural area of Bangladesh. *BMC Health Serv. Res.* **17**, 573 (2017).
43. Cirillo, D. *et al.* Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* **3**, 1–11 (2020).
44. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000).
45. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **26**, 217–222 (2005).
46. Lavanya, D. & Rani, K. U. Performance evaluation of decision tree classifiers on medical datasets. *Int. J. Comput. Appl.* **26**, 1–4 (2011).
47. Siu, S., Gibson, G. & Cowan, C. Decision feedback equalisation using neural network structures and performance comparison with standard architecture. *IEE Proc. I Commun. Speech Vis.* **137**, 221–225 (1990).
48. Buitinck, L. *et al.* Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint*arXiv:1309.0238 *(2013).*
49. Lameski, P., Zdravevski, E., Mingov, R. & Kulakov, A. Svm parameter tuning with grid search and its impact on reduction of model over-fitting. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, 464–474 (Springer, 2015).
50. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Methodol.* **36**, 111–133 (1974).
51. Stone, M. Asymptotics for and against cross-validation. *Biometrika*. **64**, 29–35 (1977).
52. Mosteller, F., & Tukey, J. W. *Data Analysis and Regression: A Second Course in Statistics* (1977).
53. Ng, A. Y. *et al.* Preventing overfitting of cross-validation data. In *ICML*, Vol. 97, 245–253 (Citeseer, 1997).
54. Dash, M. & Liu, H. Feature selection for classification. *Intell. Data Anal.* **1**, 131–156 (1997).
55. Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 5–32 (1996).
56. Tyler, R. S. & Baker, L. J. Difficulties experienced by tinnitus sufferers. *J. Speech Hear. Disord.* **48**, 150–154 (1983).
57. Vanneste, S., Joos, K. & De Ridder, D. Prefrontal cortex based sex differences in tinnitus perception: Same tinnitus intensity, same tinnitus distress, different mood. *PLoS ONE* **7**, e31182 (2012).
58. Basso, L. *et al.* Gender-specific risk factors and comorbidities of bothersome tinnitus. *Front. Neurosci.* **14**, 706 (2020).
59. Fioretti, A., Natalini, E., Riedl, D., Moschen, R. & Eibenstein, A. Gender comparison of psychological comorbidities in tinnitus patients-results of a cross-sectional study. *Front. Neurosci.* **14**, 704 (2020).
60. Seydel, C., Haupt, H., Olze, H., Szczepek, A. J. & Mazurek, B. Gender and chronic tinnitus: Differences in tinnitus-related distress depend on age and duration of tinnitus. *Ear Hear.* **34**, 661–672 (2013).
61. Richter, K. *et al.* Insomnia associated with tinnitus and gender differences. *Int. J. Environ. Res. Public Health* **18**, 3209 (2021).
62. Ciman, M. *et al.* Smartphones as sleep duration sensors: Validation of the isensesleep algorithm. *JMIR mHealth uHealth* **7**, e11930 (2019).
63. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
64. Buuren, S. v. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**(1), 1–68 (2010).
65. Bradley, M. M. & Lang, P. J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **25**, 49–59 (1994).
66. Kokoska, S. & Zwillinger, D. *CRC Standard Probability and Statistics Tables and Formulae* (CRC Press, 2000).
67. Tate, R. F. Correlation between a discrete and a continuous variable point-biserial correlation. *Ann. Math. Stat.* **25**, 603–607 (1954).
68. Bergsma, W. A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* **42**, 323–328 (2013).
69. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
70. Suykens, J. A. & Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **9**, 293–300 (1999).
71. Hinton, G. E. Connectionist learning procedures. artificial intelligence, 40 1-3: 185 234, 1989. reprinted in j. carbonell, editor. *Machine Learning: Paradigms and Methods*" (MIT Press, 1990).
72. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and regression trees. *Int. Group* **432**, 151–166 (1984).
73. Schlee, W. *et al.* Innovations in doctoral training and research on tinnitus: The European school on interdisciplinary tinnitus research (esit) perspective. *Front. Aging Neurosci.* **9**, 447 (2018).

Acknowledgements

This work was partly funded by the ESIT (European School for Interdisciplinary Tinnitus Research[73]) project, which is financed by European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 722046 and the UNITI (Unification of Treatments and Interventions for Tinnitus Patients) project financed by the European Union's Horizon 2020 Research and Innovation Programme, Grant Agreement Number 848261.

Author contributions

J.A. primarily wrote this paper, created the figures, tables, and trained the machine learning algorithms for gender prediction. W.S., B.L., T.P. and R.P. carefully read and revised the paper. Everybody contributed to the methodology. R.P. supervised the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL. This publication was supported by the Open Access Publication Fund of the University of Wuerzburg.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-96731-8.

Correspondence and requests for materials should be addressed to J.A.

Reprints and permissions information is available at www.nature.com/reprints.