Sensor-Based Detection of Food Hypersensitivity Using Machine Learning

Lennart Jablonski
Institute of Medical Informatics,
University of Lübeck
Lübeck, Germany
l.jablonski@uni-luebeck.de

Torge Jensen
Institute of Medical Informatics,
University of Lübeck
Lübeck, Germany
torge.jensen@student.uni-luebeck.de

Greta Marie Ahlemann
Institute of Nutritional Medicine,
University Hospital
Schleswig-Holstein, Campus Lübeck,
University of Lübeck
Lübeck, Germany
GretaMarie.Ahlemann@uksh.de

Xinyu Huang
Institute of Medical Informatics,
University of Lübeck
Lübeck, Germany
x.huang@uni-luebeck.de

Vivian Valeska Tetzlaff-Lelleck
Institute of Nutritional Medicine,
University Hospital
Schleswig-Holstein, Campus Lübeck,
University of Lübeck
Lübeck, Germany
Vivian.Lelleck@uksh.de

Artur Piet
Institute of Medical Informatics,
University of Lübeck
Lübeck, Germany
ar.piet@uni-luebeck.de

Franziska Schmelter
Institute of Nutritional Medicine,
University Hospital
Schleswig-Holstein, Campus Lübeck,
University of Lübeck
Lübeck, Germany
Franziska.Schmelter@uksh.de

Valerie Sophie Dinkler
Institute of Nutritional Medicine,
University Hospital
Schleswig-Holstein, Campus Lübeck,
University of Lübeck
Lübeck, Germany
valerie.dinkler@student.uni-luebeck.de

Christian Sina
Institute of Nutritional Medicine,
University Hospital
Schleswig-Holstein, Campus Lübeck,
University of Lübeck
Lübeck, Germany
Fraunhofer IMTE
Lübeck, Germany
Christian.Sina@uksh.de

Marcin Grzegorzek
Institute of Medical Informatics,
University of Lübeck
Lübeck, Germany
Fraunhofer IMTE
Lübeck, Germany
marcin.grzegorzek@uni-luebeck.de

ABSTRACT

The recognition of physiological reactions with the help of machine learning methods already plays a major role in many research areas, but is still little represented in the field of food hypersensitivity recognition. The present work addresses the question of how food hypersensitivity can be detected by analysing sensor data with explainable machine learning algorithms. In a first step, the Empatica E4 wristband, a wearable device that can be easily integrated into everyday life, collects raw data on various physiological patterns, and algorithms are implemented to extract a variety of features from the raw data. Subsequently, machine learning methods are used to target this classification problem and examine how food hypersensitivity can be detected using these objectively measurable features. In a subject-independent setup, an accuracy of 91% could be achieved, which provides a promising basis for a new non-invasive and objectively measurable method to detect food hypersensitivity.

iWOAR 2023, September 21–22, 2023, Lübeck, Germany

ACM ISBN 979-8-4007-0816-9/23/09.
https://doi.org/10.1145/3615834.3615845

KEYWORDS

sensor-based, time series analysis, feature engineering, explainable AI, machine learning, classification, random forest, precision nutrition, food hypersensitivity, adverse reaction to food, carbohydrate malassimilation

ACM Reference Format:
Lennart Jablonski, Torge Jensen, Greta Marie Ahlemann, Xinyu Huang, Vivian Valeska Tetzlaff-Lelleck, Artur Piet, Franziska Schmelter, Valerie Sophie Dinkler, Christian Sina, and Marcin Grzegorzek. 2023. Sensor-Based Detection of Food Hypersensitivity Using Machine Learning. In *8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence (iWOAR 2023), September 21–22, 2023, Lübeck, Germany*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3615834.3615845

1 INTRODUCTION

Approximately 30% of the German population suffers from Food Hypersensitivity (FH) [28], [20]. It is common among children [24], [19], young adults [20], and elderly people [12]. The difficulties of identifying such a widespread disease can pose two problems. Firstly, people suffering from FH may not receive the correct diagnosis, and the resulting inadequate treatment leads to a reduced quality of life. Secondly, people with other diseases may be misdiagnosed as having FH, which is particularly devastating for patients with rapidly progressing diseases that need intensive and prompt treatment. Therefore, it is crucial to make a reliable diagnosis of FH. Despite the resulting high importance of a correct diagnosis of FH, a large proportion of affected patients are not diagnosed [12] or FH has not yet been confirmed [20] due to the many difficulties in diagnostics.

FH can be divided into immunologically and non-immunologically caused reactions. While immunologically caused FH are an outcome of specific antibodies in the blood, the non-immunological reactions are mostly based on carbohydrate malassimilation (CM) such as lactose or fructose. The impaired absorption of carbohydrates is either due to an enzyme deficiency or an oversupply with limited capacity of the specific transport systems in the small intestine. If the undigested carbohydrates reach the large intestine, they are metabolised by bacteria to produce gases, including hydrogen, and thus usually trigger dose-dependent symptoms such as flatulence, abdominal pain, and diarrhoea [22], [13]. While immunological FH can be diagnosed in most cases by measurable and specific IgE antibodies in the blood, CM can only be diagnosed by hydrogen breath tests [5]. When undigested carbohydrates are metabolised by bacteria in the large intestine, the resulting hydrogen diffuses into the bloodstream and is exhaled through the lungs, where it can be analysed in the breath. The hydrogen breath test relies on the presence of hydrogen-producing bacteria. However, the colon flora of some patients does not produce hydrogen. So-called "non-H2 producers" are found in 3-25 % and lead to false-negative results of the hydrogen breath test [25], [8].

But not only the presence of non-H2 producers make it difficult to diagnose CM. The complex structure of the human body makes it challenging to determine the cause of a physiological reaction. The physiological reaction can be influenced by many different factors such as immune activation, disturbed intestinal fermentation, enteric dysmotility, post-infectious changes, and psychological disturbances [18] and occur in different ways. Sometimes people exhibit such heightened sensitivity that inhalation alone can trigger FH [23].

The current gold standard for diagnosing FH is the Oral Food Challenge (OFC) [15], [7], [11]. Other common tests usually have either high sensitivity, such as the Skin Prick Test, or high specificity, such as the Basophil Activation Test, but not both at the same time [11]. In addition, there is yet no sensor-based approach applying machine learning to diagnose FH. For this reason, the underlying aim of this work was to identify a new approach for the diagnosis of FH based on objective measurable patterns and combine those with machine learning algorithms.

In this work, the difficulty of diagnosing FH is tackled in two steps. Firstly, it is necessary to find objectively measurable quantities that can be used to draw conclusions about the physiological reactions of the human body. Secondly, these quantities must be analyzed in order to find correlations with FH. With a focus on practicality in everyday life and following the expertise of medical partners the Empatica E4 wristband (Empatica Inc., Cambridge MA, United States) [1] has been chosen for data collection. This wristband allows for collecting valuable physiological data, e.g., blood volume pulse, electrodermal activity, skin temperature, and acceleration, which can be deployed to extract predefined efficient features. For the second step, a classifier is trained to distinguish sequences with a physiological reaction from sequences without a physiological reaction. In this study, a comprehensive analysis was conducted to compare promising classifiers based on the available dataset.

The overall aim of this project is therefore to overcome the limiting factors and identify new objective and measurable biomarkers in the diognaostics of FH.

2 DATA AND PREPROCESSING

2.1 Dataset and Acquisition

The data collection was conducted in collaboration with the Institute of Nutritional Medicine at the University Hospital of the University of Lübeck. To evaluate food hypersensitivity the H2 breath test was utilised which exploits the fact that the processing of indigestible sugars produces hydrogen, which can be measured in the breathing air. In this test, subjects ingest a specific sugar dissolved in water, such as fructose, glucose, lactose, lactulose, document their symptoms and perform several breath tests. First, the subject consumes the specified sugar dissolved in water on an empty stomach and records their symptoms every 20 minutes for a period of 180 minutes in a symptom protocol. The protocol included documenting the start time and noting the subject's symptoms every 20 minutes. It is important to note that although multiple breath tests were conducted during each session, these results were not considered for the subsequent analysis. According to the medical perspective, the accuracy of diagnosing FH based on the H2 values is approximately 70%. Therefore, this paper aims at recognising the physiological reactions of the patients without using the H2 values.

Throughout the duration of the test, the subject wore an Empatica E4 wristband. The wristband included a temperature sensor that measured skin temperature in Celsius (TEMP) at a sampling rate of 4 Hz. It also featured an electrodermal activity sensor (EDA) that measured activity in microsiemens at 4 Hz. Additionally, the wristband had a photoplethysmography that captured the blood volume pulse (BVP) at a sampling rate of 64 Hz. Furthermore, a 3-axis accelerometer measured acceleration in the x, y, and z axes (ACC) at a sampling rate of 32 Hz. These sensor measurements were extracted as one-dimensional data arrays, with the ACC parameter

consisting of three separate arrays (ACC-X, ACC-Y, ACC-Z), one for each axis.

Only one session per individual subject was considered for analysis in order to simplify the allocation into different groups for subject-independent 5-fold cross-validation. Sessions where it is not possible to accurately determine the start time, such as when the documented start time of the test is earlier than the data's start time, are excluded. In addition, inaccuracies in the entries and a higher susceptibility to errors in the labeling due to many changes between positive and negative time periods led to the exclusion of further sessions. Subsequently, 55 of the original 80 sessions remained, 22 of which are negative and 33 of which are positive. All sessions in which the respondent has no symptoms for the entire test are designated as negative. Positive sessions are those in which the subject reports symptoms such as bloating, abdominal pain, diarrhoea and nausea for at least 60 minutes during the test. Individual time windows of a positive session have a negative label if no symptoms were reported at that time. To improve the generalization of the model, it is crucial to develop a population-based model. Consequently, a subject-independent evaluation facility was considered necessary. For this purpose, five groups were created, each with equal proportions of positive and negative time windows, and the remaining data were removed. Each of the five groups consists of four negative and six positive sessions. After the session selection process, the dataset comprised 50 sessions from 50 unique subjects, with 23 positive females, 13 negative females, 7 positive males, and 7 negative males. The average age of the subjects was 54.4 years, with a standard deviation of 14.6 years. All data used for the analysis included the data collected with the Empatica E4 wristband and the symptom protocol.

2.2 Preprocessing

Preprocessing is a vital step that involves trimming the individual data arrays, derived from the Empatica E4, to align with the time span specified in the symptom protocol, and to ensure they are sampled at a uniform frequency.

The preprocessing stage comprises three main steps. Firstly, the data arrays are trimmed temporally to match the time recorded in the symptom protocol. Secondly, all data arrays are scaled to a specified frequency. Lastly, each time point in the data array is labeled. The objective is to generate a matrix that represents features for a specific time window based on the data.

To begin, the data arrays from the wristband are aligned temporally with the symptom protocol. This alignment is achieved by comparing the wristband data arrays' Unix timestamps with the documented start time of the test in the symptom protocol. Subsequently, the data arrays are trimmed to cover the 180-minute time window from the start time. Additionally, the first 15 minutes of each session, which are deemed irrelevant by medical professionals, are discarded.

Since the data arrays are recorded at different frequencies, interpolation is used to scale the data to a specified frequency. This ensures that each time point of each data array has an exact corresponding value in the other data arrays.

Because supervised learning techniques are utilised, it is necessary to assign labels to the training and testing data. These labels are determined based on the symptom protocol, where the subjects recorded their symptoms every 20 minutes. To ensure that the symptoms were due to the testing, only sessions of subjects who exhibited symptoms for a minimum of four consecutive time points, amounting to at least 60 consecutive minutes, were considered. The labeling of these sessions begins with the first recorded positive symptom and ends with the last recorded positive symptom. However, due to the 20-minute interval, it is not possible to precisely determine the exact start and end times of the symptoms. To achieve labeling that is as temporally accurate as possible, a specific temporal radius around the starting and ending points of the positively classified data is removed. This ensures a reduction in potential mislabeling by minimising the impact of uncertainty regarding the exact beginning and ending of symptoms.

2.3 Feature Engineering

Typically, feature extraction is employed to transform data into a lower-dimensional space. 19 handcrafted features [17] are utilised, listed in Table 1. After the preprocessing, each session is partitioned into time windows of a specified duration, and the features were computed for each of these time windows. The features were calculated for each channel (ACC-X, ACC-Y, ACC-Z, BVP, EDA, TEMP), resulting in a feature matrix consisting of 114 columns. The 115 column represents the classification label.

For transitions from positive to negative sections and vice versa, all time windows containing both positive and negative labels were excluded in the feature extraction. This ensures strong labeling where the extracted features are based solely on time windows where the label is either positive or negative so that the classification label is still correct for the features extracted.

Feature Selection plays a crucial role in enhancing model accuracy by eliminating features that decrease accuracy, as well as removing redundant and irrelevant features. This results in a reduced feature set, leading to improved stability of the model performance, decreased complexity and computation time. To achieve this, Recursive Feature Elimination with Cross Validation (RFECV) [21] is employed, which consists of a Random Forest classifier and a stratified 5-fold cross-validation. This approach ensures the mitigation of biases and maintains subject independence while selecting features. RFECV starts by training the estimator on the initial feature set. It then determines the importance of each feature based on an attribute of the estimator and removes the least important features, according to a given step size, from the feature set. This process is repeated recursively until the desired number of features to select is reached. Once RFECV concludes, a ranking of the features can be extracted. In this case, RFECV is used with one feature to select and a step size of one, resulting in a ranking where each rank contains only one feature.

3 CLASSIFICATION APPROACHES

Different classifiers were investigated to distinguish sequences with a physiological reaction from sequences without a physiological reaction. In particular, so-called traditional machine learning methods such as Support Vector Machines, Random Forests, and k-Nearest Neighbors were considered for several reasons.

Table 1: 19 extracted features used for the FH detection

Maximum	Minimum
Zero Crossings	Percentile 20
Percentile 50	Percentile 80
Interquartile range	Mean
Standard Deviation	Mean First Order
Mean Second Order	Normalised Mean First Order
Normalised Mean Second Order	Spectral Energy
Spectral Entropy	Fundamental Amplitude
Peak Amplitude	Peak Frequency
Auto-Correlation	

In the medical domain, the interpretability of the classifier plays a vital role, making it crucial to choose methods that offer high explanatory capabilities. The mentioned methods were identified as suitable choices due to their superior explainability relative to typical deep learning approaches. Moreover, considering the limited size of the available dataset, these selected methods demonstrate the potential to yield satisfactory and dependable outcomes, whereas deep learning models often necessitate extensive amounts of training data to achieve optimal performance. Lastly, it is worth noting that deep learning methods impose significantly higher computational demands compared to traditional machine learning approaches. Previous works have shown the outstanding performance of random forests in detecting reactions of the human body from physiological data [14]. Random forests are resistant to overfitting and robust to outliers due to the randomness in selecting subsets of features for each tree and aggregating many trees into a forest. Especially due to successes in classification tasks, the popularity of random forests has increased greatly [10]. Random forests work efficiently even with large amounts of data and provide very good accuracy among current algorithms [26]. Moreover, other two famous traditional classifiers, i.e., Support Vector Machine and k-Nearest-Neighbors, are taken into account as well, which can as comparative models to demonstrate the effectiveness of the proposed Random Forest. A brief description of the models is as follows:

Random Forest (RF) [6] is an ensemble supervised learning method used for classification and regression tasks. When executed, it produces numerous random decision trees, where each tree assigns the input data to the class with the highest likelihood. Subsequently, the input data is assigned to the class that has received the majority decision from the ensemble of decision trees.

Support Vector Machine (SVM) [27] is a supervised learning approach suitable for both classification and regression tasks. In the case of classification, the algorithm aims to create a maximum margin hyperplane that effectively separates labeled training data. New data is mapped to the same feature space and classified based on its position in relation to the hyperplane. While the kernel trick enables the generation of non-linear hyperplanes, in this work a support vector machine with an RBF (Radial Basis Function) kernel for classification.

K-Nearest-Neighbor (KNN) [9] is a supervised learning algorithm applicable for classification and regression tasks, though mostly used for classification tasks. Its principle is based on the assumption that data points of the same class tend to be located close to each other. To determine the class of a new data point, the algorithm calculates the distances between the point and its k nearest neighbors. The class assigned to the new data point is determined by the majority class among its k nearest neighbors.

Hyperparameter optimisation is a crucial stage that plays a pivotal role in enhancing the accuracy of trained models. To reduce the computational power and time required for optimising hyperparameters, HyperOpt is employed [4], which leverages Bayesian optimisation algorithms. This approach offers an efficient and high-quality solution to the challenge of hyperparameter optimisation. The Hyperopt Python package [2] implements the Tree Parzen Estimator algorithm [3] which is a sequential model-based optimisation (SMBO) approach. A sequential model-based optimisation is an approach in Bayesian optimisation. It follows a sequential trial process, where one trial runs after another and each trial builds upon the knowledge gained from previous ones. This iterative approach utilises the insights obtained from prior trials to guide the selection of hyperparameters to explore in subsequent trials.

4 EVALUATION METRICS

To ensure a fair comparison between the supervised classification methods and the comparability of the results with the current gold standard, the accuracy metric is used. Additionally, the F1 Score is utilised, as it addresses the limitations of the accuracy score [16]. In order to achieve reliable performance estimates and minimise potential biases, a stratified 5-fold cross-validation approach is implemented which involves dividing the dataset into five groups while ensuring that the distribution of positive and negative labels between the groups is equal. Each of these groups serves as a test set once while the remaining groups act as training sets. The process is repeated five times to cover all possible combinations resulting in an averaged accuracy and macro averaged F1 score. This evaluation procedure allows for a realistic assessment of the classification methods' performance. Accuracy [21] is calculated by dividing the number of correct predictions (true positives (t_p) and true negatives (t_n)) by the total number of classifications (true positives (t_p), true negatives (t_n), false positives (f_p), false negatives (f_n)), as illustrated in Eq. 1.

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1}$$

Additionally, the F1 Score (Eq. 4) is used, which represents the harmonic mean of precision (Eq. 2) and recall (Eq. 3) [21].

$$Precision = \frac{t_p}{t_p + f_p} \tag{2}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{3}$$

$$F_1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

5 EXPERIMENTAL RESULTS

All algorithms and models were implemented in Python 3.8. The classifiers and evaluation metrics were implemented using scikit-learn [21]. To evaluate the optimal preprocessing settings for achieving the best results, the dataset underwent preprocessing

Table 2: Overview of all preprocessing settings with associated test accuracy with 80% - 20% split.

removed radius	window length	64 Hz	32 Hz	16 Hz	8 Hz	4 Hz
5	1	**0.857**	0.856	0.818	0.436	0.675
5	2	0.801	0.821	0.776	0.438	0.854
5	5	0.450	0.457	0.460	0.428	0.679
5	10	0.442	0.428	0.432	0.421	0.653
5	15	0.432	0.419	0.428	0.423	0.644
5	30	0.406	0.403	0.404	0.405	0.646
5	60	0.401	0.401	0.401	0.391	0.652
5	120	0.373	0.383	0.387	0.375	0.636
5	300	0.401	0.370	0.401	0.367	0.618
5	600	0.411	0.405	0.373	0.424	0.665
10	1	0.842	0.840	0.831	0.762	0.646
10	2	0.810	0.811	0.789	0.794	0.834
10	5	0.425	0.421	0.442	0.453	0.539
10	10	0.411	0.403	0.413	0.399	0.402
10	15	0.404	0.396	0.402	0.398	0.390
10	30	0.394	0.377	0.385	0.388	0.380
10	60	0.376	0.361	0.357	0.375	0.363
10	120	0.366	0.357	0.367	0.347	0.354
10	300	0.366	0.397	0.376	0.359	0.366
10	600	0.403	0.389	0.410	0.368	0.368

Table 3: Comparison of the best preprocessing settings with subject independent stratified 5-fold cross validation.

frequency in Hz	removed radius	window length	5-fold av. accuracy	5-fold av. F1 Score
32	5	1	0.858	0.86
64	5	1	0.857	0.86
4	5	2	0.854	0.85
4	10	2	0.849	0.85
64	10	1	0.845	0.85
16	10	1	0.844	0.84
32	10	1	0.836	0.84
32	5	2	0.825	0.83
16	10	2	0.819	0.82
64	5	2	0.815	0.81
64	10	2	0.812	0.81
16	5	1	0.803	0.80
32	10	2	0.802	0.80
8	10	2	0.795	0.79
16	5	2	0.771	0.77
8	10	1	0.768	0.77
4	5	5	0.695	0.69
4	5	1	0.680	0.68
4	10	1	0.661	0.66
4	10	5	0.551	0.55
8	10	5	0.459	0.45
8	5	1	0.452	0.44
8	5	5	0.444	0.43
8	5	2	0.441	0.43

with various configurations for the three adjustment settings. Frequency adjustments included 4 Hz, 8 Hz, 16 Hz, 32 Hz, and 64 Hz, chosen based on the range of frequencies in the original data, with 4 Hz being the smallest (EDA, TEMP) and 64 Hz being the largest (BVP). Additionally, only powers of two were considered for frequency adjustment to facilitate interpolation for conversion, given that all original data also possessed powers of two. For the radius removal around transition points from positive to negative and vice versa, two options were tested: 5 minutes and 10 minutes. Finally, feature engineering involved using window lengths of 1, 2, 5, 10, 15, 30, 60, 120, 300, and 600 seconds. To test the range of possibilities, all combinations of the three settings were generated. With five frequency options, two radius removal choices, and ten window lengths, this resulted in a total of $5 \times 2 \times 10 = 100$ different settings.

In the first step, the accuracy was compared depending on the preprocessing settings for all 100 different settings in Table 2. For this purpose, the dataset was divided into subject independent groups of 80% training and 20% test data taking into account that both groups have approximately the same ratio between positive and negative data. Afterwards, the average accuracy of the best preprocessing settings was compared using subject-independent stratified 5-fold cross-validation in Table 3.

With the three best preprocessing settings, an RFECV was performed and the average test accuracy was plotted as a function of the number of features selected. Figure 1 shows this plot for a frequency of 32 Hz which leads to the highest accuracy of 91.16% whereas a frequency of 64 Hz in Figure 2 could obtain an accuracy of up to 91.15% and the frequency of 4 Hz in Figure 3 was resulting in the lowest accuracy of 89.88%. In all cases, the Auto-Correlation of the skin temperature was the single best feature. It was further investigated with a boxplot showing how well this feature separates the two classes of the above 80% - 20% split of the given dataset. The boxplot for the test data is shown in Figure 4 and for the train data in Figure 5. In both cases, the two classes are well separated by this one feature.

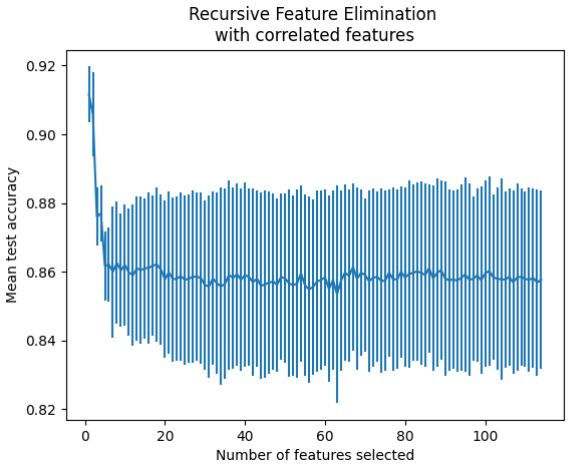

Figure 1: Test accuracy as a function of the number of features selected for the best preprocessing settings (frequency = 32 Hz, window length = 1sec) with 5-fold cross validation.

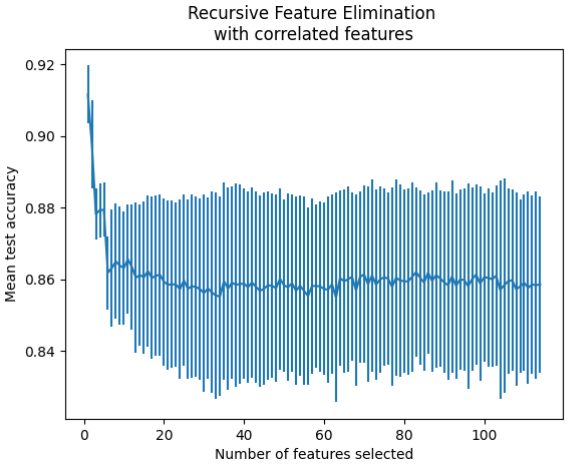

Figure 2: Test accuracy as a function of the number of features selected for the second best preprocessing settings (frequency = 64 Hz, window length = 1sec) with 5-fold cross validation.

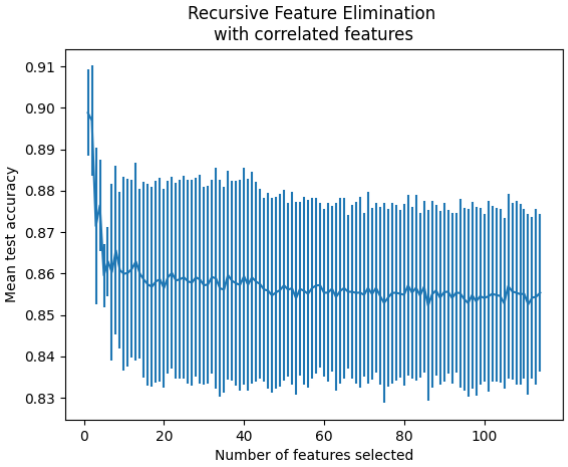

Figure 3: Test accuracy as a function of the number of features selected for the third best preprocessing settings (frequency = 4 Hz, window length = 2sec) with 5-fold cross validation.

After RFECV, an attempt was made to optimise the hyperparameters using the best preprocessing settings. However, the results could not be further improved with a Bayesian optimisation of the hyperparameters.

6 DISCUSSION

As mentioned in the introduction, there are many challenges in diagnosing FH. To address this problem, a way to make FH objectively measurable was sought. After several steps, an appropriate classifier was found that can detect intolerance in food consumption with high accuracy. As previously stated, a large number of objectively measurable features should first be found for the objective recognition of physiological reactions in order to be able

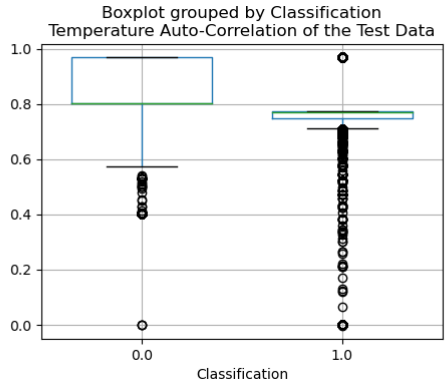

Figure 4: Separation of both classes in the test data by the Auto-Correlation of the skin temperature. The x-axis compares the two classes "negative reaction" (0.0) versus "positive reaction" (1.0) and the y-axis displays the corresponding Auto-Correlation of the skin temperature.

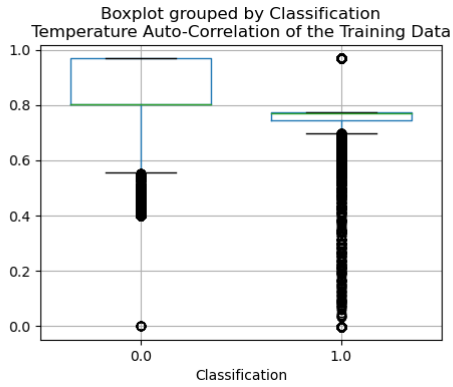

Figure 5: Separation of both classes in the training data by the Auto-Correlation of the skin temperature. The x-axis compares the two classes "negative reaction" (0.0) versus "positive reaction" (1.0) and the y-axis displays the corresponding Auto-Correlation of the skin temperature.

to select the best features later. For this reason, so many different features are extracted from the six different sensor channels during preprocessing, which must initially all be compared in terms of their accuracy in classification. In section 3 multiple reasons were given to choose RF as a classifier. Through RFECV it became clear that this particular classification problem is much simpler than expected, as it was best solved with a single feature. Random forests can be used to solve very complex problems. Nevertheless, it has been shown here that they also have the necessary reliability for very simple problems. Following the RFECV, despite the random factor in the RF, the most important feature occurred in every tree, so that this classification problem could be solved with high accuracy. In order to save computing time, no CV is applied for the comparison of all 100 different preprocessing settings. To avoid a

reduction in the quality of the results, not only the best settings, like for example a removed radius of 5 minutes with a window length of 1 second and a frequency of 64 Hz, but also neighbouring time windows, for instance a window length of 2 seconds with the above removed radius and frequency, are taken into account when comparing the best preprocessing settings with subject independent stratified 5-fold CV. The RFECV was performed with the three best preprocessing settings.

One feature turned out to be particularly interesting as it alone appears to best separate the classes. Therefore the Auto-Correlation of the skin temperature was investigated further. For comparison, the average accuracy for all features without the most important one was calculated. The latter was 84%, which is understandable when the best feature alone leads to an accuracy of 91% and, together with the subsequent features, settles at around 86%. Moreover, the separation of both classes by the best feature was displayed in a boxplot. It can be seen clearly that this feature separates the two classes well. The skin temperature is strongly influenced by the blood flow in the peripheral blood vessels. If the blood vessels are dilated, more blood flows through them so that the body releases more heat and the core temperature thus cools down. If the blood vessels contract, less blood flows through them and the body releases less heat through the skin. These processes keep the core temperature of the body stable. The body reacts to external stimuli such as heat or cold by dilating or constricting the peripheral blood vessels, which results in an increase or decrease in skin temperature. If external influences like the environmental conditions remain constant, a change in skin temperature, therefore, suggests stimuli coming from within the body. If the core body temperature changes when eating food to which there is hypersensitivity, it can be assumed that this change influences the skin temperature. Food hypersensitivity triggers a reaction of the body to food. Based on the experiments conducted, it can be assumed that this reaction affects the Auto-Correlation of skin temperature, which describes the change in skin temperature over time.

7 CONCLUSION

For the objective and everyday detection of food hypersensitivities, a number of different features were first determined and then the most important features were identified. It could be shown that a suitable selection of preprocessing settings and features can lead to an accuracy of 91%. The number of features could be drastically reduced to one feature, which reduces the computation time and thus increases the real-time suitability. All the necessary data was collected continuously and non-invasively using a simple smartwatch, making it very easy to integrate into everyday life and making all the necessary tools freely available to everyone. To further improve the performance, it could be investigated to explore other features which might be more correlated to food hypersensitivity. Moreover, other classifiers could be compared and the Bayesian optimisation of hyperparameters could be tried out with more evaluations. The data collection is still ongoing and a larger dataset will facilitate the use of more complex deep learning models such as recurrent neural networks or transformers.

ACKNOWLEDGMENTS

The scientific work leading to this paper has been financed by German Federal Ministry of Education and Research (BMBF) within the grant INDICATE-FH with the number 01EA2109A.

We thank Nele Ziebell (Institute of Medical Informatics, University of Lübeck) for her assistance in collecting, preparing and summarising the data.

REFERENCES

[1] [n. d.]. *Empatica E4 Wristband.* Retrieved May 29, 2023 from https://www.empatica.com/research/e4/

[2] [n. d.]. *Hyperopt Python library GitHub Page.* Retrieved June 06, 2023 from https://github.com/hyperopt/hyperopt

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

[4] Yamins D. Cox D. D. Bergstra, J. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *TProc. of the 30th International Conference on Machine Learning (ICML 2013)* (2013), pp. I–115 to I–23. http://proceedings.mlr.press/v28/bergstra13.pdf

[5] Barbara Braden, C. Braden, M Klutz, and Bernhard Lembcke. 1993. [Analysis of breath hydrogen (H2) in diagnosis of gastrointestinal function: validation of a pocket breath H2 test analyzer]. *Zeitschrift fur Gastroenterologie* 31 4 (1993), 242–5.

[6] L. Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[7] Mauro Calvani, Annamaria Bianchi, Chiara Reginelli, Martina Peresso, and Alessia Testa. 2019. Oral Food Challenge. *Medicina* 55 (2019). https://doi.org/10.3390/medicina55100651

[8] Stefan Ulrich Christl, Peter R. Murgatroyd, Glenn R. Gibson, and John H. Cummings. 1992. Production, metabolism, and excretion of hydrogen in the large intestine. *Gastroenterology* 102 4 Pt 1 (1992), 1269–77.

[9] T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (1967), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[10] Antonio Criminisi and Jamie Shotton. 2013. Decision Forests for Computer Vision and Medical Image Analysis. In *Advances in Computer Vision and Pattern Recognition*.

[11] R X Foong, Jennifer A Dantzer, Robert A. B. Wood, and Alexandra F. Santos. 2021. Improving Diagnostic Accuracy in Food Allergy. *The Journal of Allergy and Clinical Immunology. in Practice* 9 (2021), 71 – 80. https://doi.org/10.1016/j.jaip.2020.09.037

[12] A. Fujimori, Tomomi Yamashita, Masaru Kubota, Hiromi Saito, Nobue Takamatsu, and Mitsuhiko Nambu. 2015. Comparison of the prevalence and characteristics of food hypersensitivity among adolescent and older women. *Asia Pacific journal of clinical nutrition* 25 4 (2015), 858–862. https://doi.org/10.6133/apjcn.092015.39

[13] Miriam Goebel-Stengel, Andreas Stengel, M. Schmidtmann, Ivo R. van der Voort, Peter Kobelt, and Hubert Mönnikes. 2014. Unclear Abdominal Discomfort: Pivotal Role of Carbohydrate Malabsorption. *Journal of Neurogastroenterology and Motility* 20 (2014), 228 – 235.

[14] Muhammad Tausif Irshad, Muhammad Adeel Nisar, Xinyu Huang, Jana Hartz, Olaf Flak, Frédéric Li, Philip Gouverneur, Artur Piet, Kerstin M Oltmanns, and Marcin Grzegorzek. 2022. SenseHunger: Machine Learning Approach to Hunger Detection Using Wearable Sensors. *Sensors* 22, 20 (2022), 7711. https://doi.org/10.3390/s22207711

[15] Kirsi M. Järvinen and Scott H. Sicherer. 2012. Diagnostic oral food challenges: procedures and biomarkers. *Journal of immunological methods* 383 1-2 (2012), 30–8. https://doi.org/10.1016/j.jim.2012.02.019

[16] Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. 2014. Consistent Binary Classification with Generalized Performance Metrics. (2014).

[17] Frédéric Li, Kimiaki Shirahama, Muhammad Adeel Nisar, Lukas Köping, and Marcin Grzegorzek. 2018. Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors. *Sensors* 18, 2 (2018). https://doi.org/10.3390/s18020679

[18] Gülen Arslan Lied, Kristine Lillestøl, Ragna A Lind, Jørgen Valeur, Mette Helvik Morken, Kirsi Vaali, Kine Gregersen, Erik Florvaag, Tone Tangen, and Arnold Berstad. 2011. Perceived food hypersensitivity: A review of 10 years of interdisciplinary research at a reference center. *Scandinavian Journal of Gastroenterology* 46 (2011), 1169 – 1178. https://doi.org/10.3109/00365521.2011.591428

[19] Eva Östblom, Ann-Charlotte Egmar, A N N Gardulf, Gunnar Lilja, and Magnus Wickman. 2008. The impact of food hypersensitivity reported in 9-year-old

children by their parents on health-related quality of life. *Allergy* 63 (2008). https://doi.org/10.1111/j.1398-9995.2007.01559.x

[20] Morten Osterballe, C G Mortz, Tine K Hansen, K. E. Andersen, and Carsten Bindslev-Jensen. 2009. The Prevalence of food hypersensitivity in young adults. *Pediatric Allergy and Immunology* 20 (2009). https://doi.org/10.1111/j.1399-3038.2008.00842.x

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[22] Martin Raithel, Michael Weidenhiller, Alexander F. Hagel, Urban Hetterich, Markus Friedrich Neurath, and Peter Christopher Konturek. 2013. The malabsorption of commonly occurring mono and disaccharides: levels of investigation and differential diagnoses. *Deutsches Arzteblatt international* 110 46 (2013), 775–82.

[23] Daniel A. Ramirez and Sami L. Bahna. 2009. Food hypersensitivity by inhalation. *Clinical and Molecular Allergy : CMA* 7 (2009), 4 – 4. https://doi.org/10.1186/1476-7961-7-4

[24] Carina Venter, Brett Pereira, Jane D Grundy, C. Bernie Clayton, S. H. Arshad, and Tara Dean. 2006. Prevalence of sensitization reported and objectively assessed food hypersensitivity amongst six-year-old children: A population-based study. *Pediatric Allergy and Immunology* 17 (2006). https://doi.org/10.1111/j.1399-3038.2006.00428.x

[25] Piero Vernia, Mauro Di Camillo, Vanessa Marinaro, and Renzo Caprilli. 2001. Effect of predominant methanogenic flora on the outcome of lactose breath test in irritable bowel syndrome patients. *European Journal of Clinical Nutrition* 57 (2001), 1116–1119.

[26] Mohammed Zakariah. 2014. Classification of large datasets using Random Forest Algorithm in various applications: Survey.

[27] Yongli Zhang. 2012. Support Vector Machine Classification Algorithm and Its Application. (2012), 179–186. https://doi.org/10.1007/978-3-642-34041-3_27

[28] Torsten Zuberbier, Günter Edenharter, Margitta Worm, I. Ehlers, S Reimann, Thomas Hantke, Christoph Roehr, Karl E. Bergmann, and Bodo Niggemann. 2004. Prevalence of adverse reactions to food in Germany – a population study. *Allergy* 59 (2004). https://doi.org/10.1046/j.1398-9995.2003.00403.x