Cognitive Science Student Journal
Osnabrück University

Diagnosing Pancreatic Cancer via Urinary Biomarkers

Malte Heyen

Cognitive Science Student Journal 2023, Number 11

Heyen, M. (2023). Diagnosing Pancreatic Cancer via Urinary Biomarkers. *Cognitive Science Student Journal 2023, 11.* 1-9.

This title can be downloaded at:
http://cogsci-journal.uni-osnabrueck.de

Institut für Kognitionswissenschaft
Universität Osnabrück
49069 Osnabrück
Germany
https://www.ikw.uni-osnabrueck.de

Storage and cataloging done by Osnabrück University

The project is funded by *Stiftung Innovation in der Hochschullehre*

Abstract

This project is based on the "Urinary biomarkers for pancreatic cancer" dataset published in a case–control study by Debernardi et al. (2020). The authors analyzed urinary biomarkers from three groups of patients: healthy controls, patients with non-cancerous pancreatic conditions (e.g. chronic pancreatitis), and patients with pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer. The main goals of this project are to investigate if the levels of the urinary biomarkers creatinine, LYVE1, REG1B, and TFF1 differ significantly between the patient groups and to develop a statistical model to identify patients with pancreatic cancer based on the four urinary biomarkers. The analysis of multiple gamma regression models showed that the level of LYVE1, REG1B, and TFF1 is significantly elevated in patients who have been diagnosed with PDAC. A multinomial logistic regression model trained on 50% of the data was able to detect 84.7% of the patients with PDAC in the test set. The levels of LYVE1 and TFF1 seem to be the most relevant predictors of the risk of pancreatic cancer.

1 Introduction

This project is based on the "Urinary biomarkers for pancreatic cancer" dataset provided by Davis (2020) on Kaggle[1]. The dataset was published in a case–control study by Debernardi et al. (2020) in the journal PLOS Medicine. The authors analyzed urinary biomarkers from three groups of patients: healthy controls, patients with non-cancerous pancreatic conditions (e.g. chronic pancreatitis), and patients with pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer. Pancreatic cancer is one of the most lethal types of cancer: After diagnosis, the five-year survival rate is below 10% and the majority of cases show unresectable, already advanced or metastatic tumors at the time of diagnosis (Debernardi et al., 2020; Sarantis et al., 2020). Unfortunately, most patients are asymptomatic until a late stage (Kamisawa et al., 2016) and classical treatments like chemotherapy, surgery, and radiation do not seem to significantly enhance survival rate (Sarantis et al., 2020). However, if detected early, the chance of survival is highly improved (Sarantis et al., 2020). Thus, a reliable test to diagnose patients early in the course of the disease may have a significant impact on treatment and survival rate. The goal of this project is to develop a classifier to identify patients with pancreatic cancer based on urinary biomarkers.

2 Material and methods

The following section introduces the dataset and the research questions of interest. Further, the methods used to investigate the research questions are presented.

2.1 Dataset

The dataset contains 590 patient samples from three groups: healthy control, patients with non-cancerous pancreatic conditions (benign hepatobiliary disease), and patients with pancreatic ductal adenocarcinoma (PDAC) (Davis, 2020). According to Debernardi et al. (2020), groups were age- and sex-matched wherever possible (see Table 1). PDAC samples were collected from patients before treatment. The original dataset contains the following 14 variables:

1. sample_id: a unique string identifying each subject.
2. patient_cohort: the cohort the sample was collected from. Cohort 1 are the previously used samples.

[1] see https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

3. sample_origin: the centre where the sample was obtained.[2]

4. age: patient age in years

5. sex: patient sex (F=Female, M=Male)

6. diagnosis: patient group (1=healthy control, 2=benign, 3=PDAC)

7. stage: disease stage, only for patients with PDAC (diagnosis 3)

8. benign_sample_diagnosis: diagnosis, only for benign patients

9. plasma_CA19_9: blood plasma levels of CA 19–9 monoclonal antibody in U/ml that are often elevated in patients with pancreatic cancer. These were only assessed in 350 patients.[3]

10. creatinine: urinary biomarker of kidney function in mg/ml

11. LYVE1: urinary levels of lymphatic vessel endothelial hyaluronan receptor 1 in ng/ml, a protein that may play a role in tumor metastasis

12. REG1B: urinary levels of a protein in ng/ml that may be associated with pancreas regeneration

13. TFF1: urinary levels of Trefoil Factor 1 in ng/ml, which may be related to regeneration and repair of the urinary tract

14. REG1A: urinary levels of a protein in ng/ml that may be associated with pancreas regeneration. These were only assessed in 306 patients.[4]

From the 208 benign samples, 119 are chronic pancreatitis and from the 199 PDAC samples, 102 are stage I–II and 97 are stage III–IV. The goal of this study is to examine whether urinary biomarkers can be used to detect and predict pancreatic cancer. Thus, the main features which will be included in the analysis are the four urinary biomarkers: creatinine, LYVE1, REG1B, and TFF1. Additionally, age and sex may also impact the risk for pancreatic cancer and therefore their effect is also included in the analysis. I will exclude the biomarkers plasma_CA19_9 and REG1A for this work as they were only partly collected and are not the main focus of this project. Similarly, the variables patient_cohort, sample_origin, stage, and benign_sample_diagnosis are not relevant for this project and are therefore excluded.

	Healthy	Benign	PDAC
Patients	183	208	199
Avg. Age	56.3	54.7	66.1
Female	62.8%	48.5%	41.7%

Table 1: Number of patients, average age, and female patients per group in the dataset.

[2]The samples stem from the following centres: Barts Pancreas Tissue Bank, University College London, University of Liverpool, Spanish National Cancer Research Center, Cambridge University Hospital, and University of Belgrade (Debernardi et al., 2020).

[3]One goal of the original study by Debernardi et al. (2020) was to compare various CA 19-9 cutpoints from a blood sample to the model developed using urinary samples.

[4]Another goal of the original study by Debernardi et al. (2020) was to examine whether REG1B is a better predictor than REG1A.

2.2 Research questions

This project is twofold and will examine the following two questions:

1. Do the four biomarkers of interest differ between the three patient groups?
2. Can the four biomarkers be used as predictors in a classification model to accurately predict the risk of pancreatic cancer in patients?

With regard to the first question, I want to investigate if the distributions of each biomarker variable differ based on the patient group. As a first approach to tackle the first question, performing a two-way MANOVA with a 2x3 factorial design, including the factors diagnosis and sex was considered. However, MANOVA requires that the dependent variables (the four biomarkers) are multivariate normally distributed for each group and further assumes homoscedasticity, i.e. that the covariance matrices for each group are equal. Testing normality with the Shapiro test and homoscedasticity with Bartlett's test for all four biomarker variables revealed that these assumptions are in fact not met ($p << 0.002$ for all tests). Visual inspection of density plots (see Figure 1) and QQplots support these findings. A Generalized Linear Model based on the gamma distribution is more appropriate to capture the distributions of the four biomarkers and thus used to investigate Question 1. In order to test a classification model for risk prediction to investigate Question 2, I will use a multinomial logistic regression model.

2.3 Gamma regression

The gamma regression is used for continuous, skewed response variables and when the variance of the response grows with its mean. The gamma regression model with log-link is defined by the equation:

$$Y_i \sim Ga(v, \tfrac{\mu_i}{v}) \text{ where } log(\mu_i) = x_i'\beta \ (i = 1, 2, 3)$$

The mean μ_i is connected to the linear predictors $x_i'\beta$ via the link function. In order to investigate Question 1, I will apply four gamma regressions for each biomarker of interest individually with diagnosis and sex as independent variables.

2.4 Multinomial logistic regression

With respect to Question 2, I will define a classification model which incorporates the variables creatinine, LYVE1, REG1B, TFF1 and age as predictors to discriminate the three classes healthy, benign, and PDAC. The following section is adapted from the book "Applied Multivariate Statistics with R", chapter 10.2, by Zelterman (2015). Generalizing binary logistic regression to multinomial logistic regression, which is able to discriminate multiple classes, is based on the idea that one class is selected as a reference baseline. All other classes are then compared to this baseline. For this project, the healthy control group is a natural candidate for a baseline. The multinomial logistic regression model can then be defined by the following equations, where x is the vector containing the explanatory (predictor) variables and β contains the model coefficients:

$$\log \frac{Pr(Class = 2|x)}{Pr(Class = 1|x)} = \beta_2' x \tag{1}$$

$$\log \frac{Pr(Class = 3|x)}{Pr(Class = 1|x)} = \beta_3' x \tag{2}$$

Now, the probability to belong to class 2 or class 3 can separately be compared to the probability of belonging to the reference category. The coefficients β_2 and β_3 of the fitted model then represent the change in the log-odds ratios between the probabilities of belonging to class 2 (or class 3, respectively) or the reference class 1, when the independent variable changes by one unit (Zelterman, 2015).

3 Results

The following section presents the results of the statistical analysis performed on the dataset. As a first exploratory analysis, I applied Kernel Density Estimation (KDE). To test if there is a significant difference in the level of urinary biomarkers between the patient groups, I used gamma regression. Lastly, multinomial logistic regression was used to develop a classification model for predicting a patient's diagnosis.

3.1 KDE for each biomarker per group

As a first exploratory analysis, I applied Kernel Density Estimation (KDE) to estimate the distribution for each biomarker and visualized the density per group. Visual inspection of Figure 1 indicates that for LYVE1, REG1B, and TFF1, the patient groups seem to have different distributions. For the PDAC group, the biomarker levels for LYVE1, REG1B, and TFF1 are elevated. For the benign group, it seems that at least the levels of LYVE1 and TFF1 are moderately higher than in the healthy control. For creatinine, the groups appear to have similar distributions.

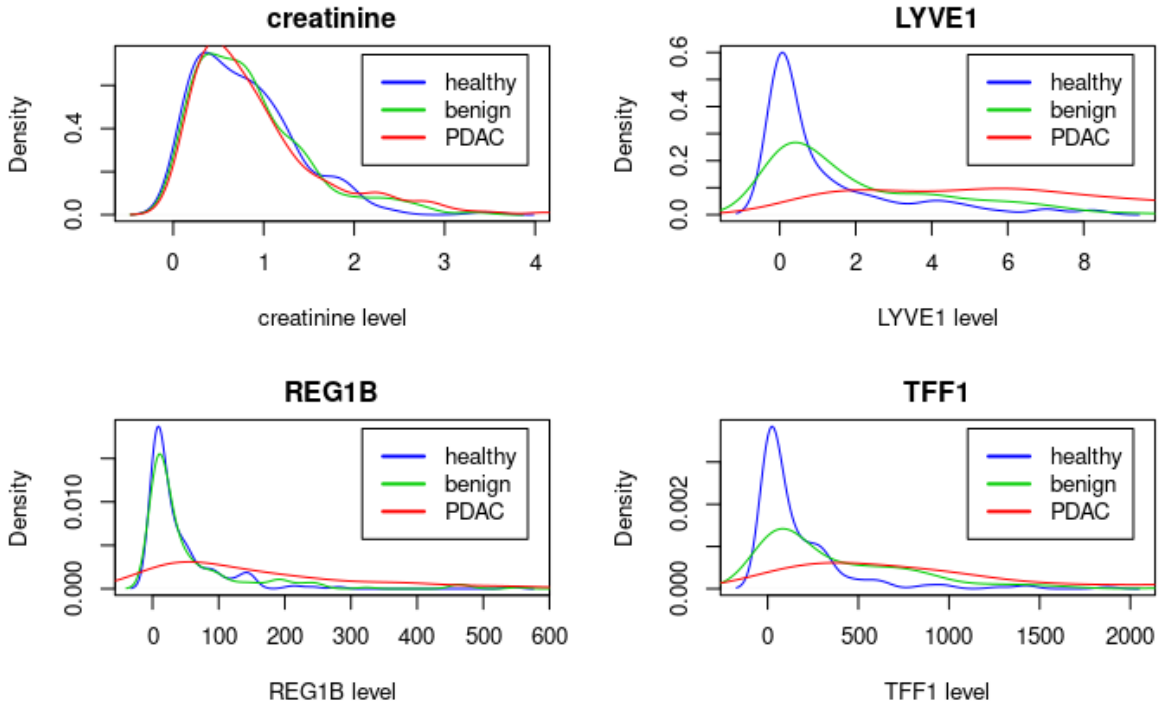


Figure 1: Distributions per patient group based on KDE for the biomarkers creatinine, LYVE1, REG1B, and TFF1.

3.2 Results gamma regression

When looking at the interactions between the dependent variables diagnosis and sex, as shown in Figure 2, no interactions can be identified between them. This was also shown when fitting the four gamma regression models, as the interaction coefficient between the dependent variables does not have a significant effect on the biomarker levels. Thus, the interaction effect was excluded in the four model definitions: $biomarker\_level \sim diagnosis + sex$. A Pearson X^2 test on the dispersion parameters after fitting showed no lack of fit (all models $p > 0.4$).

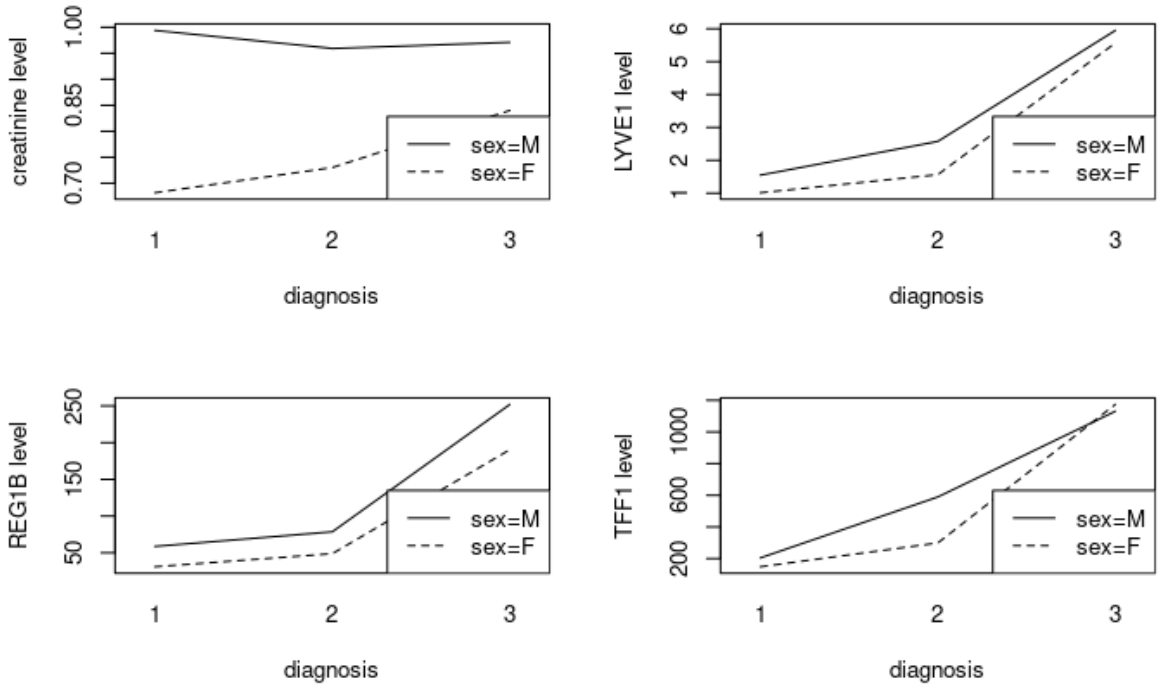


Figure 2: Effect of diagnosis and sex on the mean level of the biomarkers creatinine, LYVE1, REG1B, and TFF1. There seems to be no interaction present between diagnosis and sex.

As shown in Table 2, three out of the four urinary biomarkers of interest (namely LYVE1, REG1B, and TFF1) significantly differ between patient groups, even when considering sex as an additional explanatory factor. Their level is enhanced in patients who have been diagnosed with PDAC. The levels of LYVE1 and TFF1 also seem to be significantly elevated in patients with benign pancreatic diseases, even though not as much as in patients with PDAC. The level of REG1B seems to be slightly elevated in benign patients as well, however not strongly significantly. The level of creatinine does not seem to be significantly different between patient groups, however, male patients seem to have a slightly higher level of creatinine than female patients. In general, all urinary biomarkers are significantly higher in male patients than in female patients.

3.3 Results multinomial logistic regression

Before training the model, the four features were normalized. In order to evaluate the performance of the model, the data was split into 50% training set and 50% evaluation set. After fitting the model on the training set, the confusion matrix and additional

	1	2	3
1	60	25	4
2	40	50	9
3	18	17	72

Table 3: Confusion matrix for the predicted and true classes.

	diagnosis=2	diagnosis=3	sex=M
creatinine	0.03012	0.09812	0.26082***
LYVE1	0.48819***	1.52193***	0.32884**
REG1B	0.3923*	1.6436***	0.4588***
TFF1	0.9003***	1.8758***	0.3257**

Table 2: Coefficients for the factors diagnosis and sex for each gamma regression model. The four gamma regression models are for creatinine, LYVE1, REG1B, and TFF1.
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

performance metrics were calculated based on the evaluation set. The overall accuracy is 0.6169 ($\pm$0.0581). Based on the confusion matrix (Table 3) and the performance measures per patient group (Table 4), it is evident that the performance of the model actually differs between groups. The model is better at discriminating class 3 from the other two classes than discriminating between classes 1 and 2. The Sensitivity (true positive rate) and Specificity (true negative rate) are higher or equal for class 3 than for the other two patient groups.

	Class: 1	Class: 2	Class: 3
Sensitivity	0.5085	0.5435	0.8471
Specificity	0.8362	0.7586	0.8333
Pos Pred Value	0.6742	0.5051	0.6729
Neg Pred Value	0.7184	0.7857	0.9309
Prevalence	0.4000	0.3119	0.2881
Detection Rate	0.2034	0.1695	0.2441
Detection Prevalence	0.3017	0.3356	0.3627
Balanced Accuracy	0.6723	0.6510	0.8402

Table 4: Performance statistics for all three patient groups. Class 1 is the healthy control group, Class 2 is the benign group, and Class 3 is the PDAC group.

This is also evident in Figure 3, which shows the predicted posterior probabilities for each patient sample of belonging to group 2 or 3. Ideally, every class would be distributed in one corner of the triangle. The overlapping distribution of class 1 and class 2 patients around the middle of the x-axis also shows that the model has difficulties distinguishing between patient group 1 and group 2.

Analysing the (exponential) model coefficients in Table 5 shows that especially LYVE1 and TFF1 seem to be good predictors for pancreatic cancer. An interpretation of the exponential coefficient estimates suggests that an increase of one standard deviation in the level of TFF1 increases the risk of PDAC by 10 compared to the healthy control. An increase by one standard deviation in the level of LYVE1 increases the risk of PDAC by 5 compared to the healthy control. However, TFF1 is also a good predictor for non-cancerous (benign) pancreatic diseases and may therefore not be a good indicator to discriminate between patients with benign pancreatic diseases and patients with PDAC. To discriminate between benign and PDAC patients, the LYVE1 level may be more critical.

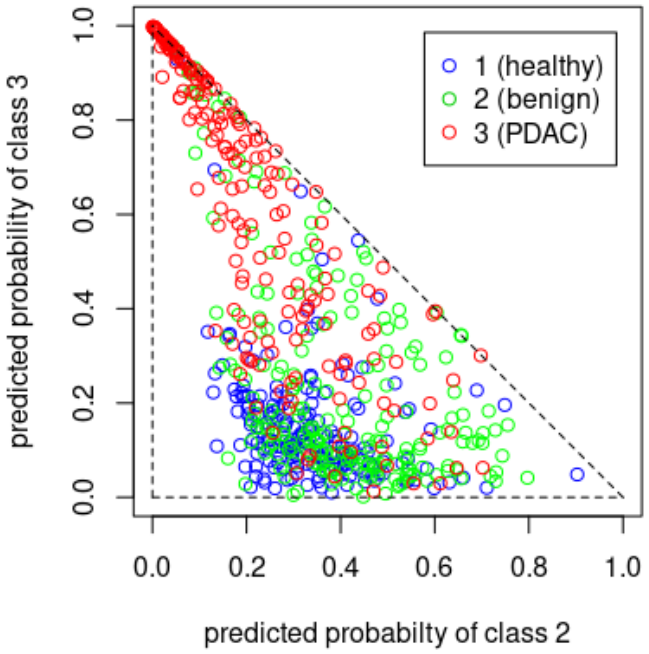


Figure 3: Predicted probabilities for all patients to belong to class 2 or class 3.

	(Intercept)	age	creatinine	LYVE1	REG1B	TFF1
Class 2 (benign)	11.7805140	0.9756719	0.5376012	1.746908	1.060096	10.95982
Class 3 (PDAC)	0.3209646	1.0277276	0.4153227	5.077497	1.757681	10.10292

Table 5: Exponential coefficients of the fitted multinomial regression model.

4 Conclusion

Based on KDE and the analysis of multiple gamma regression models, it was established that three out of the four urinary biomarkers of interest, namely LYVE1, REG1B, and TFF1, do actually differ between patient groups, even when considering sex as an additional explanatory factor. Their level is enhanced in patients who have been diagnosed with PDAC. The levels of LYVE1 and TFF1 also seem to be significantly elevated in patients with benign pancreatic diseases, even though not as much as in patients with PDAC. The level of REG1B is also elevated in the PDAC group. The level of creatinine does not seem to be significantly different between patient groups, however, male patients seem to have a slightly higher level of creatinine than female patients.

The multinomial logistic regression model performed better at discriminating between PDAC patients and the two other groups than discriminating between healthy and benign patients (see Figure 3). Prior to the analysis, I expected that benign and PDAC patients would be more difficult to discriminate. The classifier was able to detect 84.7% of the patients with PDAC in the test set. Analyzing the model coefficients shows that especially LYVE1 and TFF1 seem to be good predictors, resulting in a 5-fold and 10-fold risk increase of PDAC respectively per increase of one standard deviation compared to the healthy control. TFF1 additionally seems to be a good predictor for non-cancerous (benign) pancreatic diseases. To discriminate between benign and PDAC patients, the LYVE1 level is more critical.

References

Davis, J. (2020). *Urinary biomarkers for pancreatic cancer*. Kaggle. Retrieved December 31, 2020, from https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer

Debernardi, S., O'Brien, H., Algahmdi, A. S., Malats, N., Stewart, G. D., Plješa-Ercegovac, M., Costello, E., Greenhalf, W., Saad, A., Roberts, R., et al. (2020). A combination of urinary biomarker panel and pancrisk score for earlier detection of pancreatic cancer: A case-control study. *PLoS Medicine*, *17*(12). https://doi.org/10.1371/journal.pmed.1003489

Kamisawa, T., Wood, L. D., Itoi, T., & Takaori, K. (2016). Pancreatic cancer. *The Lancet*, *388*(10039), 73–85. https://doi.org/10.1016/S0140-6736(16)00141-0

Sarantis, P., Koustas, E., Papadimitropoulou, A., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Pancreatic ductal adenocarcinoma: Treatment hurdles, tumor microenvironment and immunotherapy. *World journal of gastrointestinal oncology*, *12*(2), 173.

Zelterman, D. (2015). *Applied multivariate statistics with r*. Springer. https://doi.org/10.1007/978-3-319-14093-3

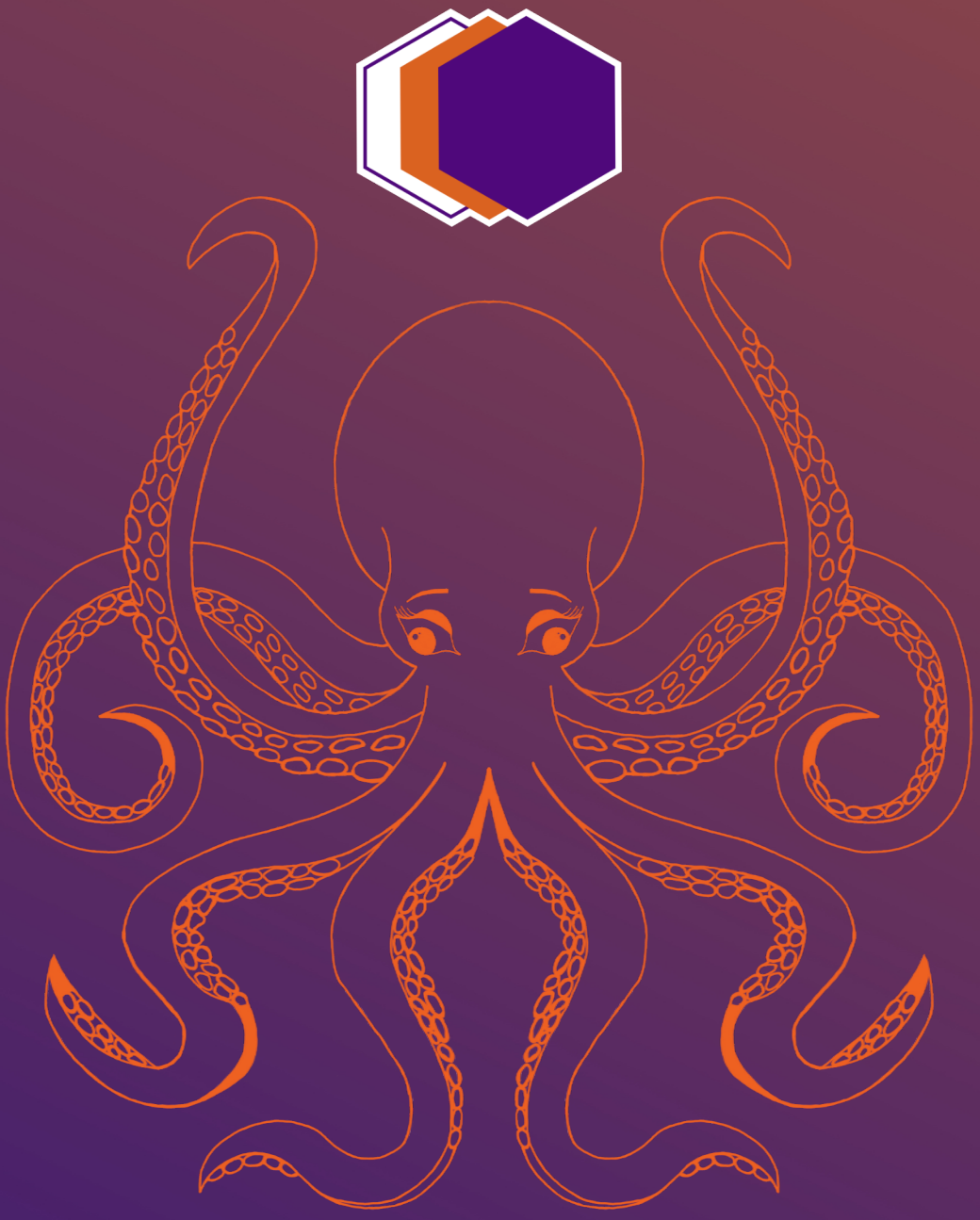
About the Journal

The 'Cognitive Science Student Journal' aims at giving its readers an insight into current research and cutting-edge topics at our institute from a student perspective as well as students a platform to publish their work. Its editorial board consists of seminar participants and instructors of the Institute of Cognitive Science.

Cognitive Science is taught as an interdisciplinary research field at University Osnabrück, investigating cognition and the mind as a joint research effort of Artificial Intelligence, Neuroscience, Computational Linguistics, Psychology, Neuroinformatics, and Philosophy of Mind.

The journal can be accessed via:
http://cogsci-journal.uni-osnabrueck.de

Find us on social media:
https://www.instagram.com/cogscistudentjournal/
https://www.linkedin.com/company/cognitive-science-student-journal/



Cognitive Science Student Journal 2023, Number 11