Explainable Depression Detection Using Handwriting Features

Francesco Prinzi*[1,2], Gennaro Raimo[3], Pietro Barbiero[4,2], Gennaro Cordasco[3,5], Pietro Lio[2], Salvatore Vitabile[1], and Anna Esposito[3,5]

[1] Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Palermo, Italy
{francesco.prinzi, salvatore.vitabile}@unipa.it
[2] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
{pl219}@cam.ac.uk
[3] Department of Psychology, University of Campania "L. Vanvitelli", Caserta, Italy
{gennaro.raimo, gennaro.cordasco, anna.esposito}@unicampania.it
[4] Università della Svizzera Italiana
{barbip}@usi.ch
[5] Istituto Internazionale per gli Alti Studi Scientifici (IIASS), Vietri sul Mare, Italy

Abstract. Depression is considered one of the most prevalent diseases worldwide, with a rapid increase in recent years. An interesting area of research for depression detection is the analysis of handwriting and drawing. Although machine learning models have shown promising results to support the diagnostic process in several fields, their lack of transparency inhibits their actual use. For this reason, the work aims to develop explainable machine learning models for depression detection using handwriting-based features. The research involved 138 participants, equally divided into healthy and sub-clinical groups, according to their score on a DASS-21 (Depression, Anxiety and Stress Scale). The same protocol, consisting of seven handwriting and drawing activities, was submitted to each participant. Decision Tree and XGBoost algorithms were compared with the explainable-by-design Entropy-based Logic Explained Network (e-LEN). A 10-repeated 10-fold cross-validation was employed for performance evaluation. XGBoost showed a higher AUROC 0.750 ± 0.134 compared with e-LEN 0.723 ± 0.143 and DT 0.681 ± 0.119. However, e-LEN enabled a significant reduction in complexity compared with XGB. In particular, the e-LEN model exploits on average 41 logic predicates to perform the predictions, while XGBoost employs 303 nodes on average. Moreover, the e-LEN model enabled an explanation by providing the logic rules for clinician model validation. Specifically, ductus, time and pressure features were more predictive. It is therefore possible to use these techniques and methodologies to speed up and improve the identification of depression.

Keywords: Depression Detection · Explainable AI · Machine Learning · Handwriting Features.

* Corresponding author. Email: francesco.prinzi@unipa.it

1 Introduction

With a sharp rise in the past two decades, depression has emerged as the most common mental disorder globally in recent times. Depression, also called clinical depression, major depressive disorder, or major depression, affects about 5% of adults. As people age, the prevalence of depression tends to increase, impacting about 5.7% of adults aged 60 and older. Approximately 350 million people suffer from depression worldwide [8]. Numerous circumstances, both positive (like pregnancy) and negative (like traumatic occurrences) can be linked to this syndrome. Women are almost 50% more likely than men to experience depression. More than ten percent of expectant and new mothers experience depression [23]. More than 75% of people in low- and middle-income countries do not receive mental health therapy, although there are effective treatments available [21]. The societal stigma associated with mental illnesses, a lack of qualified healthcare workers, and inadequate funding for mental health services all hinder the efficacy and the use of therapies. Worldwide, almost 800,000 suicides occur each year as a result of this lack of access to treatment [14]. Cognitive, physical, emotional, and behavioral symptoms are related to depression, as listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [9].

To reduce the costs associated with this condition, efforts have been made to deploy new technology for early diagnosis and detection based on these last symptoms [11]. A particular field of study has examined the behavioral signs of depression and the examination of handwriting and drawings [10]. Numerous traits of individuals, such as markers of neurodegenerative diseases [22], depressive states [19], and negative moods [6], have been successfully identified via the use of statistical methodologies. However, nowadays, a statistical analysis can be approached through machine learning algorithms.

Although machine learning methods have been shown to support the diagnostic process in several fields, their lack of transparency inhibits the actual use [12]. In fact, in some clinical settings, global and local explanations are a key requirement to validate and justify the models and their decisions [15]. Several post-hoc explanation methods were proposed to explain shallow and deep learning architectures. These methods aim to estimate the model features' importance (global explanation) and analyze the contribution of each feature for each specific prediction (local explanation) [16]. Despite these methods are widely used, they do not provide how the involved features interact to perform the prediction. The explainable-by-design Entropy-based Logic Explained Network (e-LEN) [1] was offered as a solution to this problem, providing an explanation through the First-Order Logic formalism. In fact, in order to facilitate the creation of straightforward *logic explanations*, e-LEN integrates constraints in both the architecture and the learning process and computes the First-Order Logic predicates involved in the prediction process to yield the explanations.

For this reason, in this work, the e-LEN architecture was compared with XGBoost and Decision Tree models for depression detection using handwriting features. Considering model explainability as important as their accuracy, the work aims to establish the most important features for depression detection, as

well as provide the predicated leveraged in the prediction process. The predicates computed by e-LEN enable the clinical validation of the model findings.

2 Materials and Methods

2.1 Dataset Description

One hundred thirty-eight participants were included in the study, and they were split into two groups based on the DASS-21 score depression subscale: healthy and sub-clinical [4]. All participants completed the same protocol, consisting of 7 handwriting and drawing tasks. The performed tasks are shown in Figure 1 and the procedure is widely described in [19]. As in previous studies, the pen strokes were separated into three categories [7, 5]: on-paper, in-air (but near the paper), and idle (far from the paper). In particular, 17 features were extracted from the seven tasks and divided into five categories:

- *Pressure*: the values of the pen's pressure on the paper when doing a particular activity (considering just on-paper traits). Features are maximum *(max#)*, minimum *(min#)*, standard deviation *(stdev#)*, average *(avg#)*, 90th percentile *(p90#)*, and 10th percentile *(p10#)* values of pressure.
- *Time*: total time spent to complete the task in each pen status. Features are the time of on-paper *(tDown#)*, in-air *(tUp#)*, and not recognized *(tIdle#)* pen status, and their sum *(tToal#)*.
- *Ductus*: the number of strokes in each pen status. Features are on-paper *(nbDown#)*, in-air *(nbUp#)*, and not recognized *(nbIdle#)* amount of traits.
- *Inclination*: the average inclination of the diagonals of the bounding boxes containing the strokes *(SlopeA#)*.
- *Space*: the area that the strokes task covers (taking into account only characteristics found on paper). Features are a value obtained by deriving the area of the rectangle formed by the distance of the tract from the axes (BB#); the mean lengths of empty spaces among consecutive strokes *(spaceA#)*; sum of the lengths of empty spaces between consecutive on-paper strokes *(spaceT#)*.

The symbol $'\#'$ represents the task number. A total of 121 features were collected including features related to age and gender.

2.2 Machine Learning Methods

Features preprocessing and selection Features were normalized before all the processing steps. The correlated features using the Spearman Correlation coefficient were discarded, considering $|\rho| < 0.9$ as threshold. The two groups' distributions were compared using the Mann-Whitney test. Finally, the Sequential Feature Selector (SFS)[20] was employed for feature selection. The accuracy was the metric to maximize.

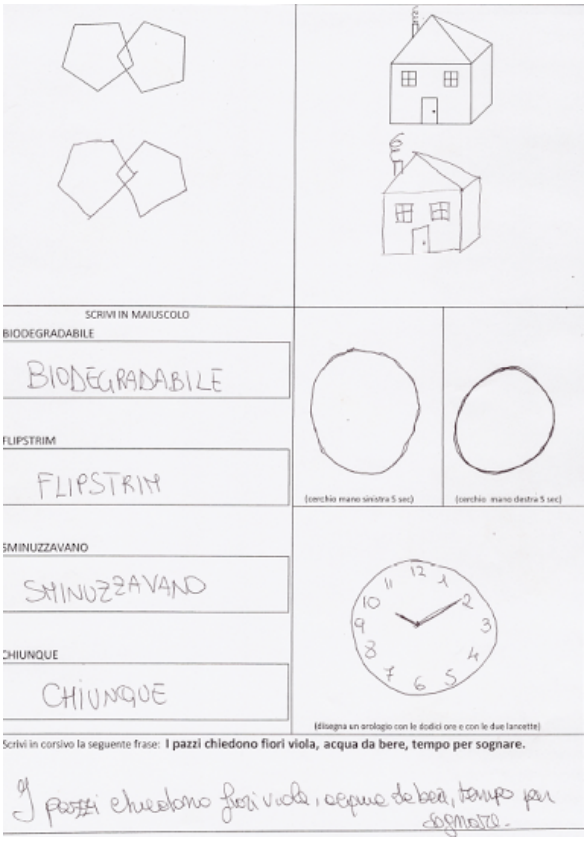

Fig. 1. Performed handwriting and drawing tasks.

Entropy-based Logic Explained Network The main purpose of the work is to provide an explainable model capable of enabling model validation by comparing the findings with the clinical literature [17]. This can be achieved by implementing two main aspects: 1) using an intelligible input, i.e., the intrinsic meaning of each feature is known; 2) using a classifier that is explainable-by-design, or applying post-hoc explanation methods to explain black-box algorithms [18].

Regarding the first aspect, our input is represented by handcrafted intelligible features. Regarding the classifier, the implemented e-LEN [1] was proposed as an explainable-by-design algorithm to provide both high performance and an explainable neural network. The e-LEN model uses the First-Order Logic formalism to determine which concepts in the prediction process are most relevant. The Entropy layer was implemented to compute: (i) a truth table T_i describing how concepts are used by the network to perform predictions for the i-th class; and (ii) the embedding h_i (like any other linear layer). Because of this, the model loss function considers both the maximization of the concept entropy and the

minimization of a standard loss function. Specifically, the formal definition of entropy is:

$$\mathcal{H}(\alpha^i) = -\sum_{j=1}^{k} \alpha_j^i \log \alpha_j^i \tag{1}$$

where the significance of each concept j and class i is denoted by a_j^i. Therefore, the loss function is defined as:

$$\mathcal{L}(f, y, a_1, ..., a_r) = L(f, y) + \lambda \sum_{i=1}^{r} \mathcal{H}(\alpha^i) \tag{2}$$

where $L(f, y)$ is the loss function for supervised learning (f is the network and y is the target), a^i represents the importance of the concepts related to the i class (for all r categories), and the hyperparameter λ is used to adjust the relative importance of low-entropy solutions in the loss function.

Eventually, the e-LEN model provides the global explanation and local explanation. The global explanation consists of the most common predicates found for each class. For this reason, two different sets of predicates are provided for healthy and depressed patients, respectively. Each predicate is connected to the other between an $\wedge$ (and) operator, composing a rule. All rules are connected by the $\vee$ (or) operator. In the end, the overall explanation for each class is represented by the set of all predicates.

Models Training The XGBoost methods and Decision Tree classifiers were compared with the e-LEN model. Three linear layers composed of 200, 100, and 40 neurons each were used to implement the e-LEN, preceding one entropy layer with 200 neurons. Following each layer, the ReLU was employed as the activation function. Ten iterations of a 10-fold cross-validation were conducted to evaluate the models' performance. Accuracy, Specificity, Sensitivity, and Area Under the Receiver Operating Characteristic (AUROC) were computed to evaluate the models' performance.

In addition, the model complexity was calculated during the cross-validation process. The complexity of e-LEN was determined by summing the calculated predicates (separately for each class). The complexity of the Decision Tree model was determined by the number of nodes created during training. In XGBoost, the number of nodes produced by the trees added to the ensemble was considered.

3 Results

3.1 Experimental Results

Fifty estimators, a 0.3 learning rate, and a binary logistic as the loss function were used to train the XGBoost algorithm. Using the Adam optimizer with a learning rate of $0,001$ and a Binary Cross-Entropy with Logit as the loss function, the e-LEN model was trained over 200 iterations.

The performance of the three models computed taking into account the stratified 10-fold cross-validation repeated ten times is displayed in Table 1. XGBoost showed a higher AUROC 0.733 ± 0.112 compared with e-LEN 0.723 ± 0.148 and DT 0.681 ± 0.118. Figure 2 shows the AUROC curves. Furthermore, XGB was equally capable of predicting depression positivity and negativity considering its sensitivity and specificity balancing.

Table 1. Performance achieved by e-LEN, Decision Tree (DT), and XGBoost (XGB) models.

Model	Accuracy	Specificity	Sensitivity	AUROC
e-LEN	0.695 ± 0.116	0.730 ± 0.167	0.661 ± 0.175	0.723 ± 0.148
DT	0.681 ± 0.120	0.693 ± 0.192	0.668 ± 0.168	0.681 ± 0.120
XGB	0.733 ± 0.112	0.733 ± 0.165	0.733 ± 0.150	0.750 ± 0.134

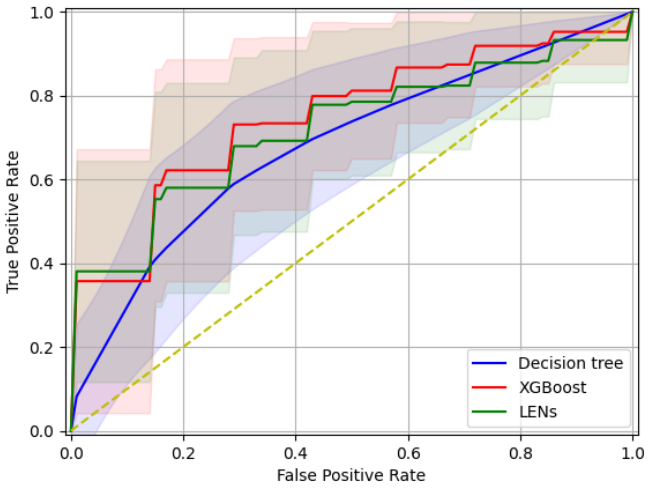

Fig. 2. AUROC curves computed for e-LEN, Decision Tree (DT) and XGBoost (XGB).

Regarding the impact of the employed features, those related to the time were the most selected via SFS for the three trained models. The number of idle strokes features resulted important as well. The common features subset selected for the three models was composed of:

– *Time*: total amount of time used to finish the task (tDown3, tDown5, tIdle6, and tUp3) in each pen status;

– *Ductus*: the total number of traits (nbIdle1, nbIdle7) that are not recognized;
– *Pressure*: the pressure a pen applies to paper (on-paper traits) while a certain task is performed (stdev1).

3.2 Explainability and Complexity

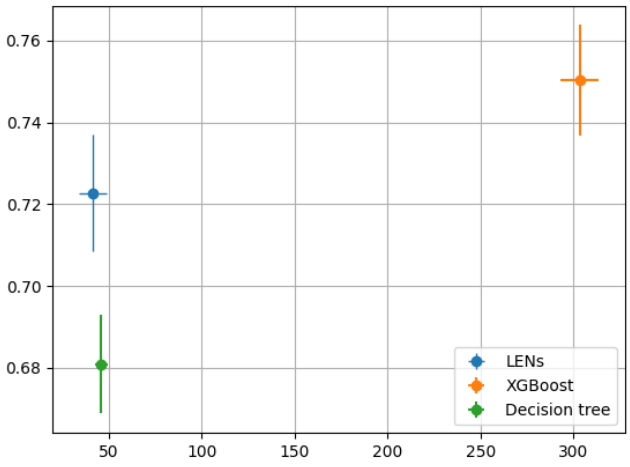

Fig. 3. Complexity plot. The e-LEN, Decision Tree (DT), and XGBoost (XGB) models' complexity scores are plotted on the X axis, while the AUROC values are on the Y axis.

The computed e-LEN complexity was significantly lower compared with XGB complexity. Figure 3 shows the complexity of the three models and their related AUROC. An average of 41 logic predicates were produced by the e-LEN to perform the predictions, compared with the 303 nodes used by XGBoost and 46 for the Decision Tree. For this reason, XGBoost produces higher performance but exploits a more complex model. Moreover, the e-LEN model enables a logic explanation and provides the rules to allow the clinician model validation. The most frequent predicates produced by e-LEN were:

– Predictive predicates of the "healthy" class:
$(tDown5 \wedge \neg tUp3 \wedge stdev1) \vee (tUp3 \wedge nbIdle7 \wedge tDown5) \vee$
$\vee (tDown5 \wedge \neg slopeA6 \wedge \neg tUp3) \vee (tDown5 \wedge \neg slopeA6 \wedge stdev1) \vee$
$\vee (tUp3 \wedge stdev1 \wedge nbIdle7 \wedge \neg slopeA6) \vee (tUp3 \wedge \neg nbIdle7 \wedge slopeA6 \wedge \neg stdev1)$
– Predictive predicates of the " sub-clinical" class:
$(slopeA6 \wedge \neg tUp3 \wedge nbIdle7 \wedge \neg stdev1) \vee (tDown3 \wedge \neg stdev1 \wedge \neg tUp3) \vee$

$$\vee(\neg slopeA6 \wedge \neg stdev1 \wedge \neg nbIdle7) \vee (tUp3 \wedge \neg slopeA6 \wedge tDown3 \wedge \neg nbIdle7) \vee$$
$$\vee(nbIdle7 \wedge \neg tUp3 \wedge stdev1 \wedge \neg tDown3 \wedge \neg slopeA6)$$

There are three to five predicates in each rule. The presence of predicates indicates that an above-average value of that feature is predictive for the class under investigation. The negation $\neg$ means that a value lower than the average is predictive. The predicates connected through the $\wedge$ means the combination of predicates simultaneously is predictive for the class under consideration.

4 Discussion and Conclusion

Analyzing the results obtained, the implemented machine learning approaches proved competitive performance to detect depression using objective physical features. In terms of AUROC, the XGBoost model outperformed its competitor. Nonetheless, the benefits of employing an explainable-by-design technique like e-LEN balance out the performance gap promoting the models' explainability. The interpretable extracted rules of e-LEN allow a clinical validation of the trained models, considering the interpretability as the capacity to convey meaning to a human being in a language that they can comprehend [3, 13].

When considering handwriting and drawing traits, the most predictive features were time, space, and ductus (measured as the number of strokes). The findings make it clear that all of the characteristics considered to be important in determining a person's depression level are strongly correlated with time. This arises from "indirect" features like ductus as well as "direct" features like the amount of time needed to finish the work. As previously indicated, ductus is concerned with the specific strokes used to accomplish the task. In this instance, it turns out that the ductus' most predictive attribute is idle, which is defined as all strokes that the graphics tablet is unable to identify. As a result, we may presume that the participant took the pen off the tablet and pushed it away. It follows that depressed people move more slowly than normal people, as do their cognitive processes. The psycho-motor retardation theory — which postulates that people with depression slow down both cognitively and physically — is strongly supported by these findings [2]. In contrast to earlier research [19], this machine learning technique offers a notable enhancement. Specifically, while looking at the rules that were extracted to improve the explainability of depression versus non-depression, it becomes clear that each of the tasks included in the protocol needs to be thoroughly and inclusively analyzed avoiding a distinction between drawing and writing tasks.

In light of all these findings, the handwriting and drawing methodology is considerably more precise and helpful for quickly and precisely recognizing depressive states and signals when supported by data-driven methods.

Acknowledgements The research leading to these results received funding from the EU-H2020 program grant No. 823907 (MENHIR), the Unicampania Giovani Ricercatori (DR 509/2022) program (DR 834/2022, SALICE), and

the EU NextGenerationE, PNRR Mission 4 Component 2 Investment 1.1 – D.D n.1409 del 14-09-2022 PRIN 2022 – Project code "P20222MYKE" - CUP: B53D2302598000 (IRRESPECTIVE)

References

1. Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., Melacci, S.: Entropy-based logic explanations of neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 6046–6054 (2022)
2. Bennabi, D., Vandel, P., Papaxanthis, C., Pozzo, T., Haffen, E.: Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications. BioMed research international **2013** (2013)
3. Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S.: Benchmarking and survey of explanation methods for black box models. arXiv preprint arXiv:2102.13076 (2021)
4. Bottesi, G., Ghisi, M., Altoè, G., Conforti, E., Melli, G., Sica, C.: The italian version of the depression anxiety stress scales-21: Factor structure and psychometric properties on community and clinical samples. Comprehensive psychiatry **60**, 170–181 (2015)
5. Cordasco, G., Buonanno, M., Faundez-Zanuy, M., Riviello, M.T., Likforman-Sulem, L., Esposito, A.: Gender identification through handwriting: an online approach. In: 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000197–000202. IEEE (2020)
6. Cordasco, G., Scibelli, F., Faundez-Zanuy, M., Likforman-Sulem, L., Esposito, A.: Handwriting and drawing features for detecting negative moods. In: Italian Workshop on Neural Nets. pp. 73–86. Springer (2017)
7. Cordasco, G., Scibelli, F., Faundez-Zanuy, M., Likforman-Sulem, L., Esposito, A.: Handwriting and drawing features for detecting negative moods. Quantifying and Processing Biomedical and Behavioral Signals 27 pp. 73–86 (2019)
8. Depression, W.: Other common mental disorders: global health estimates. Geneva: World Health Organization **24** (2017)
9. Edition, F., et al.: Diagnostic and statistical manual of mental disorders. Am Psychiatric Assoc **21**, 591–643 (2013)
10. Esposito, A., Raimo, G., Maldonato, M., Vogel, C., Conson, M., Cordasco, G.: Behavioral sentiment analysis of depressive states. In: 2020 11th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000209–000214. IEEE (2020)
11. Greenberg, P.E., Fournier, A.A., Sisitsky, T., Simes, M., Berman, R., Koenigsberg, S.H., Kessler, R.C.: The economic burden of adults with major depressive disorder in the united states (2010 and 2018). Pharmacoeconomics **39**(6), 653–665 (2021)
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
13. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
14. Large, M.: Study on suicide risk assessment in mental illness underestimates inpatient suicide risk. Bmj **352** (2016)

15. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
16. Prinzi, F., Militello, C., Scichilone, N., Gaglio, S., Vitabile, S.: Explainable machine-learning models for covid-19 prognosis prediction using clinical, laboratory and radiomic features. IEEE Access **11**, 121492–121510 (2023)
17. Prinzi, F., Orlando, A., Gaglio, S., Vitabile, S.: Breast cancer classification through multivariate radiomic time series analysis in dce-mri sequences. Expert Systems with Applications **249**, 123557 (2024)
18. Prinzi, F., Orlando, A., Gaglio, S., Vitabile, S.: Interpretable radiomic signature for breast microcalcification detection and classification. Journal of Imaging Informatics in Medicine pp. 1–16 (2024)
19. Raimo, G., Buonanno, M., Conson, M., Cordasco, G., Faundez-Zanuy, M., McConvey, G., Marrone, S., Marulli, F., Vinciarelli, A., Esposito, A.: Handwriting and drawing for depression detection: A preliminary study. In: Applied Intelligence and Informatics: Second International Conference, AII 2022, Reggio Calabria, Italy, September 1–3, 2022, Proceedings. pp. 320–332. Springer (2023)
20. Raschka, S.: Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software **3**(24) (Apr 2018). https://doi.org/10.21105/joss.00638
21. Semrau, M., Alem, A., Ayuso-Mateos, J.L., Chisholm, D., Gureje, O., Hanlon, C., Jordans, M., Kigozi, F., Lund, C., Petersen, I., et al.: Strengthening mental health systems in low-and middle-income countries: recommendations from the emerald programme. BJPsych Open **5**(5), e73 (2019)
22. Taleb, C., Khachab, M., Mokbel, C., Likforman-Sulem, L.: Feature selection for an improved parkinson's disease identification based on handwriting. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). pp. 52–56. IEEE (2017)
23. Woody, C., Ferrari, A., Siskind, D., Whiteford, H., Harris, M.: A systematic review and meta-regression of the prevalence and incidence of perinatal depression. Journal of affective disorders **219**, 86–92 (2017)