UNIVERSITY OF TWENTE.

**Faculty of Electrical Engineering,
Mathematics & Computer Science**

Classification of eating gestures using a wrist worn IMU and the deep learning model InceptionTime.

van Loh, Soenke Ulf

**B.Sc. Thesis
July 2022**

Commitee:
van Beijnum, Bert-Jan, dr.ir.
Klaassen, Randy, dr. MSc
Haarman, Juliet
Mevissen, Sigert

Biomedical Signals and Systems Group
Faculty of Electrical Engineering,
Mathematics and Computer Science
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

Contents

1 Abstract

This paper deals with the classification of eating related gestures using 3 axis accelerometer and gyroscope data from a wrist worn Inertial Measurement Unit (IMU) for the purpose of dietary monitoring. The data is used to train the convolutional neural network (CNN) InceptionTime. It is gathered in an experiment consisting of 9 participants and contains 8 classes which the network needs to classify. The data is fed to the network as a multivariate time series (MTS) which means that the 6 different channels of measurements are treated as one time series in classification. The results of the experiments are compared to results from the master thesis of Sigert Mevissen as an extension of his work [1]. It is confirmed that the approach of using a CNN for classification of MTS is applicable in the case of eating gesture recognition. On a set containing all 9 participants, 69% F1 score is achieved. When combining the eating and non-eating gestures into a binary classification, this increases to 80%. In three leave one subject out (loso) tests, F1 score of 63% on average are achieved. The SVM from the master thesis, trained on another dataset, achieved 82% F1 score on a full and 18% F1 score on loso test sets.

2 Introduction

Obesity and overweight are an increasing problem. While it is well known that countries like the USA have problems with it, more countries follow this trend. By now, some scientists speak about a global epidemic [2]. The World Health Organization summarizes it as: “Worldwide obesity has nearly tripled since 1975. [. . .] 39% of adults aged 18 years and over were overweight in 2016, and 13% were obese. [. . .] Obesity is preventable” [3]. There are multiple diseases which are directly connected to obesity, like cardiovascular diseases or diabetes [2], [3]. The two main reasons for obesity and overweight are an increased intake of sugars and fats and a decrease in energy expenditure [3]. Therefore, to decrease the problem, a balance of energy consumption and usage should be established. One way to help people do that is tracking the food intake and physical activity. Here we will focus only on tracking the food intake. A common technique to track food intake is to use journals where people note down what they eat and when.

However, this technique, developed in the 1940s, shows low accuracy’s over time “due to manual logging labor requested from respondents” [4]. The results are therefore, especially for long term measurements, influenced by willpower, memory and cognition of the person that is logging them. Additionally, food intake is such a common thing in our everyday life that it is easy to forget logging it every time [4], [5].

Therefore, automatic dietary monitoring is getting more and more attention, trying to develop sensor systems which can track dietary intake or help to track it [1], [4]–[6]. A common component of most of these systems is an Inertial measurement unit (IMU) placed on the wrist of the participant. This could for example be a smartwatch, these contain such units. The main goal is to find out if CNN are a suitable approach of detecting gestures from a multivariate time series (MTS) consisting of 3-axis accelerometer and gyroscope data. Additionally, it is investigated how they perform compared to support vector machines (SVM), a more classical approach of machine learning which

uses human engineered features from the 6 axis of measurement. This approach is used in the Master Thesis of Sigert Mevissen [1]. Parts of this thesis are build upon his work.
In the master thesis, the data from the existing dataset is classified using a combination of human feature engineering and support vector machine (SVM). Only the results will be displayed in table 1 for comparison purposes. The data here consists of 13s pre- and 13s appended to the data point which is to be classified. No resampling was done prior to that.

dataset	accuracy	precision	recall	F1
SVM	-	0.88	0.77	0.82

Table 1: metrics for Sigerts data classified on SVM

From figure 1 the ratio between full and loso set regarding the F1 score can be calculated to be $0.18/0.5 = 0.36$. This will be used for comparison of generalizability of the two networks.

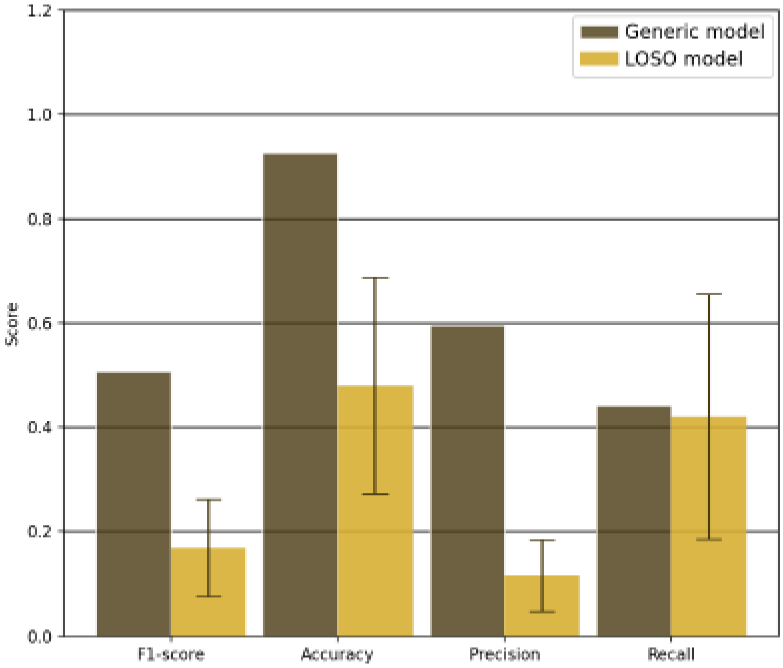


Figure 1: The F1-score, accuracy, precision and recall of eating gestures of the generic SVM model compared to the loso SVM model. The LOSO model scores are averaged with the standard deviation as error bars. 7s were pre- and appended to the datapoints.

3 Related Work

There are two main ways in which eating detection is desired to help with dietary monitoring. The first approach consists of a complex sensor system which can involve different classes of sensing such as acoustic sensors for swallowing, chemical sensors in the mouth in order to classify properties of the food, cameras for image classification

and IMUs for gesture recognition of the arm/hand etc. [4], [6], [7]. These systems have to deal with high amounts of input data and in most cases these are classified independently and then combined in algorithmic approaches like finite state machines, similar to the work of Sigert Mevissen [1].This approach becomes very complex as there is a high variability in classes which need to be predicted. The final goal is to do the journaling for the person, to some extent remove the person from the loop. The other common method discussed in the literature is a system which uses a smaller amount of sensors, such as cameras and IMUs. Here the goal is rather to remind the subject of food journaling, for example by sending a reminder when a gesture is classified as eating [5]. This needs more involvement of the person itself, but miss classifications of the networks can be compensated to some extent. The number of classes in the simplest case can be a binary classification of "eating" and "not eating". However, for these systems, same as for the more complex ones, the number of variables such as sound, visual environment and movements is complex. In all the mentioned papers the most used classification approach for the IMU data, mostly accelerometer measurements, is to use support vector machines (SVM) for the classification.
An alternative to using SVM's for the classification of movement data are deep learning networks. Deep learning approaches such as CNN's for the analysis of univariate time series (UTS) and multivariate time series (MTS) are already established. In their paper, "Deep Learning for Classifying Physical Activities from Accelerometer Data. Sensors (Basel)" [8] Nunavath V. et al. investigate different networks for the classification of accelerometer data using deep learning approaches. They achieve accuracy and F1 scores of beyond 90% with different deep learning architectures, also including various CNN architectures. In their paper "InceptionTime: finding AlexNet for time series classification" [9] Ismail Fawaz et al. propose a special CNN structure for the classification of time series. They compare their network to other state-of-the-art time series classification methods using the UCR-Archive. The UCR archive is a collection of different annotated univariate time series [10]. They show that their architecture is comparable or even exceeds other state-of-the-art approaches for classification of time series. Ruiz AP et al. do a similar comparison in their paper "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances." [11] On the UEA archive. Similar to the UCR archive, the UEA archive is a collection of multivariate time series [12]. In their paper, they also include InceptionTime, and it belongs to the top performing networks.

4 Materials and Methodology

In order to extend the Work from the Master thesis, the following things will be done:

1. Train the InceptionTime ensemble network on data from the master thesis
2. Collect eating and non-eating related gestures in an experiment in order to build a dataset for gesture classification which is more balanced and consistently sampled
3. Training the InceptionTime ensemble network on the newly gathered data

4.1 Learning from existing data

Before data from the experiment in the master thesis [1] can be fed into a CNN, it needs some preprocessing. The available data consists of 3 axis accelerometer and 3 axis gyroscope data, both in raw and high pass filtered form. The data was collected with a smartwatch connected to a phone via a Web Socket. Even though the CNN is supposed to be trained on raw data, an inspection of the data and some preprocessing is done to ensure that the data has the right shape. The measurements are 83Hz on average. Therefore, a data frame with 83 entries per second of measurement for each axis of measurement is expected. Additionally, a structure which supplies labels for these data points is supplied. Since 83Hz was an average value, measurements differed in samples per second. As the network needs consistent length inputs, the data is resampled to 50Hz using linear interpolation. The classes in the dataset are unbalanced. Since the experiment is done not only for eating gestures but also aimed at chewing and swallowing. Due to the additional sensor measurements, the “other” class, which describes no specific movement with the hand, is a majority class of the data for gesture recognition. The two ways used to counteract this behavior are removing samples of the majority class and applying weights to each class in the loss function when training the network.

4.2 Experiment

A new experiment is designed to gather data especially for gesture classification. The goal is to have more consistent measurements regarding samples per second and to have a more balanced dataset since the experiment is only aimed at gesture detection and no chewing or swallowing tasks are implemented. To have good sampling quality, the “MetaMotionS” [13] sensor is used in the experiment. It is configured to measure 3 axis accelerometer and gyroscope data at 50Hz. The measurement axis is therefore the same as in the prior experiment, and the sampling frequency is the same that is used after resampling. The data from the sensor is streamed directly to the phone, which is used to record the video for later annotation of the data. The data is saved in two CSV files, one for the accelerometer and one for the gyroscope data. Besides the data, the CSV files also contain timestamps. Using the timestamps, the two measurement series can be synchronized. There is a slight offset between them ($< 8ms$) which is removed.

The experiment consists of 7 different gestures, 3 eating related ones:

- eating with hand
- eating with spoon
- eating with fork

and 4 non-eating related:

- random arm movement
- using the phone
- writing on paper
- scratching the head

All gestures are performed sitting down at a table to have a controlled and comparable environment for all participants. All data points which do not belong to a specific movement of the list are regarded as the “other” class. The gestures are chosen because they represent typical movements that are done sitting down at a table in everyday life. An eating gesture involving knife and fork was excluded as people are not consistent in holding knife or fork with their dominant hand and someone holding the knife vs someone holding the fork in the hand with the sensor would be hard to compare. The experiment is performed on 9 participants of varying age and gender.

# participants	9	Age mean	39.625
# male	6	Age std	22.129

Table 2: Information about participants of the experiment

All participants performed the experiment with their dominant (right) hand. In the beginning, every participant clapped their hands 3 times to synchronize the recorded video with the recorded movement data. From the synced video, annotation dictionaries for labeling of the data are created. Every full second in which a gesture is performed is annotated with that gesture. For example, if someone started bringing a chip to their mouth at 10.5 seconds, the whole 10th second is annotated as that gesture. When multiple seconds are concatenated either the majority label of the window or the original, so not appended or prepended second, is used for labeling.

4.3 Processing experiment data

Before the data can be fed into the CNN for training, preprocessing must be done. This consists of two main steps. First, the shape of the data must be changed, and then the data must be z normalized. For the network to train on the data, it needs a predefined shape. The raw data is a 2D frame 2a where:

- each row represents a timestamp
- each column represents a value for that timestamp

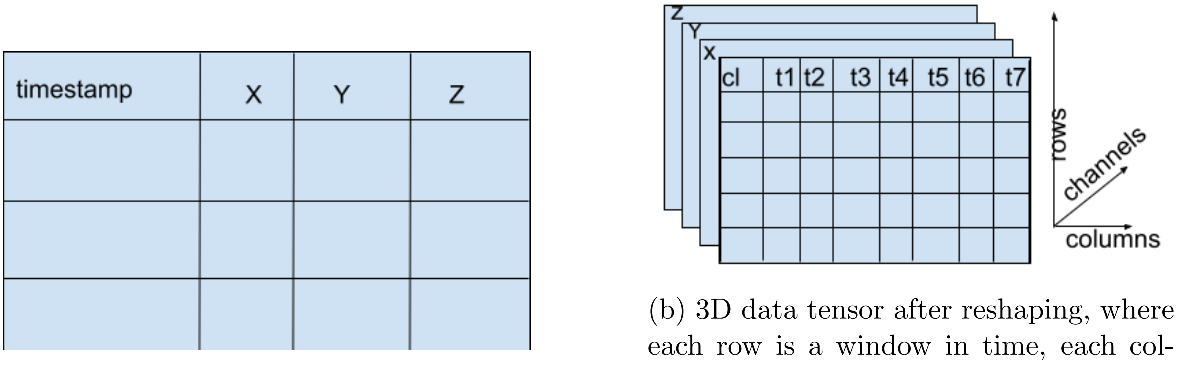
For training a 3 dimensional shape 2b is needed where:

- each row represents a time span (certain amount of measurement values)
- each column represents one measurement point in that time span, except the first one which represents the class of measurement
- each channel represents one axis of measurement

During this process, multiple seconds can be included in a time span and the sampling frequency can be altered using up/down sampling. The final shaping step before the data can be used for training is to make train and test splits of the data. For this, the “train-test-split” function from the sklearn library is used. A stratified split is done, this ensures that all classes are equally distributed over train and test sets and that the data is shuffled. Multiple persons train and test sets can be concatenated. All data sets consists of time series where 4s are prepended to a data point and the data is resampled to 25Hz. Labeling is done based on the majority label. The final data sets used are:

1. full set with stride 1
2. full set with stride 5
3. loso set 1
4. loso set 2
5. loso set 3

A loso set describes a set where the test data is one person's data and the train data is all other persons' data. The number of loso sets is limited due to limited computation time available in training. The loso sets are done to test how well the network can generalize to eating gestures in general and not only to a person's behavior. After the data has the right shape, it is z-normalized. Z-normalization is a common step in data processing for convolutional and deep learning networks. It is done for all data sets in the Inception Time paper as well. In z-normalization, the data is shifted in a way that its mean becomes zero and its variance 1. Z-normalization is only fitted on the train sets and afterwards applied to train and test sets. This prevents leakage of information from the test sets into training.



(a) 2D data tensor from sensor, where each column represents a measurement value and each row represents a measurement point in time

(b) 3D data tensor after reshaping, where each row is a window in time, each column a measurement, besides the first one which is the class of that measurement, in that window and each channel an axis of measurement

Figure 2: Data tensor shapes

4.4 The CNN network

The network which is used for classification of the data is "InceptionTime" [9]. It is a convolutional neural network (CNN).

CNN's make use of convolution filters. These take in a certain length of data points, dependent on the kernel size. They learn which weight to apply to these filters to recognize patterns in the data. The outputs of these layers are then mostly connected to Pooling layers to reduce the number of outputs. The last layer is often a single dense layer, which represents the different output classes.

The inception time network was chosen due to its performance on the UCR-Archive [10]. Additionally, Ruiz A.P. et al. found it to be under the top performing networks for multivariate time series classification in their paper “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances” [14].

4.4.1 InceptionTime Architecture

InceptionTime is based on the popular ResNet architecture [15] which revolutionized image classification. It is an ensemble network of convolutional neural networks. Every one of these consists of multiple inception blocks. The last layer is a dense layer which has one neuron per output class. This structure is depicted in 3. Like ResNet there are residual connections which bridge inception blocks to prevent the vanishing gradient problem. Every inception block follows the same architecture. The architecture of an inception block can be seen in 4. Every inception block takes a multivariate time series of length m and depth d as input. By using a bottleneck 1D convolution layer with stride 1 and kernel size 1 the depth is reduced to $d‘ < d$ while maintaining length m. This is then run through 1D convolutions with different kernel sizes in parallel. In order to add more variability, a Max Pooling layer is applied to the original input of size m, d and after that using the bottleneck reduced to $m, d‘$. All outputs of the convolutions and Max Pooling are concatenated into a new MTS.

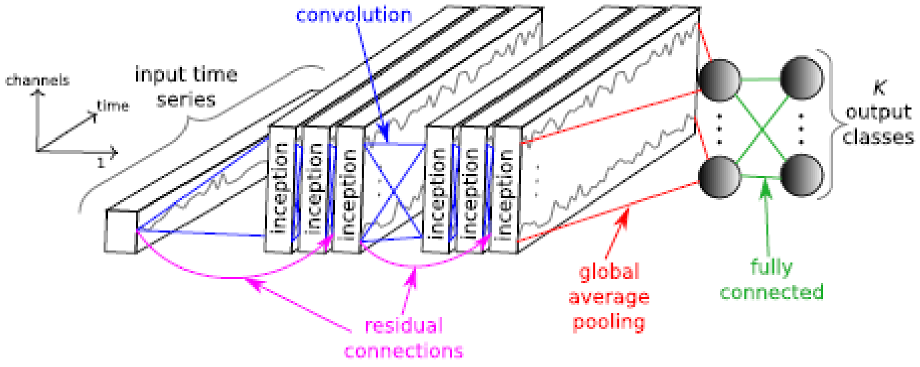


Figure 3: Overview of the Inception Network

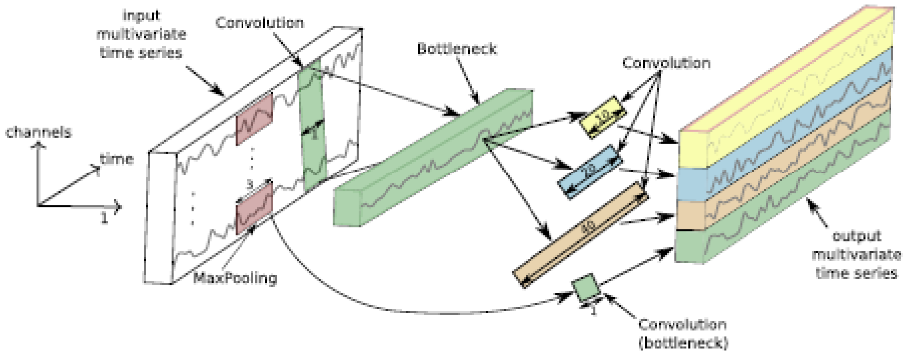


Figure 4: Overview of an Inception Block

The figures 3 and 4 are taken from the InceptionTime paper [9].

4.4.2 CNN for MTS

The Inception Time network was originally designed for UTS applications; however, due to every inception block starting with the bottleneck layer, MTS with any channel depth can be used as inputs as well. This leads to the first step being a 1x1 convolution with d‘ filters across the input channels. Therefore, only the very first operation in the network differs between UTS, where no bottle necking must be applied to the input, or MTS input. This leads to the fact that adding channels in the input does not increase the computation time of the network notably even though the amount of information is increased.

4.5 Metrics

To compare the different results from the experiments, scores and metrics are used, which can be calculated using the labels of a test set and the according predictions the network makes based on that test set. This can be visualized as a confusion matrix. The metrics used are accuracy, precision, recall and F1 score. These can all be calculated from the confusion matrix. The main metric for comparing the network results are the F1 score, since they include precision and recall of the networks and exclude accuracy. Accuracy is needed as a metric for describing the networks and their results, especially when looking at the unbalanced data sets and their influence.

4.6 Training setup

The network is trained on both data from the experiment that was done for the master thesis and data from the experiment designed especially for gesture recognition. A comparison of 6 data sets overall. First the CNN which is trained on the data from Sigerts experiment. It is trained and tested on all participants. For the new experiment, the 5 data sets described in 4.3 are used. For all these trainings the standard parameters of the InceptionTime Network were used, besides for the one on Sigerts data which used weights in training due to the uneven distribution of classes.

5 Results

5.1 Data

During the thesis, a dataset is created to train the inception time network on classifying eating gestures.

Every participant performed 8 different gestures. 3764 seconds of data were gathered. Around 40% of this data fall into the “other” class, which describes no concrete gesture. All remaining gestures are 7-10% of the dataset each. Therefore, the dataset is unbalanced, which must be considered when analyzing the training results later. The data was recorded at 50Hz.

Additionally, the dataset from the master thesis is used. Here, around 80% of all samples belong to the “other” class. The dataset contains 4 different gestures besides the ”other” class. To reduce the effect of the imbalance the dataset was resampled so that, by dropping data points of the majority class, 55% belong to the “other” class for the training of the classifier. The data was recorded at 83Hz; however, there were big differences in between seconds regarding the amount of data points. 12 seconds had less than 61 data points, 26 seconds less than 71 data points and 317 seconds less than 81 data points. Some seconds also had more than 100 data points. The set was resampled using linear interpolation.

5.2 Training Results

The metrics that are shown and described are from the ensemble network, meaning a combination of 5 independently initialized and trained InceptionTime Networks. All metrics are macro averaged over all classes, meaning that the metrics are calculated per class and then the (unweighted) average is calculated.

Table 3 shows the metrics discussed in section 4.5 for all 5 data sets from the new experiment. There is a difference in the scores, especially between the full set with stride 1 and the other data sets. Additionally, there is a significant fall off in precision, recall and F1 when comparing the new dataset to the old one.

dataset	accuracy	precision	recall	F1
Full (stride 5)	0.68	0.68	0.7	0.69
Full (stride 1)*	0.92	0.93	0.93	0.93
Loso 1	0.58	0.63	0.58	0.61
Loso 2	0.7	0.77	0.67	0.72
Loso 3	0.61	0.55	0.55	0.55
Loso averaged	0.63	0.65	0.6	0.63
Master thesis data	0.53	0.26	0.29	0.27

Table 3: Metrics for the 5 data sets from the new experiment and data from the master thesis

*There is possibility of information leakage between train and test set, therefore this result needs to be cautiously evaluated

The difference in the full data sets is in the stride that was used in order to build the windows. The one with stride 1 has 5 times more data points. Since there is overlap in the windows, it is possible that there are data points from the train set also partially present in the test set. There are no full-duplicated windows from train to test set.

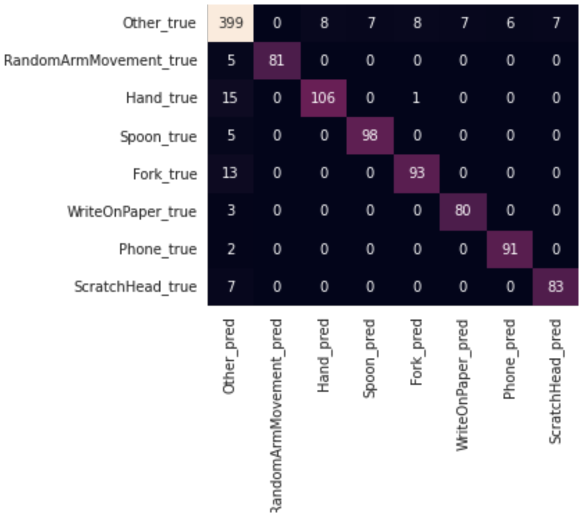
In the loso sets, there is a difference of up to 12% in accuracy and 17% in F1 score. The network predicts different persons with different F1 scores. Precision and recall deviate less than 5% points in 3 out of the 4 tests. The F1 score is always in the middle of precision and recall.

The Figures 7 and above show the confusion matrices of the 6 different data sets. One thing that all these matrices have in common is that most of the false positives and false negatives occur for the "other" class. The network misclassifies gestures as "other" and also "other" as gestures, mostly when it makes a mistake. Besides predicting the "other" class, there is an increased miss prediction in between eating gestures. Miss predictions between eating and non-eating related gestures are low compared to miss predictions in between eating gestures. An extreme case that should be mentioned is

that in loso set 3 7a all phone usages are miss predicted as the other class. Two out of the nine test persons used the phone with two hands. The test person from loso set 3 was one out of these two. A comparison of the 3 loso set F1 scores to the full test set with stride 5: 0.55/0.69 = 0.8, 0.61/0.69=0.88 and 0.72/0.69=1.04 on average 0.91.
From confusion matrix 7b the dataset from the master thesis is very unbalanced, so accuracy will not be used as a score for comparison. Precision and recall are 3% points apart. The F1 score is 42% points lower than for the full dataset from the new training data with stride 5. It is also 28% points lower than the lowest score out of the 3 loso sets. In the confusion matrix, there is no main diagonal.
Additionally, to the macro averaged scores from table 3 one can look at a combining all eating gestures in one class, all non-eating gestures in one class and calculate the metrics for these. This can be done from the confusion matrices. The according results are in table 4.

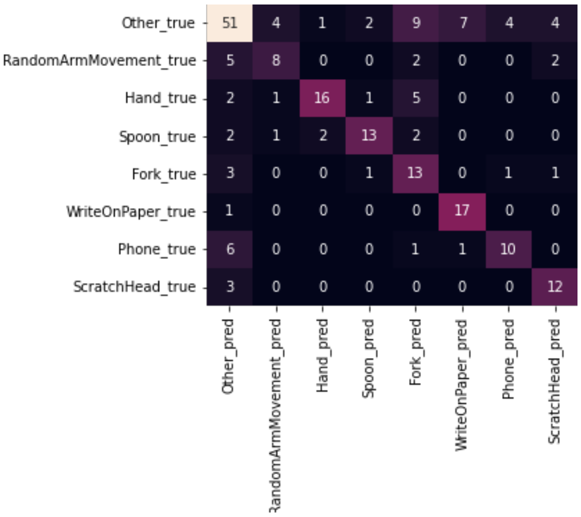
dataset	precision	recall	F1
Full (stride 1)*	0.93	0.9	0.91
Full (stride 5)	0.78	0.83	0.8
Loso 1	0.63	0.55	0.59
Loso 2	0.74	0.64	0.69
Loso 3	0.72	0.62	0.66
Loso averaged	0.7	0.6	0.65
Master thesis data	0.04	0.1	0.06

Table 4: Metrics for the 5 data sets from the new experiment and data from the master thesis when all eating gestures and non-eating gestures are combined in a binary classification problem
*There is possibility of information leakage between train and test set, therefore this result needs to be cautiously evaluated



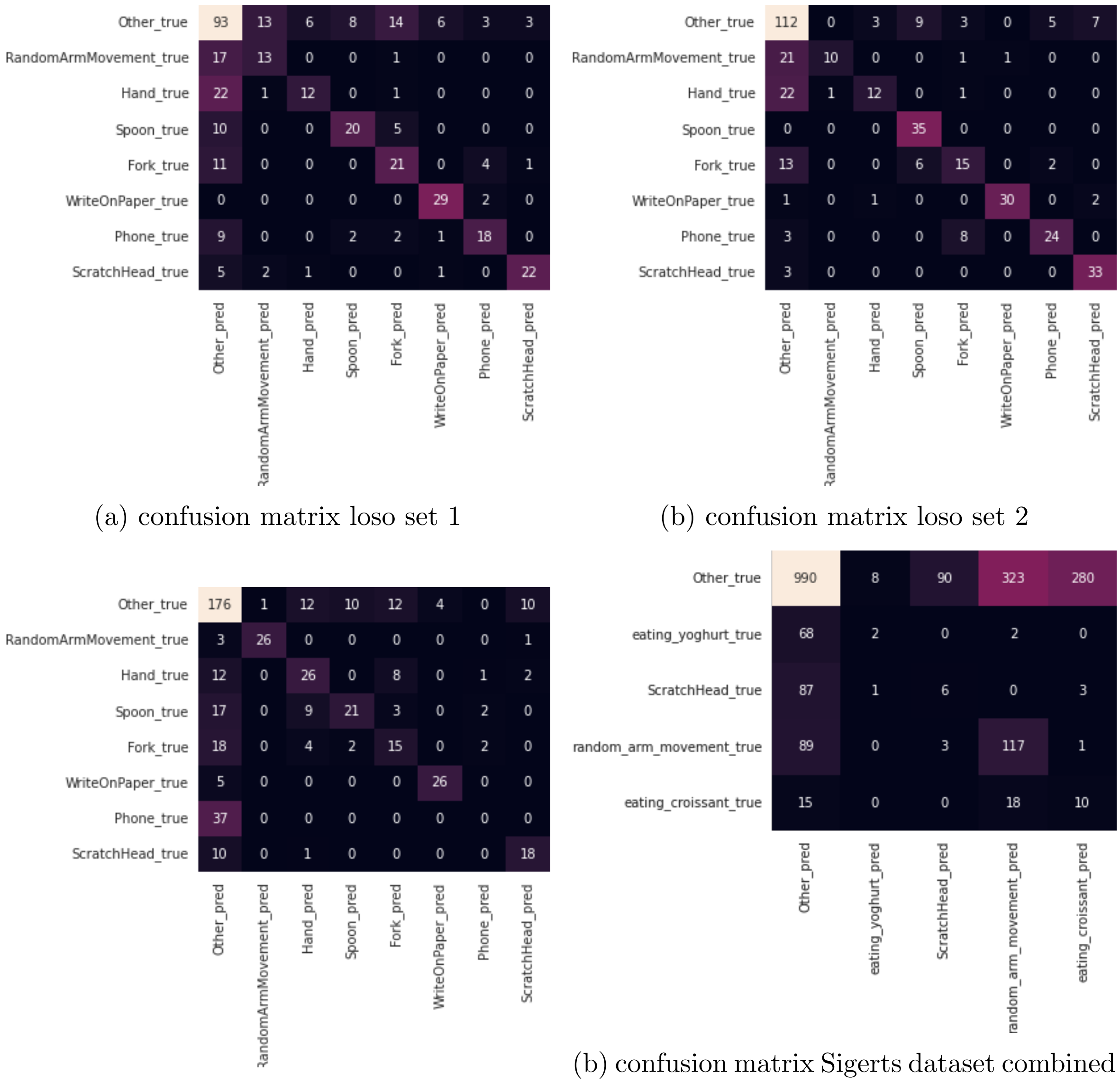
	Other_pred	RandomArmMovement_pred	Hand_pred	Spoon_pred	Fork_pred	WriteOnPaper_pred	Phone_pred	ScratchHead_pred
Other_true	399	0	8	7	8	7	6	7
RandomArmMovement_true	5	81	0	0	0	0	0	0
Hand_true	15	0	106	0	1	0	0	0
Spoon_true	5	0	0	98	0	0	0	0
Fork_true	13	0	0	0	93	0	0	0
WriteOnPaper_true	3	0	0	0	0	80	0	0
Phone_true	2	0	0	0	0	0	91	0
ScratchHead_true	7	0	0	0	0	0	0	83

(a) confusion matrix full set stride 1



	Other_pred	RandomArmMovement_pred	Hand_pred	Spoon_pred	Fork_pred	WriteOnPaper_pred	Phone_pred	ScratchHead_pred
Other_true	51	4	1	2	9	7	4	4
RandomArmMovement_true	5	8	0	0	2	0	0	2
Hand_true	2	1	16	1	5	0	0	0
Spoon_true	2	1	2	13	2	0	0	0
Fork_true	3	0	0	1	13	0	1	1
WriteOnPaper_true	1	0	0	0	0	17	0	0
Phone_true	6	0	0	0	1	1	10	0
ScratchHead_true	3	0	0	0	0	0	0	12

(b) confusion matrix full set stride 5



(a) confusion matrix loso set 1

(b) confusion matrix loso set 2

(a) confusion matrix loso set 3

(b) confusion matrix Sigerts dataset combined with InceptionTime

Figure 7: Confusion matrices for the data sets in table 3

6 Discussion

6.1 Evaluation of training results

Using the metrics and confusion matrices from section 5.2 one can judge how well a network learns from data and what possible problems there are. That the “other” class is the most miss predicted class regarding false positives and false negatives is due to the imbalance in the classes in the dataset. There are simply more instances of “other” to be learned from, and also more data points of the “other” class that need to be classified. Additionally, there were no concrete instructions for the ”other” class, as it just depicts all movements in between the gestures. This can lead to higher variance in movements for this class. In the loso sets, for different people, different classes were miss predicted more or less severely. This indicates that there is individuality in the gestures from person to person, which the network could not generalize. If for all people, the

same class would have been miss predicted mostly, that would indicate that this class is harder to predict than others. That there are more miss predictions in between the eating gestures is due to the similarity of movement, where every gesture contains the movement of moving food towards the mouth. The absolute miss prediction of the phone class in loso person 3 is due to the different behavior of using the phone with two hands, while most participants only use one hand. This leads to a different angle of the wrist during the whole gesture.
From the confusion matrix in figure 7b) it can be assumed that the network did not learn sufficient patterns from the data. There is no main diagonal which would indicate true positive predictions. This is probably due to the quality of data regarding the sampling stability and the imbalance of the dataset. As the network is using accuracy as an internal metric for training, the imbalance can influence the training behavior negatively. Adding weights to the classes did not seem to circumvent this problem.

6.2 Comparison of different data sets and networks

The best performing network is the CNN on the full dataset with stride 1 5a. With stride 5 5b however it lies below the performance of the SVM 1 and even more so when comparing the SVM to the CNN on Sigerts dataset. Regarding generalizability, both the CNN and the SVM lose percentages on average when doing loso experiments. For the CNN the ratio of recall and precision stays close to the full set, the precision in the SVM drops a lot compared to that of the recall. Therefore, the CNN classifies better onto unseen subjects.
That the CNN with stride 1 performs much better than the other networks can be due to the fact that a time series in the train and test set can share seconds from the original dataset. This means that, for example, in the train set a window including seconds 3-8 of test person 1 can be included while in the test set a window 4-9 of test person 1 could be included in the most extreme case. This cannot be the case for any of the other data sets tested. Alternatively, it could also be due to the difference in data points that the network can train on, which is 5 times higher for the stride 1 approach. The loso sets perform as good as the full set with stride 5. This is an indication that indeed the amount of data points the network has to train on makes a significant difference. In Sigerts experiment, the loso set performed significantly worse than the full sets. These results indicate that a more advanced splitting algorithm which would negate the sharing of information between test and train set but contain more data points in both sets would result in F1 scores somewhere in between the stride1 and stride 5 approach.
Regarding the two data sets, the CNN learned better from the newly gathered data than from Sigerts data. The main differences between the data sets are the number of classes and the distribution of data points per class. Generally, a higher number of classes should rather indicate that the newer dataset has a lower F1 score due to higher probability of miss classification. The amount of data points in the majority class is higher in Sigerts dataset. This makes it more difficult to learn from all classes equally for the classifier. Weights were applied during training, but this did not have any significant effect on the results. Lastly, there is the difference in sampling stability. There were close to no deviations in the new data regarding sampling stability, while for the old data the sampling was rather unsteady with high deviations in between seconds.

It cannot be said which of these is the exact reason for the difference in performance between the two data sets.
Additionally, the evaluation on the multi class problem also results for a binary classification problem of eating vs not eating were calculated. The results for this can be seen in table 4. The networks are still trained on the multi class problem and only the metrics were recalculated, therefore the results can differ if the network would also be trained on the binary problem. From the results one can see that for the stride 1 set the results are almost identical, even 2% lower in F1 score. For the loso sets and the stride 5 sets, the F1 scores improved. In the stride 5 set the F1 score increased from 69% to 80% and in the loso sets on average from 63% to 65%. This supports the thesis that a significant amount of miss predictions happened in between eating gestures. This can be beneficial, especially for the case where the main goal of an application is to be a reminder for the user to log their food, as a classification of the specific eating gesture is not needed for that. The results of the CNN trained on Sigerts dataset show that this network is completely unusable. The F1 score dropped from 27% to 6%. This further supports the result that the network did not pick up patterns for classification from the data, but only fit to the majority class.

6.3 Implications for automatic eating detection

There are two main approaches in eating detection. The first one tries to combine a lot of sensor data, classify it and then uses algorithmic approaches in order to fuse the data into a final classification. This classification can include information like type of food, nutritional values and amount of food. In this paper, the 6 different axis of IMU measurements were fused and classified as one MTS. The majority approach in the literature is to analyze statistical features of every axis independently and then classify them using an SVM. The results of both approaches were comparable in F1 scores. This can be further investigated regarding fusion of multiple types of sensors, like what is done in weather forecasting, for example. This could further enhance the predictions for dietary monitoring.
In the other case, the classifications are used in order to remind the user to take notes. This experiment gave no conclusive idea that using CNN for the classification of eating gestures provides any significant benefit above the already established method of SVMs. Similar, from literature, there was no clear indication that this would be the case. However, there is also no counter indication that CNN can perform on similar levels in such tasks as SVMs do.

7 Conclusions and recommendations

7.1 Conclusions

The goal of the paper is to answer the question 1. Are CNN's a suitable alternative to SVMs in combination with human feature engineering for the classification of eating gestures and 2. How well do they perform in comparison to the SVM approach.

Yes, CNN's can be an alternative to other machine learning approaches like SVMs or random forests. There are indications that for more complex problems they could

be beneficial, while performing worse on unbalanced and limited data. CNN's seem to generalize better to unseen data. This, however, could also be due to the difference in data sets used in comparing them to the SVM. There is no evidence that CNN's, especially the InceptionTime ensemble, are in general a better or worse approach than machine learning approaches like SVMs for the classification of eating gestures.
This paper can be a starting point for the introduction of multivariate time series analysis for sensor data in the work of automatic eating detection.

7.2 Recommendations

To make the comparison of SVM and CNN more accurate, the SVM would need to be reevaluated on the new data and included in the comparison of results.
One could look into other CNN networks than InceptionTime which may be even better at discriminating time dependent data. One possible network for that which was found in the "MTS Backeoff" [11] is the "ROCKET" classifier.
More data needs to be gathered, as deep learning and convolutional networks best learn from big data sets, where they can extract different levels of features. This could include long term trials where people use systems outside controlled environments for extended periods of time.

References

[1] S. Mevissen, “A wearable sensor system for eating event recognition using accelerometer, gyroscope, piezoelectric and lung volume sensors,” Ph.D. dissertation, University of Twente, 2021.

[2] Y. C. Chooi, C. Ding, and F. Magkos, “The epidemiology of obesity,” *Metabolism*, vol. 92, pp. 6–10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002604951830194X

[3] WHO, “Obesity and overweight,” 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[4] G. Schiboni and O. Amft, “Automatic Dietary Monitoring Using Wearable Accessories BT - Seamless Healthcare Monitoring: Advancements in Wearable, Attachable, and Invisible Devices,” T. Tamura and W. Chen, Eds. Cham: Springer International Publishing, 2018, pp. 369–412. [Online]. Available: https://doi.org/10.1007/978-3-319-69362-0{_}13

[5] X. Ye, G. Chen, Y. Gao, H. Wang, and Y. Cao, “Assisting Food Journaling with Automatic Eating Detection,” in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 3255–3262. [Online]. Available: https://doi.org/10.1145/2851581.2892426

[6] E. Thomaz, “Practical Food Journaling,” in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp ’13 Adjunct. New York, NY, USA: Association for Computing Machinery, 2013, pp. 355–360. [Online]. Available: https://doi.org/10.1145/2494091.2501089

[7] W. Hong and W. G. Lee, “Wearable sensors for continuous oral cavity and dietary monitoring toward personalized healthcare and digital medicine.” *The Analyst*, vol. 145, no. 24, pp. 7796–7808, jan 2021.

[8] V. Nunavath, S. Johansen, T. S. Johannessen, L. Jiao, B. H. Hansen, S. Berntsen, and M. Goodwin, “Deep Learning for Classifying Physical Activities from Accelerometer Data.” *Sensors (Basel, Switzerland)*, vol. 21, no. 16, aug 2021.

[9] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, “InceptionTime: Finding AlexNet for time series classification,” *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020. [Online]. Available: https://doi.org/10.1007/s10618-020-00710-y

[10] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, Batista Gustavo, and Hexagon-ML, “The UCR Time Series Classification Archive,” 2018.

[11] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances.” *Data mining and knowledge discovery*, vol. 35, no. 2, pp. 401–449, 2021.

[12] A. Bagnall, H. A. Dau, J. Lines, M. Flynn, J. Large, A. Bostrom, P. Southam, and E. Keogh, “The UEA multivariate time series classification archive, 2018,” 2018. [Online]. Available: https://arxiv.org/abs/1811.00075

[13] “MetaMotionS.” [Online]. Available: https://mbientlab.com/store/metamotions/

[14] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, “The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances,” *Data Mining and Knowledge Discovery*, vol. 35, no. 2, pp. 401–449, 2021. [Online]. Available: https://doi.org/10.1007/s10618-020-00727-3

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

A ConsentForm

Participant:

By signing this consent form, I acknowledge the following:
1. I am sufficiently informed about the investigation by means of a separate information sheet. I have read the information sheet and then had the opportunity to ask questions. These questions have been adequately answered.
2. I voluntarily participate in this study. There is no explicit or implicit compulsion for me to participate in this research. It is clear to me that I can terminate participation in the investigation at any time, without giving any reason. I don't have to answer a question if I don't want to.
3. I am aware that I can withdraw from the study up to 24 hours after the examination without giving a reason.
4. I am aware that the personal information collected about me that could identify me, such as height, weight, gender and age, will not be shared outside the research team.

	Yes	No
5. I consent to process the data collected from me during the investigation as included in the attached information sheet.		
6. I give permission to make video recordings including audio during the experiment.		
7. I am aware of the possible risks as described in the information brochure and have made my objections known.		
8. I give permission that the data of this research as prescribed by research guidelines will be stored for a maximum of ten years in a password-protected location.		
9. I give permission to publish the data and results of this research anonymously (without audio and video) for future research		

Signed in duplicate:

Name researcher Signature

……………………………………. …………………………………….

I have given an explanation of the research. I declare myself willing to answer to the best of my ability any questions that arise about the research.

Participant Name Signature

……………………………………. …………………………………….

Date:

For additional information you can mail to Sigert Mevissen (s.j.mevissen-1@utwente.nl) or Sönke van Loh (s.u.vanloh@student.utwente.nl). For complaints or independent or additional comments for an independent body, the participant may report this to the Ethics Committee of EEMCS (ethicscommittee-cis@utwente.nl).

B Information Brochure

Research

Obesity is an increasingly common problem in the world. There are many ways to do something about this, where one method works better than the other. An important aspect in losing weight is regulating the food intake, such as the amount of food, the type of food and the regularity of the food. There are apps to keep track of what you eat to find out what you ingest in a day. The problem with this, however, is that this does not always happen properly; people enter too much or too little or forget to add meals and snacks. You want to know objectively what, when and how much food someone gets. This research ties in with that, specifically with the issue of when someone is eating. This could be used, for example, to remind people to register what has just been eaten.
You will wear a sensor during the experiment that can potentially be used to show when the wearer is eating.

1) The sensor is an inertial measurement unit, which measures the rotational speed and acceleration in 3 axes. Movements towards the mouth can be used as an eating indication.

Explanation of the Experiment:

Task	#number of times/ duration	Additional information
Clap	3 times	-
Random Arm movement	30 sec	For example imagine that you collect dishes from the table
Eat chips out of bowl	8 times	-
Eat yoghurt with spoon	8 times	-
Eat fruit with fork	8 times	-
Write on paper	30 sec	Copy for set duration from a book
Use the phone	30 sec	Either scroll through social media or note down something
Scratch head/ adjust hair	8 times	The main goal of this task is to move your hand towards your head without it beeing an eating gesture
Lay hands on table	-	This is the end of the experiment. The researcher will indicate when the experiment is finished and the hand can be freely moved afterwards

During the experiment, you will eat or drink the following:

Type of Food	Possible implications
chips	allergies
Yoghurt/vegan yoghurt	Lactose intollerance, allergies
Fruit (banana or apple)	Glucose intollerance, allergies

During the experiment, a video (with audio) will be made that will be annotated. This is necessary to be able to link the arm movements in time to the data from the sensors.
The examination takes about 10 minutes.

Risks

The subject is at risk of an allergic reaction to the food or drink to be consumed. If you have any objection to eating or drinking one of the mentioned foods due to, for example, an allergy or other dietary restrictions, you can indicate this. The packaging of the food/drink to be consumed will be stored so that you can check what is in it. Indicating that you cannot or do not want to consume one or more things can be done without giving a reason. If you have any questions about this, please ask them. The research has been reviewed and approved by an ethics committee.

Personal data

All results collected will be used solely for research purposes, will never be shared with third parties and will be anonymous if published. Personally identifiable information will be retained for up to ten years, as usual research guidelines prescribe. It concerns the video images with audio and the name on the consent form. The personal data is stored securely until it is no longer usable for the research. The data will only be accessible to the researchers working on the project.

Participation

Participation in this study is completely voluntary and the participant can stop at any time, without the need for a reason and without consequences. After the experiment, the participant can also withdraw within 24 hours after the experiment.

Contact

If a question or need for additional information arises during or after the research, the participant can contact Sigert Mevissen (s.j.mevissen-1@utwente.nl) or Sönke van Loh (s.u.vanloh@student.utwente.nl)

If the participant wishes to submit a complaint or has additional comments for an independent body, the participant may report this to the Ethics Committee of EEMCS (ethicscommittee-cis@utwente.nl).

C Reflection on the scientific and societal dimensions of the work done in this thesis

Obesity and the eating habits which cause obesity are an ever growing societal problem. Therefore, it is strictly necessary that a multitude of ways is researched which can help people tackle those problems. Since sensors like IMUs, cameras and so on become more and more integrated in our everyday life it is beneficial to use those if applicable. In order to make use of the data these sensors produce their data needs to be analyzed, mostly in the case of eating detection classified. This helps subjects by reminding them of tracking their food intake or doing the tracking for them. Deep neural networks and convolutional neural networks are introduced into alot of problems which involve finding patterns in data, be it pictures, timeseries or audio. The biggest thresholds in the field are computing power and amounts of data and both seem to become less problematic. Both, an increase in sensors and therefore data and an increase in computing power is observable. This is why it is important to test problems for their applicability to deep learning approaches. This thesis tries to identify if deep learning is applicable to the detection of eating gestures from IMU timeseries data. What is important in order for it to work, for example regarding the data that one needs and how does it perform compared to the common approach used so far in literature. The results were not astonishing and did not exceed current results from literature. However they did show that the approach of deep learning is applicable and can be pursued in this field.