

BACHELOR THESIS  
ARTIFICIAL INTELLIGENCE

**Radboud University**



---

Identifying Patients at Risk for Suicidal Ideation and Key Factors Responsible  
by Means of a Self-Explaining Neural Network

---

Author

Lukas L. Deis  
s4588827  
L.Deis@student.ru.nl  
DeisLukas@gmail.com

Supervisors  
Radboudumc

Dr. Rose Collard  
Department of Psychiatry  
rose.collard@radboudumc.nl

Supervisor  
Radboud University

Dr. Pim Haselager  
Department of Artificial Intelligence  
& Donders Institute  
w.haselager@donders.ru.nl

Dr. Peter Mulders  
Department of Psychiatry  
& Donders Institute  
peter.cr.mulders@radboudumc.nl



**Radboudumc**  
university medical center

---

14 February 2021

### **Abstract**

Suicide is a major cause of death in all of Europe, and it is on the rise.

Worldwide, suicide is the second most common cause of death in the age group of 15 to 29 years (Bachman, 2018). Suicide experiences a treatment gap of around 50%, meaning that only half the people that die by suicide receive treatment beforehand (Bruffaerts et al., 2011). Thus, it is worthwhile to investigate whether an automated solution could be used to identify people in the general population that may be at risk for suicide. Not only being able to detect suicidal ideation, but also giving insight into the underlying issues, is key to better treatment (WHO, 2019b). Therefore this should be done using explainable methods. A self-explaining neural network (SENN) was trained to predict if a person suffers from suicidal ideation and state which factors were important in that prediction. For this research data from the MIND-SET study was used, which is a study by the Radboudumc. It includes 705 participants, of which 574 suffer from common psychiatric disorders and 131 are healthy controls. The best performing model had an accuracy of 85.3% on the test set with a sensitivity of 79.1% and a specificity of 89.1%; the positive predictive value was approximately 81.538% and the negative predictive value was 87.5%. Most important risk factors are from two questionnaires. One is the Outcome Questionnaire, designed to capture a low quality of life. The second is the Inventory of Depressive Symptomatology, designed to give insight into depression. Some factors seem to significantly reduce the risk, too, such as a score describing a good mental health. Perhaps surprisingly, most other relevant risk-reducing factors stem from the measurement of autism characteristics.

**Contents**

<b>Abstract</b>	<b>2</b>
<b>1 Introduction</b>	<b>6</b>
<b>2 Preliminaries</b>	<b>9</b>
2.1 Neural Networks	9
2.1.1 Concept	9
2.1.2 Activation functions	10
2.1.3 Loss and gradient descent	11
2.1.4 Data	12
2.2 Identity Mapping	13
2.3 Self Explaining Neural Networks	13
2.3.1 Mechanism	13
2.3.2 General Structure	15
2.3.3 Different components	16
2.3.3.1 Conceptizer	16
2.3.3.2 Parameterizer	17
2.3.3.3 Aggregator	18
2.3.3.4 Output	18
2.3.4 Critique	19
2.4 MIND-SET	19
2.5 Utilized Measurements	21
2.5.1 Accuracy	21
2.5.2 Sensitivity and Specificity	21
2.5.3 Positive and Negative Predictive Value	22
2.6 Suicidal Ideation	23
2.7 Used Abbreviations	23
2.7.1 Included Questionnaires	23
2.7.2 Important Questions	24
2.8 Related work	25

<b>3 Method</b>	<b>27</b>
3.1 The Data	27
3.1.1 Augmentation and Amount	27
3.1.2 Variables	27
3.1.2.1 The Target Question	27
3.1.3 Participants	28
3.1.4 Splitting the Data	29
3.2 Assessments and Measures	30
3.3 Baseline Model	30
3.4 Model	<b>31</b>
3.4.1 Utilizing a Self-Explaining Neural Network (SENN)	31
3.4.2 Implementation	32
3.4.2.1 General Implementation	32
3.4.2.2 Designed Output	32
3.4.2.3 Structure	33
3.4.2.4 Implementation Conceptizer	34
3.4.2.5 Implementation Parameterizer	34
3.4.2.6 Implementation Aggregator	36
3.4.2.7 Preprocessing	37
3.4.3 Specific adaptations to the model	37
3.4.4 Tackling Class Imbalance	39
3.5 Stability	39
<b>4 Results</b>	<b>40</b>
4.1 The Performance	40
4.1.1 Important Models	40
4.1.2 All Models	41
4.2 Most important factors	42
4.2.1 Underlying structure	43
4.2.2 Risk raising factors	44
4.2.2.1 Best Performing model - No.9	44
4.2.2.2 Second best performing model - No.7	46

IDENTIFYING PATIENTS AT RISK FOR SUICIDAL IDEATION	5
4.2.3 Risk reducing factors	47
4.2.3.1 Best Performing model - No.9	48
4.2.3.2 Second best performing model - No.7	49
4.3 Additional Model	50
<b>5 Conclusion</b>	<b>52</b>
5.1 Performance	52
5.2 Important Factors	52
5.2.1 Risk Factors	53
5.2.2 Protective Factors	53
<b>6 Discussion</b>	<b>55</b>
<b>7 Limitations and Ethical implications</b>	<b>57</b>
<b>8 Future Research</b>	<b>60</b>
<b>9 Possible Applications</b>	<b>61</b>
<b>10 Acknowledgements</b>	<b>61</b>
<b>11 References</b>	<b>62</b>
<b>12 Appendix</b>	<b>69</b>
12.1 MIND-SET Publications	69
12.2 Full Code and Results	73
<b>13 Addendum</b>	<b>74</b>
13.1 How many models are there in total?	74
13.2 Are the results still reproducible given that the models were lost?	74

## 1 Introduction

Suicide is a major cause of death all over the world (WHO, 2019a), being the second common cause of death in the age group of 10 to 30 years (Bachman, 2018). That issue is even more pronounced in the European Union (Eurostat, 2020a). Being able to detect suicidal ideation and understanding important factors is key in treating it better (WHO, 2019b). While some risk factors for suicidal ideation are known, they vary between countries (Nock et al., 2008) and diagnostic tools based on known risk factors are not sufficiently accurate (Runeson et al., 2017).

Suicide experiences a treatment gap of around 50%, meaning that only half the people that die by suicide receive treatment beforehand (Bruffaerts et al., 2011). Thus, it is worthwhile to investigate whether an automated solution could be used to identify people in the general population that may be at risk for suicide. This project aims to expand our understanding of how suicidal ideation can be predicted by identifying risk-factors and taking the first steps to an automated solution. The general application of automated solutions, specifically machine learning, has been shown to carry a wide range of benefits to diagnosis, treatment and research of mental health (Shatte et al., 2019).

One specific example of an application of machine learning are neural networks. They offer high performance, but usually are not transparent and thus hard to interpret. The field of explainable AI aims to solve this issue by providing insight in a number of ways. While explainable AI methods are rarely used, in this case they are crucial. For one, explainability provides a major benefit to clinical applications.

Because suicidal ideation can not be treated directly, the root causes and underlying disorders have to be identified such that they can be treated instead. The explanations could provide the necessary insight. For another, the reasons that the network expects a person to suffer from suicidal ideation could give insight into general risk factors for suicide, helping to identify and reach people at risk.

That is why in this case so called self-explaining neural networks (SENNs) are used, which do provide explanations for the predictions made (see: section [2.3 Self Explaining Neural Networks](#)).

The following research question has been formulated:

*How accurately can a neural network that was trained on data from common psychological tests identify subjects that suffer from suicidal ideation, and what are the most important factors based on the provided explanations?*

In order to answer the research question, 10 SENNs are trained to predict suicidal ideation from questionnaires designed to evaluate different aspects of mental health. From those ten networks, the most accurate two are selected. Their explanations are analyzed to find out what the most important predictors are and the performance of those networks is noted. Naturally, it is desirable for the found factors to be applicable to a broad range of people. That is why a dataset including healthy controls has been chosen.

The explanations given by the SENNs are expected to be focussed on quality of life, general mental health as well as age and gender, because they have been identified as risk factors in the Netherlands, where the data was collected (GGD, 2017). Predicting the model's accuracy is difficult because of the absence of previous examples, but it is expected to perform above the baseline model which is explained in section [3.3 Baseline Model](#).

In the following chapters it will become clear how a SENN was used to identify people that experience suicidal ideation as well as factors relevant to that prediction. First, some preliminary knowledge, including recurring abbreviations will be explained. Second an explanation of what data was used and how it was treated will be given. That is followed by an explanation of used measurements, including the baseline model, and a description of the utilized model. Afterwards the results are listed and conclusions are drawn before the discussion. The paper ends with a look at limitations including ethical concerns, future research and possible applications.

## **2 Preliminaries**

### **2.1 Neural Networks**

Neural networks are computational models. They are able to learn complex and hidden structures in data without any other input while outperforming other models as well as humans in a variety of tasks (LeCun et al., 2015). Interpretability of neural networks is often a concern but there are ways to gain insight into their inner workings (Melis and Jaakkola, 2018). In this section, some relevant concepts to understand neural networks are explained.

#### **2.1.1 Concept**

Neural networks are inspired by how biological neurons work. They are usually made up of several layers, containing a number of artificial neurons. The different layers are connected by means of artificial synapses, for which different implementations exist. Commonly they are modeled as weights, analogous to the biological strength of connections. Every neuron has an activation that is calculated on the basis of the activation of all incoming synapses, weighted by the strength of the synapses.

### 2.1.2 Activation functions

The activation of a neuron can be calculated in a number of ways.

One of the simplest would be the linear activation function. In this case, the activation of a neuron is a weighted sum of all inputs and their strengths, formally:

$$\text{LinearActivation} = \sum_{i=1}^N w_i \cdot a_i$$

where  $N$  is the number of incoming connections,  $w_i$  is the weight of a specific incoming connection and  $a_i$  is the activation of a specific neuron.

Linear functions have two important drawbacks. One is that the algorithm commonly used to train neural networks requires the activation function to have derivatives that have a relation to the input, which linear functions do not. This will be explained in section [2.1.3 Loss and gradient descent](#). The other drawback is that to model non-linear relationships, non-linear functions are necessary.

Usually that issue is solved by adding another step after calculating the activation with the linear activation function. Two commonly used activation functions are the sigmoid and rectified linear activation functions. The sigmoid activation function can be described as:

$$\text{SigmoidActivation} = \frac{1}{1 + e^{-a}}$$

Where  $a$  is the activation calculated by the function for linear activation above. It has the advantage of always scaling outputs to a value between 0 and 1, normalizing the output.

Rectified linear functions are mathematically simpler and thus can be trained faster. During training, some rectified linear functions will suffer from something called “the dying ReLU problem” where a rectified linear unit (ReLU) stops providing any activation because it was set too low once (Agarap, 2018). This is avoided by the leaky rectified linear function which is defined in such a way that it always maintains at least a small activation. Common would be a value of 0.01. The function can be formalized as:

$$\text{ReLU Activation}(a) = \begin{cases} a & \text{if } a > 0 \\ 0.01a & \text{otherwise} \end{cases}$$

Where  $a$  is again the activation calculated by the function for linear activation above.

### **2.1.3 Loss and gradient descent**

When training a neural network, one is essentially minimizing a specified mathematical function. This function is called the loss function. The loss function can be chosen to reflect various properties but usually it describes how close the output of the network was to the desired output.

After calculating the loss, backpropagation can be performed. Backpropagation is an algorithm that efficiently calculates the loss function’s gradient with respect to the weights. Effectively, this means that for every weight in the network it is known how it should change to predict a specific outcome better. This can be repeated for any number of samples and is the most basic step of training a neural network.

In a process called gradient descent the network learns by repeatedly calculating the loss and performing backpropagation through the model, updating the weights to minimize the loss. At some point the weights do not change significantly anymore, they converge. If one has gone through all the data and the weights have not converged yet, the process can be repeated; this is called an epoch. How many epochs are advisable depends on the situation.

An increasing number of epochs usually increases performance. However, with an increasing number of epochs one runs a higher risk of overfitting. Overfitting describes the process where the network gets adapted to patterns that are in the training data but do not generalize. That means that the network is learning information specific to individual samples of the training set, not about what the training set should represent. The performance on the training set would increase but decrease for (more) general data.

#### **2.1.4 Data**

One requires a certain amount of data to train a neural network. How much exactly is intractably hard to determine though, so usually one uses as much data as is available. That can result in utilizing datasets with more than one trillion entries (Fedus, 2021). It is hard to get to a point where more data does not add anything any more. Therefore, weights are often considered converged if they only change less than a specified threshold instead of not at all. Exactly defining that threshold can be hard, as not every threshold is achievable with the available amount of data.

Even knowing how much data was necessary in other cases does not help, because the necessary amount of data depends on the complexity of the (differing) problem, variance in the data, and a plethora of other factors. However, especially in the medical field, people do not have access to such large amounts of data anyways (Baro, 2015). Models trained on small datasets can still perform above chance, but it can be assumed that they would perform better if more data was available. Therefore, the amount of data is often a concern.

## **2.2 Identity Mapping**

Essentially, neural networks provide a mapping from input to output with many steps in between. The most basic mapping is the identity mapping, which simply sets the output to be equal to the input. As this is a constant function, it does not require any training.

## **2.3 Self Explaining Neural Networks**

### **2.3.1 Mechanism**

Self-explaining neural networks (SENNs) propose a general way of utilizing deep learning while maintaining a high level of explainability. They achieve that by expanding on the concept of a linear regression model (LRM). LRMs are known to be quite easily interpretable as they simplify data by assuming linear separability.

Their output can be visualized as a straight line that divides data points into two classes. However, their performance is limited as very few relations are truly linear.

In SENNs, to model non-linear relationships properly instead of actually relying on linear regression, the different parts of a LRM are rebuilt with neural networks. While conceptually the same on an abstract level, the implementation utilizing neural networks allows for the weights to rely on the input and thus the modeling of any kind of relationship. In a SENN the important factors are determined by a module called conceptizer. Instead of the weights for the direct connections between input and output, the relevance of each factor is given by a module called the parameterizer. Another module called the aggregator is what combines the concepts with the weights and produces the prediction.

Neural networks can be trained in a supervised or unsupervised fashion. While the supervised way enables us to embed prior knowledge of the issue, it requires a lot of work and expert knowledge. In contrast, when working in the unsupervised fashion the network is forced to learn the structure from the bottom up and may encounter relationships that were previously unknown during that process. In addition, it does not require any labeling of the data which saves a lot of time. SENNs utilize both supervised and unsupervised learning.

The conceptizer is an autoencoder, a kind of neural network designed to extract important features from data without any supervision. This enables the SENN to determine itself which (combinations of) features are relevant and should be focussed upon. Meanwhile, the SENN is ultimately supposed to classify the output according to labels that we give to it, making it a supervised form of learning. The loss of the aggregator is calculated based on the given labels and the parameterizer.

### 2.3.2 General Structure

Figure 1 shows the general structure of a SENN which is composed of the three previously mentioned components::

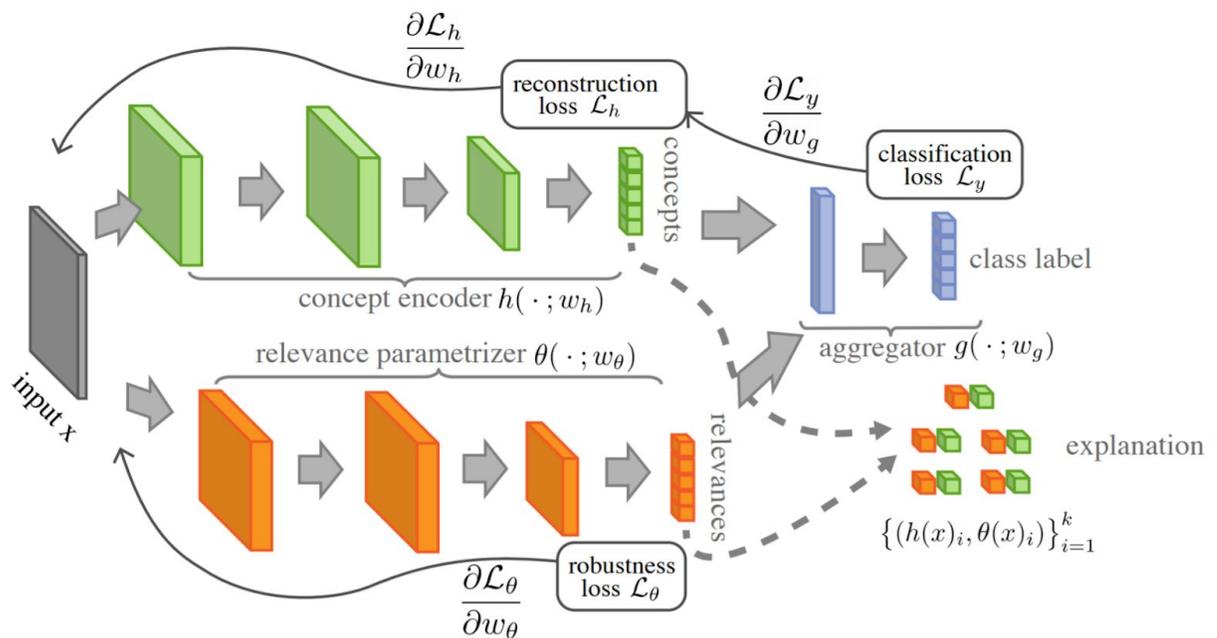


Figure 1: Graphical representation of the SENN model by Melis and Jaakkola, 2018.

### **2.3.3 Different components**

The SENN can be split into 3 sub-models:

- 1) Conceptizer, shown as green part above
- 2) Parameterizer, shown as orange part above
- 3) Aggregator, shown as blue part above

They will be explained more in-depth in the next section.

#### ***2.3.3.1 Conceptizer***

The conceptizer can be considered as a separate deep neural network. The input is equal to the input of the whole SENN. The output is a number of concepts that can be extracted from the input. To achieve that, this network is defined to be an autoencoder. They are used to learn efficient data encodings. The loss function in autoencoders is usually chosen to reflect how well the input can be restored from the output of the network, facilitating the focus on important information in the input. The loss function of autoencoders is usually called the reconstruction loss.

Simply put, the network is trained to compress the input into the most essential features, also called concepts. The number of concepts can be changed to adapt to different problems with the only restriction being that the number has to be equal to the number of output nodes of the parameterizer.

The conceptizers output is used to generate the final classification in the aggregator but is also an output on its own. It encodes the concepts on the basis of which a prediction was made and is hence called 'concepts'.

### ***2.3.3.2 Parameterizer***

Similarly to the conceptizer, the parameterizer can be considered as a separate deep neural network and its input is equal to the input of the whole SENN. In contrast to the conceptizer though, it is not an autoencoder and thus utilizes a different loss function, here called robustness loss. That loss function depends on what issue is at hand and is thus not specified in the general model. However, the robustness loss should be chosen such that incentivizes a proportional relationship between changes in the input and changes in the output. Consequently, similar inputs to this network should generate similar outputs of the parameterizer and big changes in the input should result in big changes in the output.

This property is important to the robustness of this network's output which is supposed to reflect the relevance of the different concepts determined by the conceptizer. If the relevance of different concepts was allowed to change a lot for minor changes in the input, the interpretability would be limited. Every concept, represented by a node, is assigned a relevance score by position. That is why the output of the parameterizer must be equally as long as the output of the conceptizer conceptizer.

The parameterizer's output is used to generate the final classification, in the aggregator but is also an output on its own. It encodes the relevance of every concept to the made prediction and is hence called 'relevances'.

### **2.3.3.3 Aggregator**

The aggregator can be described with a fixed set of mathematical terms. It combines the outputs of the conceptizer and parameterizer to produce the prediction.

Firstly, the concepts are entry-wise multiplied with the relevances. Formally this is known as the Hadamard product, following the formula  $(A \circ B)_i = (A)_i(B)_i$  where  $A$  and  $B$  are the vectors of concepts and relevances and  $i$  is the index within them.

Secondly, to provide a binary prediction, the resulting vector  $v$  is summed up,

which can be formalized as:  $\sum_{i=1}^n v_i$  (where  $n$  is the number of elements in  $v$ ). That

operation yields a singular value  $o$  which is passed through a sigmoid function to scale it to a value between 0 and 1, keeping the output readable. This final operation can be

formalized as:  $S(o) = \frac{1}{1 + e^{-o}}$

### **2.3.3.4 Output**

The SENN provides three outputs:

For one, it provides the prediction calculated by the aggregator. The output between 0 and 1 can be binarized to 0 (negative prediction) or 1 (positive prediction) by using a threshold. For another, the concepts and relevances are provided. They are merged, such that it is clear how relevant every individual concept is to the prediction.

### **2.3.4 Critique**

SENNs, contrary to most other self explaining models, do not suffer significant performance impacts compared to a regular deep neural network (Melis and Jaakkola, 2018). It was considered that SENNs have recently been criticized for offering unreliable explanations (Zheng et al., 2019). That critique is based on the way the important concepts are determined in this technique. Specifically, it has been shown that small changes in the input can cause large changes in the output, casting doubt on how reliable the explanations are. This issue was traced back to the conceptizer, the only component not explicitly enforcing continuity in its output.

It is also unclear how understandable, even if perfectly accurate, the concepts would be in this case. That, in combination with the previously mentioned critique, led to the following decision: The conceptizer will not be implemented in the way that was proposed. Instead, this part of the SENN will be replaced with an identity mapping of the input such that every parameter is treated as a separate concept.

## **2.4 MIND-SET**

The data used in this research were already collected by the Radboudumcs Department of Psychiatry and the Donders Institute. It is made available for research purposes as MIND-SET (Measuring Integrated Novel Dimensions in Neurodevelopmental and Stress-related Mental Disorders). A list of other publications that utilize the same dataset is included as 13.1 MIND-SET Publications.

The purpose of MIND-SET is understanding underlying common and unique mechanisms of psychiatric comorbidity (Collard, 2021).

The data-collection for MIND-SET was planned to run from June 2016 to the end of 2020. As this research was begun in September 2020 the collection was not completed yet. Instead of the planned total of 800 participants, the utilized dataset contains a total of 705 participants (aged between 18 and 76). There are two groups of participants. Patients that are diagnosed with one or more of the following disorders: ADHD, autism spectrum disorder, mood disorder, anxiety disorder and addiction disorder (Radboudumc, n.d.). Next to the patients, a psychiatrically healthy control group is included (n=131).

While the dataset includes a variety of information, for this project only two types of data will be used: Demographics and neuropsychological data, including diagnoses of the disorders mentioned above, but mostly common psychological tests used at psychological treatment facilities. The utilized tests are: OQ, SD, ASI, IDS, CAARS, AQ, SF, PID, NEMESIS, MATE; their purpose is briefly explained in section [2.7.1 Included Questionnaires](#). By limiting the utilized information to this set, all non-structured data is excluded. That enabled the uniform implementation of the beneficial adaptations to the model, explained in section [3.4.3 Specific adaptations to the model](#). Within those tests subjects evaluate themselves how they feel and how much they agree with statements about themselves as well. The precise format of the questions varies per questionnaire. Additionally, results of simple tasks that measure characteristics like attentiveness are included.

## 2.5 Utilized Measurements

All utilized measurements are based on the same, four variables that are collected while testing the model.

True Positives (TP): The number of samples that were correctly identified as positive cases.

True Negatives (TN): The number of samples that were correctly identified as negative cases.

False Positives (FP): The number of samples that were falsely identified as positive cases.

False Negatives (FN): The number of samples that were falsely identified as negative cases.

### 2.5.1 Accuracy

Accuracy gives an idea about the performance of a model on all aspects, encoding how many of all made predictions were correct. It is calculated with the following formula:  $\frac{TP + TN}{TP + FP + TN + FN}$

### 2.5.2 Sensitivity and Specificity

The sensitivity gives an idea of how well the model identifies positive cases. It is the ratio of correctly identified positive cases to all positive cases, which can be expressed in the formula:  $\frac{TP}{TP + FN}$

The specificity, in turn, gives an idea of how well the model identifies negative cases. It is the ratio of correctly identified negative cases to all negative cases, which can be expressed in the formula:  $\frac{TN}{TN+FP}$

### 2.5.3 Positive and Negative Predictive Value

In the medical field, models are often judged by their positive and negative predictive values (PPV and NPV respectively). Those two measurements essentially encode the trust that one should put into a positive or negative diagnosis by the model. For easy comparison within the medical field, the PPV and NPV of the best performing model will be included.

The PPV or “precision” gives an idea of the chance that a disease is present given a positive prediction. Put differently, how many of the cases that are identified as positive are indeed positive cases, using the formula:  $\frac{TP}{TP+FP}$

The NPV or “recall” gives an idea of the chance that a disease is not present given a negative prediction. Put differently, how many of the cases that are identified as negative are indeed negative cases, using the formula:  $\frac{TN}{TN+FN}$

## 2.6 Suicidal Ideation

There are different definitions of suicidal ideation (Harmer et al., 2020) and unifying them is beyond the scope of this research. Agreed upon is, that in the majority of cases persons think of suicide before going as far as committing it (Harmer et al., 2020). For the purposes of this research the following was decided:

The thought of ending one's life, which may or may not go as far as planning the suicide, is referred to as suicidal ideation. Suicidal ideation is treated as a symptom of numerous psychiatric disorders, not a separate one.

## 2.7 Used Abbreviations

### 2.7.1 Included Questionnaires

OQ	Outcome Questionnaire, meant to measure issues with the general state of one's mental health
IDS	Inventory of Depressive Symptomatology, meant to assess the severity of depression
SD	Demographics, meant to capture general demographics and basic background
ASI	Addiction Severity Index, meant to help assess the severity of drug use
CAARS	Conners' Adult ADHD Rating Scales, meant to assess the presence and severity of symptoms related to ADHD
AQ50	Autism spectrum Quotient, meant to quantify the expression of ASD (autism spectrum disorder) related traits

SF20	20 item Short Form, a general, short questionnaire to get a fast and broad idea of overall health
PID	Personal Inventory DSM-5, meant to assess different personality traits
NEMESIS	Netherlands Mental Health Survey and Incidence Study, designed to get an idea about mental health disorders over time
MATE	Measurements in the Addictions for Triage and Evaluation, meant to get insight into substance abuse

### 2.7.2 Important Questions

OQ Total	(low) quality of life
IDS Total	severity of depression
SF20 question 19	being in especially good health
IDS question 2	having issues sleeping through
IDS question 6	being irritable
IDS question 10	feeling sad
CAARS question 2	being always busy as if driven by a motor
CAARS question 8	(still) throwing temper tantrums
OQ question 14	working or studying too much
OQ question 18	being lonely
OQ question 34	having muscle ache
OQ question 36	being nervous often
OQ IR	issues maintaining relationships with others
OQ IR Av	issues maintaining relationships with others, in relation to the average
OQ SD	general symptomatic distress
OQ SD Av	general symptomatic distress in relation to the average

NEMESIS question 10a1	having experienced sexual trauma
SF20 PG	psychiatric health
SF20 PG Scale	psychiatric health point scale
SF20 question 8	having had physical pain recently
SF20 question 16	feeling extremely sad
SF20 question 18	feeling as healthy as everybody else
AQ50 question 5	noticing sounds that others do not
AQ50 question 18	it is hard for others to throw in a word while you talk
AQ50 question 25	being able to deal with a broken routine
AQ50 question 35	usually being the last one to get a joke
AQ50 question 40	having enjoyed games that involve pretending as a kid
AQ50 question 45	having trouble to understand the goals of others
AQ50 question 50	finds it easy to play games that involve pretending with kids

## 2.8 Related work

The high performance that can be achieved with neural networks could be an important step to providing an automated solution to this issue. Yet neural networks are rarely used in this field (Shatte et al., 2019).

However, suicidal ideation has successfully been predicted on the basis of twitter posts and text mining before. In that study a sensitivity and specificity around 80% were achieved, but the important factors could not be identified (Roy et al., 2020) and the applicability beyond twitter is unclear.

There is another investigation into how precisely a number of mental health issues can be predicted on the basis of textual medical history at a psychiatrist. Again the important factors were not identified (Tran & Kavuluru, 2017).

To identify people that are at risk of suicidal ideation, especially those that currently would not get treatment, it is necessary to create a system that works with more general inputs than twitter-posts or medical history. That is only possible with insight into the underlying concepts.

Those underlying concepts have been investigated before. Using Bayesian networks, risk factors for suicidal ideation have been identified in a dataset of depressed people (Galiatsatos et al., 2015). Within that dataset the researchers achieved an accuracy of 83.51%. The identified, most relevant factors were mood depression, loss of interest or pleasure, unworthiness or guilt, living in a city and concentration in thoughts.

While Bayesian networks offer good explainability their performance is often sub-par to neural networks and deep learning approaches. Also, using a bigger dataset, closer to the general society, more insight could be gained into the important factors. If the potentially higher performance of the neural network translates to the performance of the explanations better explanations might be found.

## 3 Method

### 3.1 The Data

#### 3.1.1 Augmentation and Amount

Due to the relatively small size of the dataset, the performance of the network is expected to be suboptimal. This is further explained in section [2.1 Neural Networks](#). Although that was planned, due to time constraints the data was not augmented to enlarge the dataset.

#### 3.1.2 Variables

The participants filled in a maximum of 451 items concerning mental health as well as some demographics, for details on this please refer to section [2.4 MIND-SET](#).

##### 3.1.2.1 *The Target Question*

The risk for suicidal ideation will be extracted from question eight of the OQ, asking for it directly. The question (translated from Dutch) reads:

“I am considering to end my life”; the possible answers (also translated) are: “Never”, “Rarely”, “Sometimes”, “Often”, “Almost all the time”.

Any answer other than “Never” was counted as experiencing suicidal ideation.

A similar question was asked in the IDS (question 18), where the participants have to say which statement they agree with the most. The options are as follows (translated from Dutch): “I do not think about suicide or death.”, “I feel like my life is

empty and question if it is still worth the effort”, “Several times a week I think about suicide or death” and “Several times a day I think about suicide or death; or I made plans to commit suicide”.

The question from the OQ was chosen over the question from the IDS for its stricter focus on suicidal ideation; as thinking about suicide and death is not necessarily the same as considering it. The answers to IDS question 18 were entirely removed from the used data because they are still very similar to the target question, making it too easy for the model to predict the data, removing the need to learn the other present relations.

### **3.1.3 Participants**

This research utilizes data from the ongoing MIND-SET study (Measuring Integrated Novel Dimensions in Neurodevelopmental and Stress-related Mental Disorders) which is executed at the Radboudumcs Department of Psychiatry and the Donders Institute. The inclusion criteria to that study are explained in more detail in section: [2.4 MIND-SET](#). Generally, it can be said that most participants in the study have a psychiatric condition. Also, a psychiatrically healthy control group, consisting of 131 participants is included.

The used dataset contains 705 participants, of which 619 filled in the target question for training (see section [3.1.2.1 The Target Question](#) above for details). The intuitive decision was that participants who did not answer the target question could not be classified and their data could not be used. Thus, the plan was to ignore all data collected from those participants.

However, those patients were not actually removed but counted as not experiencing suicidal ideation. Effectively their lack of an answer was interpreted as answering “Never” to the target question, which is explained in section [3.1.2.1 The Target Question](#). While there is no hard evidence that these 86 participants actually did not experience suicidal ideation, they were judged to have a low risk to do so by the researchers collecting the data, otherwise they would have been asked to fill in the question. Possible effects of this are assessed in section [6 Discussion](#). The upside of this is, that the size of the dataset was not reduced by 12.2%.

### 3.1.4 Splitting the Data

To make sure that the model was not trained and tested on the same data, it was split three ways:

- 67.5% was used for training → 475 participants
- 07.5% was used for validation → 53 participants
- 25.0% was used for testing → 177 participants

### 3.2 Assessments and Measures

Accuracy is the one singular measurement that gives an idea about the performance of the model on all aspects, which is why it is focussed on in the research question. The sigmoid output of the model (range (0,1)) is binarized into positive predictions (experiences SI) and negative predictions (does not experience SI). The threshold for that process is automatically determined such that the accuracy is maximized. As the dataset is not balanced, a relatively high accuracy of 64.4% can be achieved by bluntly predicting the same value all the time (see: section [3.3 Baseline Model](#)). To prevent relying on a possibly deceptively high accuracy, the sensitivity and specificity have been included as measurements as well.

In a medical context, measurements often called precision and recall in an AI context, are usually called positive predictive value (PPV) and negative predictive value (NPV) respectively. People with a medical background are used to judging the performance of a model by the PPV and NPV. To enable quick comparisons, those two measurements have been included for the most accurate model.

### 3.3 Baseline Model

To put the accuracy of the model into perspective it is important to have a baseline. Only that way it can even be said if the model is better than chance or not. Given that the prevalence of suicidal ideation in this dataset is 35.6% a model could always predict that a patient is not suffering from suicidal ideation and be 64.4% accurate. Recall from section [2.5 Utilized Measurements](#) that there are separate

measurements for how well the model works for positive cases ( $Sensitivity = \frac{TP}{TP+FN}$ ) and negative cases ( $Specificity = \frac{TN}{TN+FP}$ ). While this baseline model would have a specificity of 100% it would have a sensitivity of 0%, which is something that should be avoided for the created model.

### 3.4 Model

Neural networks are usually hard to interpret due to their black-box nature. However, to identify the relevant factors, any prediction the network makes needs to be backed up by reasoning. Such an output is also advantageous for any application in treatment of suicidal ideation, as suicidal ideation is not treated itself usually. Rather the root causes have to be identified such that they can be treated. The reasons that the network expects a person to suffer from suicidal ideation could give insight into these underlying issues, enabling or simplifying the treatment.

#### 3.4.1 Utilizing a Self-Explaining Neural Network (SENN)

SENNs are a special kind of neural network designed by Melis and Jaakkola (2018). They showed that, while the widespread perception is that explainability and performance of neural networks oppose each other, this is not the case for SENNs.

The advantage of a SENN over a regular deep neural network is, that it will explain decisions based on training data. For example, it could not only say that a patient is likely to experience suicidal ideation but also what that specific combination of answers was relevant to that decision. This allows experts to reason

about a diagnosis and formulate specific questions to look into the SENNs judgement. As the weights within the SENN are determined by deep learning, the performance can be expected to be comparable or relatively close to other deep learning approaches (Melis and Jaakkola, 2018). For more information on SENNs in general, please refer to section [2.3 Self Explaining Neural Networks](#).

### **3.4.2 Implementation**

#### ***3.4.2.1 General Implementation***

The model is implemented in Python, using TensorFlow 2.0. The functional API was used to maintain as much overview as possible while providing the necessary flexibility. For more information on the code, please refer to section [12.2 Full Code and Results](#).

#### ***3.4.2.2 Designed Output***

The output of the network is a value between 0 and 1 as well as a relevance score for every input feature between -1 and 1. A negative value means that the variable is lowering the risk-score, while a positively valued parameter increases it; a value of 0 would mean that the variable is completely irrelevant.

To provide a classification, the output is binarized using the threshold that yields the maximal accuracy.

### 3.4.2.3 Structure

The implemented version looks as follows:

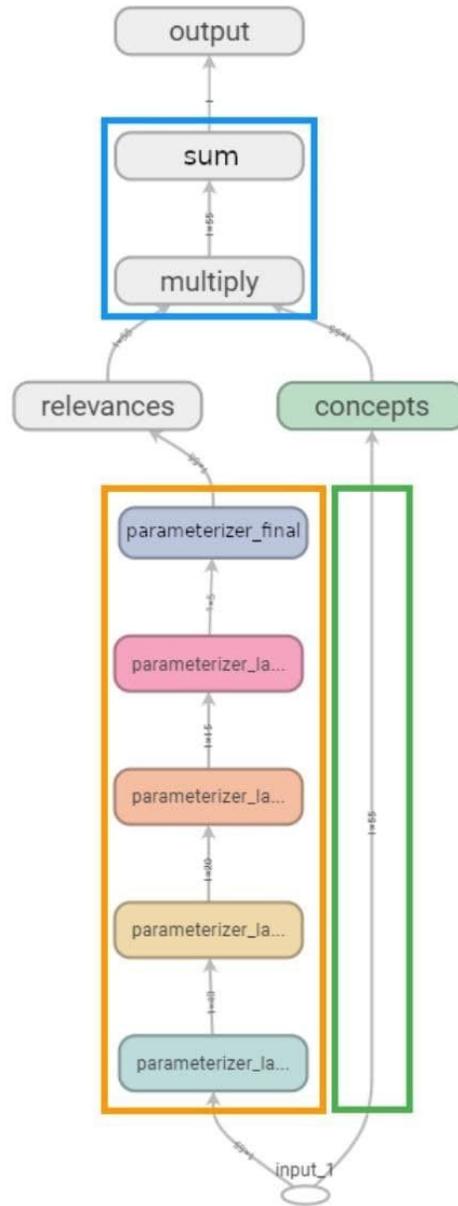


Figure 2: Graphical representation of the implemented model.

The three modules of a SENN are marked as follows: Aggregator: blue, Parameterizer: orange, Conceptizer: green. It should be noted that the Conceptizer indeed is empty, as in this implementation its output is supposed to be equal to its input.

The input at the bottom is passed through the parameterizer, which starts with 4 parameterizer-layers, each including a linear, dropout and leaky-ReLU activated layer before ending in a final layer, consisting of a linearly activated and a dropout layer. The resulting values are multiplied with the concepts, which are equal to the input, before being summed and passed through a sigmoid layer that produces the output.

The SENN can be split into 3 sub-models:

- 4) Conceptizer, shown as green part above
- 5) Parameterizer, shown as orange part above
- 6) Aggregator, shown as blue part above

In the following sections their separate Implementations will be explained.

#### ***3.4.2.4 Implementation Conceptizer***

The Conceptizer is replaced by an identity mapping of the input, meaning that the output of this module is equal to its input. This decision is explained in section [3.4.3 Specific adaptations to the model](#). As identity mappings are not trainable (for more information see section [2.2 Identity Mapping](#)), and loss functions are only used to incentivise certain things during training, no loss function was specified for this part of the SENN.

#### ***3.4.2.5 Implementation Parameterizer***

The original model does not specify what kind of deep learning is used for the Parameterizer. In this case the following structure was chosen:

The following block was repeated 4 times:

- 1) A fully connected layer with linear activation
- 2) A dropout layer
- 3) A fully connected layer with leaky ReLu activation

That block was followed by another fully connected layer with linear activation and a dropout layer. The dropout layers were configured with a dropout rate of 10% and only active during training. They prevent the model from overfitting, by randomly ( $p=10\%$ ) setting any factor in the layer to 0. While the layers with linear activation are a simple way to construct a neural network, more complex relations in the data have to be captured by non-linearly activated units. Therefore, every other trainable layer instead utilizes a leaky ReLU activation function, where leaky ReLU was chosen over regular ReLU to prevent running into issues with vanishing gradients.

The original model does not specify a loss function for the parameterizer, as this needs to be adapted to the problem at hand. Because this is a binary classification problem, binary cross-entropy was selected as the loss-function to train this part of the neural network. It is equal to:

$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

In this specific case, the variables can be interpreted as follows:

$y_i$  is the real label (0: does not experience suicidal ideation, 1: does experience suicidal ideation),  $p(y_i)$  describes the estimated probability of experiencing suicidal ideation and  $N$  is the number of participants in the dataset. The further the output of the network (between 0 and 1) is from the real label (either 0 or 1) the higher and more extreme the loss gets.

Binary cross-entropy is the standard loss function for binary classification problems such as this for multiple reasons, relating to it being a maximum likelihood estimator which comes with a number of benefits (Goodfellow et al., 2016). This loss can only be calculated over the final output of the aggregator. Thus, the loss is calculated there as classification loss and then propagated back through the network, including this module.

While no other efforts have been taken to utilize the other advantages of binary cross-entropy here, future research utilizing this model may profit from an additional property of binary cross-entropy: It usually provides good statistical calibration, enabling the models output to be interpreted as risk-scores instead of just predictions (oW\_♦, 2019).

#### ***3.4.2.6 Implementation Aggregator***

The Aggregator multiplies every concept from the Conceptizer with the corresponding weight from the Parameterizer. The resulting values are then summed together before being passed through a sigmoid function. That keeps the networks' output readable by ensuring it is between 0 and 1. While the aggregator is not trainable itself, the binary cross-entropy is defined as loss over its output. This is done such that the loss can be propagated back through the network and utilized to train the parameterizer.

While using binary cross-entropy as the loss-function should provide a fairly good statistical calibration of the Parameterizers output and the value now is between 0 and 1, it was not tested if the model as a whole is properly statistically calibrated. Therefore, this value will not be treated as a probability.

#### ***3.4.2.7 Preprocessing***

Before the data from the dataset can be worked with, it must be converted to numerical values. As that is a task that will be necessary for any possible follow-up research too it was decided that this should happen within the network.

The input gets passed into TensorFlows Preprocessing layers. They are trained separately but on the same training data as the general network afterwards. There are different preprocessing layers for different types of input.

Dates are converted to the time that passed since that date, converting birth-dates to ages. All numerical inputs are normalized to make sure that something is not seen as more important just because it was measured on a bigger scale. Textual answers are translated by building up dictionaries of possible inputs (the choices of answers for questions) and outputting one-hot vectors encoding the answers; usually a question has 5 possible answers to capture the levels of severity.

#### **3.4.3 Specific adaptations to the model**

One critical aspect of the proposed model of SENNs was altered: the way in which the concepts are learned. As the data is categorical, the whole part of the

model that is responsible to learn the categories was replaced by the preprocessed input data. This has been done before without stating the reasons explicitly (Hussain et al., 2020), but there are good arguments to follow this example. For one there is critique as to how reliable the concepts are if they are determined by a neural network (Zheng et al., 2019). This is explained in more detail in section [2.3.4 Critique](#).

For another, any complex relationship between the different input variables would be hard to interpret. Therefore, a relevance score is calculated for all the singular inputs instead, allowing us to identify singular, general, risk-factors. The proposed way is feasible in certain cases such as optical character recognition, where 451 parameters would fit into a picture of 21 x 22 pixels. Concepts found in such pictures can relatively easily be understood by humans as they themselves can be displayed in pictures. For example, in a model for optical character recognition of numbers one could expect a concept about roundish shapes, made up of pictures of the numbers 6,8 and 9. That concept could be represented by averaging several, very representative pictures together into one, pronouncing the focused upon feature. For pictures, such 'prototypes' can be used to make humans understand what kind of features make up a concept. However, if a prototype or concept is represented by a list of 451 more or less differing values that is not trivial.

To replicate the results of such a study, one would have to have access to all the input represented in such a prototype. Now, the singular identified variables can be tested and applied individually and immediately.

### **3.4.4 Tackling Class Imbalance**

In the data, the prevalence of suicidal ideation is 35.6%, meaning that only 35.6% of the cases should be classified as positive. That class-imbalance could have led to a low sensitivity of the model. To reduce that effect, the minority class was oversampled so that there is an equal number of positive and negative cases in the training data. Due to that, some participants in the training data will be duplicates.

### **3.5 Stability**

Splitting the data into separate sets for training, validating and testing as well as upsampling the minority class includes several random variables that can not easily be controlled. By design, the different datasets change randomly with every split. During upsampling, from all given samples of the minority class, samples are drawn at random. Given the varying inputs, all outputs such as performance and relevance scores have to be expected to vary between runs. The procedure will be repeated 10 times to get some indication of how stable the model is. Specifically, 5 different splits for the training sets will be utilized, each will be used to train 2 models. As a higher accuracy implies that a model captures the structure of the data better, it is expected that the best performing networks will also have the most accurate explanations. Thus, for further evaluation the two models with the highest accuracy will be chosen.

## 4 Results

### 4.1 The Performance

In this section a number of measurements are utilized to present the performance. For an explanation of the measurements please refer to section [2.5 Utilized Measurements](#). The metric for performance of a model is the accuracy.

The ten models trained on data from common psychological tests had an accuracy between 75.1% and 85.3% when identifying subjects that suffer from suicidal ideation. For more information on all models, please refer to Figure 3 below.

Only the parameters deemed most important (11 most risk increasing and 6 most risk reducing parameters are shown and evaluated in this results section.

For reference, a full list of all outputs can be found online at:

<https://gitlab.com/deislukas/senn-identifies-si>.

#### 4.1.1 Important Models

The baseline model (which is explained in detail in section [3.3 Baseline Model](#)) would have an accuracy of 64.4% while the best performing model (No. 9) had an accuracy of 85.3%; next to a sensitivity of 79.1% and a specificity of 89.1%. As explained in section [3.4.4 Tackling Class Imbalance](#), the sensitivity was expected to be lower than the specificity given that the prevalence of suicidal ideation is low, with 37.9% in the test set for this model. That accuracy is not only higher than the baseline models but also what was achieved in the only comparable study that was found, 83.51% (Galiatsatos et al., 2015).

The second best model (No.7) had an accuracy of 83.1%, providing a sensitivity of 79.4% and a specificity of 85.3%. The model with the lowest accuracy was trained on the same train-test split as the best performing model. It provided an accuracy of 75.1% being on par with the best models specificity of 87%, but providing only 59.7% sensitivity. That test set had a prevalence of 38.4%.

#### 4.1.2 All Models

	Prevalence	Accuracy	Sensitivity	Specificity	Precision (PPV)	Recall (NPV)
Model No.1	0.401	0.808	0.803	0.811	0.74	0.86
Model No.2	0.379	0.797	0.776	0.809	0.712	0.856
Model No.3	0.418	0.814	0.73	0.874	0.806	0.818
Model No.4	0.356	0.768	0.73	0.789	0.657	0.841
Model No.5	0.401	0.797	0.761	0.821	0.74	0.837
Model No.6	0.407	0.751	0.861	0.676	0.646	0.877
Model No.7	0.384	0.831	0.794	0.853	0.771	0.869
Model No.8	0.362	0.791	0.703	0.841	0.714	0.833
Model No.9	0.379	0.853	0.791	0.891	0.815	0.875
Model No.10	0.435	0.751	0.597	0.87	0.78	0.737
Baseline	0.356	0.644	0	1	0	0.644

*Figure 3: A table of all included performance metrics. As the performance may be influenced by the prevalence, which differs between the test-sets, the prevalence is included as well. For reference, the hypothetical baseline model is included while the two most accurate models are marked in green.*

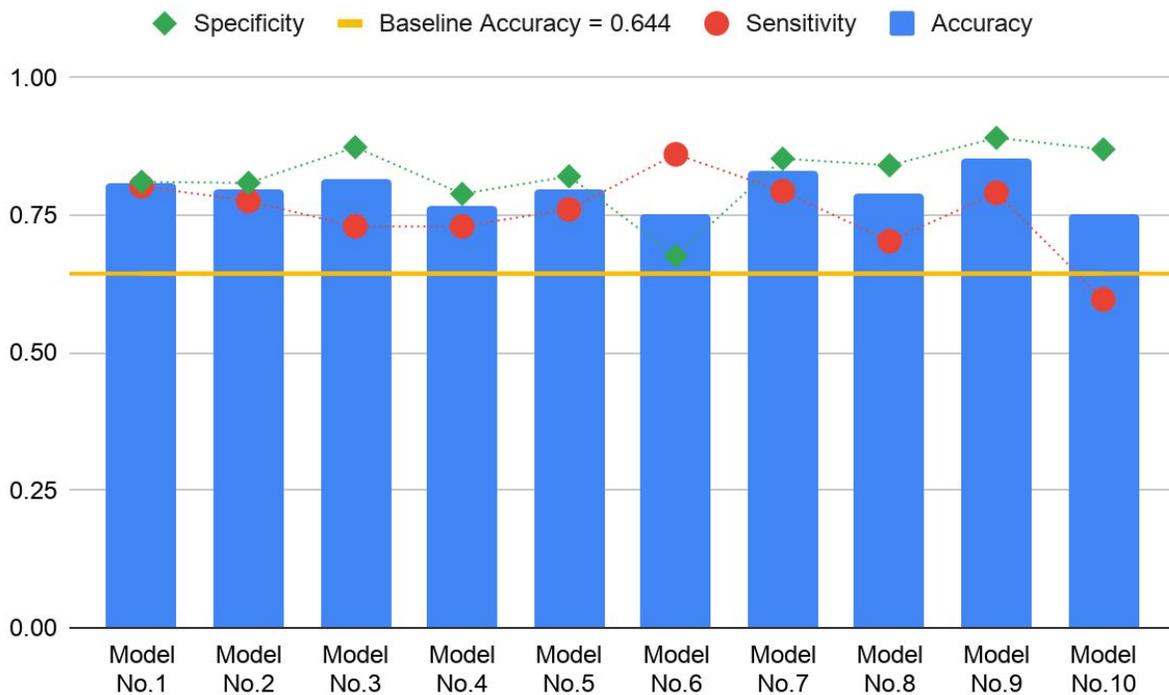


Figure 4: Accuracy of the different models in comparison. The baseline accuracy, sensitivity and specificity are included for reference.

## 4.2 Most important factors

In section [2.7.2 Important Questions](#) all questions mentioned in the upcoming graphs are explained. The relevant factors may differ for positive and negative cases. Thus, the relevance scores for positive predictions (does experience suicidal ideation) and negative predictions (does not experience suicidal ideation) are displayed separately.

#### **4.2.1 Underlying structure**

To look into what the most important factors are, an analysis was performed within two clusters. The clusters were based on if the network predicted a participant to experience suicidal ideation or not. The relevance scores associated with the parameters were compared, assuming that especially informative factors would be more relevant to one group than the other. However, it seems that while there are differences, they are very small; within a few percent points of the relevance score.

*(Data not shown)*

As explained in section [3.4.2.2 Designed Output](#), how influential a parameter was is measured in a normalized relevance score. So the score with the highest influence is seen as 100%.

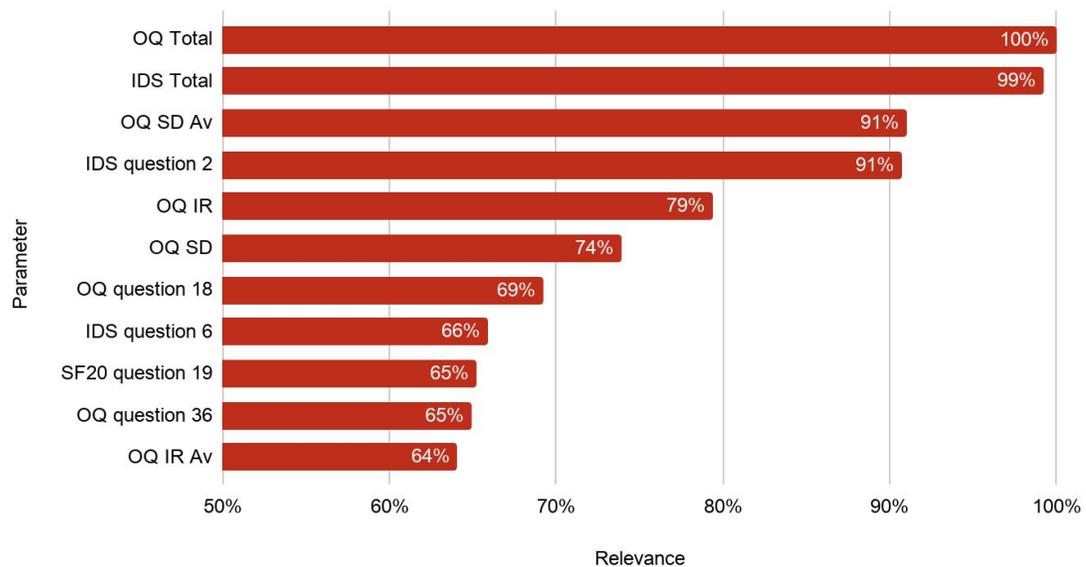
Usually, the questionnaires are evaluated by combining the points for different answers. While the exact formula may differ per questionnaire, usually this is done by simply summing up all points related to a specific score, where more extreme answers equate to more points.

## 4.2.2 Risk raising factors

### 4.2.2.1 Best Performing model - No.9

For both groups, a high OQ total score seems to be the most predictive of high risk. That score is associated with a generally low quality of life.

Relevance of factors to negative predictions (risk raising)



*Figure 5: Important factors that raise the risk to negative predictions in model 9. They can be interpreted as (in that order): low quality of life (100%), severity of depression (99%), general symptomatic distress in relation to the average (91%), having issues sleeping through (91%), issues maintaining relationships with others (79%), general symptomatic distress (74%), being lonely (69%), being irritable (66%), feeling like one is in especially good health (65%), being nervous often (65%), issues maintaining relationships with others in relation to the average (64%)*

In generally high-scoring samples, some of the same parameters were used, but they were seen as more important. Some parameters however seem to only be considered for high-risk patients.

## Relevance of factors to positive predictions (risk raising)

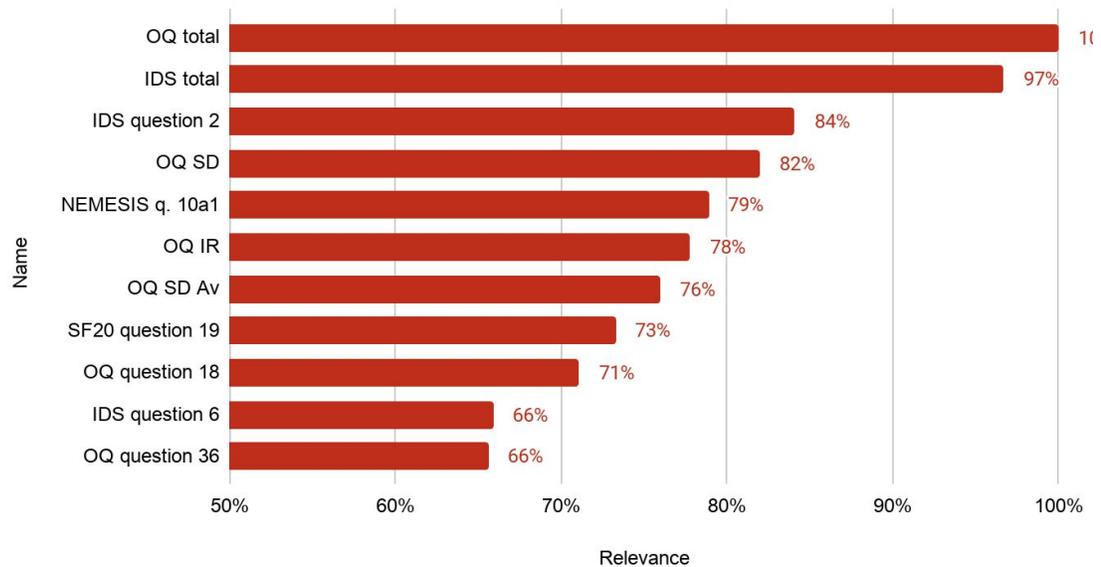


Figure 6: Important factors that raise the risk to positive predictions in model 9. They can be interpreted as (in that order): low quality of life (100%), severity of depression (97%), having issues sleeping through (84%), general symptomatic distress (82%), having experienced sexual trauma (79%), issues maintaining relationships with others (78%), general symptomatic distress in relation to the average (76%), feeling like one is in especially good health (73%), being lonely (71%), being irritable (66%), being nervous often (66%)

Some of those most important parameters are from the OQ, which is commonly used in cases related to suicidal ideation. Its total score is, although mildly, influenced by a direct question for suicidal ideation that is removed from the input and used as the target. However, the choice was made to leave the OQ total score in the input data, as it is one of the most commonly used tests for this purpose and the final score is influenced by many factors. While parameters from the OQ dominate the picture the important parameters include several questions from the IDS and SF20. One question of each the NEMESIS as well as CAARS are included too. For information on the questionnaires please refer to section [2.7.1 Included Questionnaires](#).

#### 4.2.2.2 Second best performing model - No.7

Relevance of factors to negative predictions (risk raising)

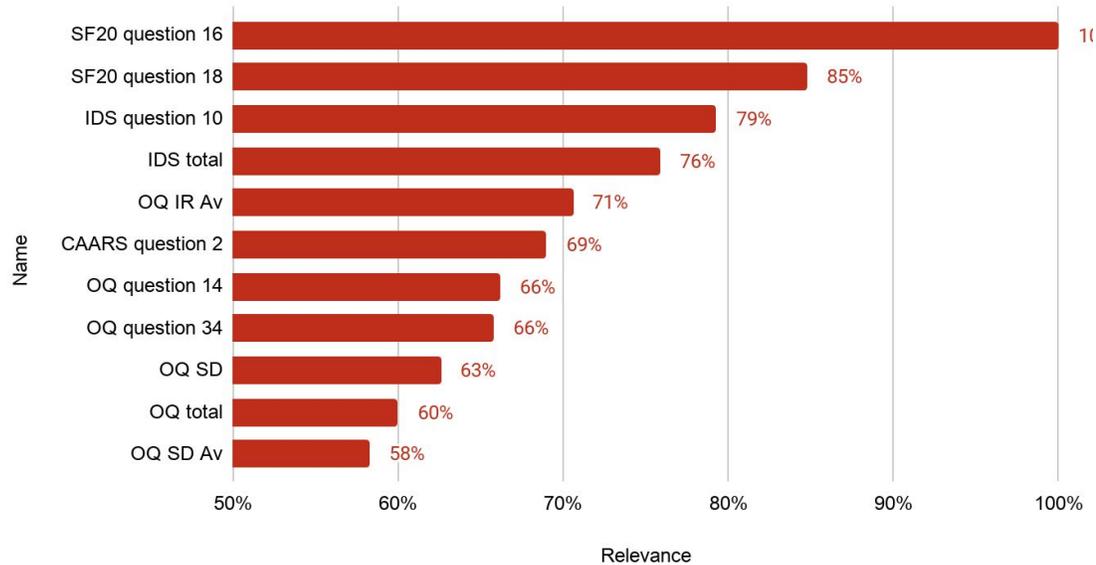
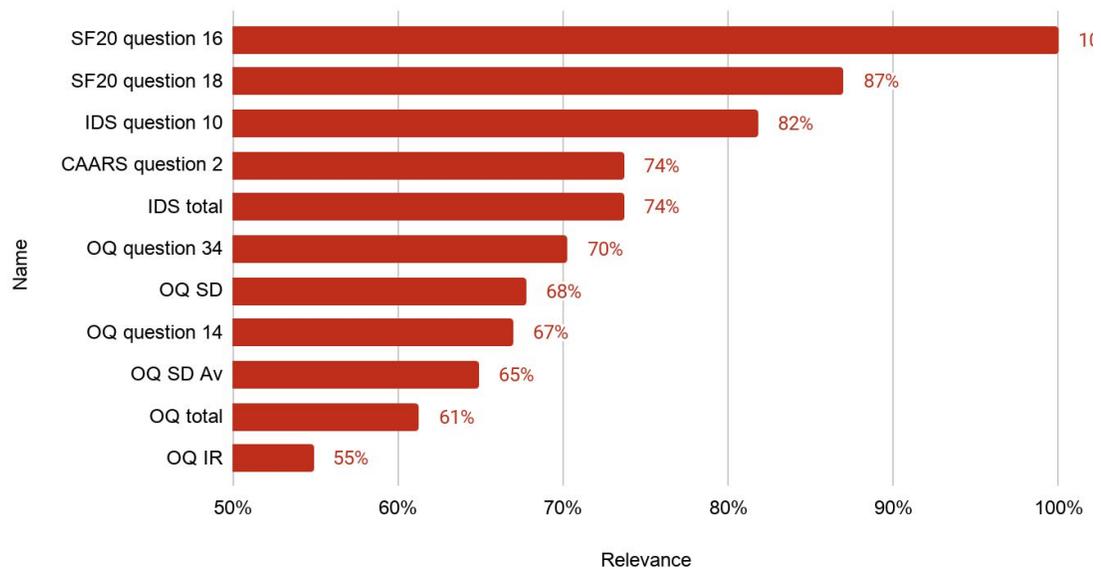


Figure 7: Important factors that raise the risk to negative predictions in model 7. They can be interpreted as (in that order): feeling extremely sad (100%), feeling as healthy as everybody else (85%), feeling sad (79%), severity of depression (76%), issues maintaining relationships with others in relation to the average (71%), being always busy as if driven by a motor (69%), working or studying too much (66%), having muscle ache (66%), general symptomatic distress (63%), quality of life (60%), general symptomatic distress in relation to the average (58%)

## Relevance of factors to positive predictions (risk raising)



*Figure 8: Important factors that raise the risk to positive predictions in model 7. They can be interpreted as (in that order): feeling extremely sad (100%), feeling as healthy as everybody else (87%), feeling sad (82%), being always busy as if driven by a motor (74%), severity of depression (74%), having muscle ache (70%), general symptomatic distress (68%), working or studying too much (67%), general symptomatic distress in relation to the average (65%), quality of life (61%), issues maintaining relationships with others (55%)*

### 4.2.3 Risk reducing factors

Risk reducing factors do not add to the risk, but reduce it. Thus they are assigned negative relevance scores and also displayed as such below. While the order and relevance scores do differ a bit, the parameters seem to be more or less the same again for both clusters in both models. In the best performing model, for positive and negative predictions, question 25 of the AQ50 seemed to be the factor reducing the risk most with a relevance score of -86.8% and -87.7% respectively. That question asks about how well the participant can deal with a broken routine, i.e. their flexibility.

**4.2.3.1 Best Performing model - No.9**

Relevance of factors to negative predictions (risk reducing)

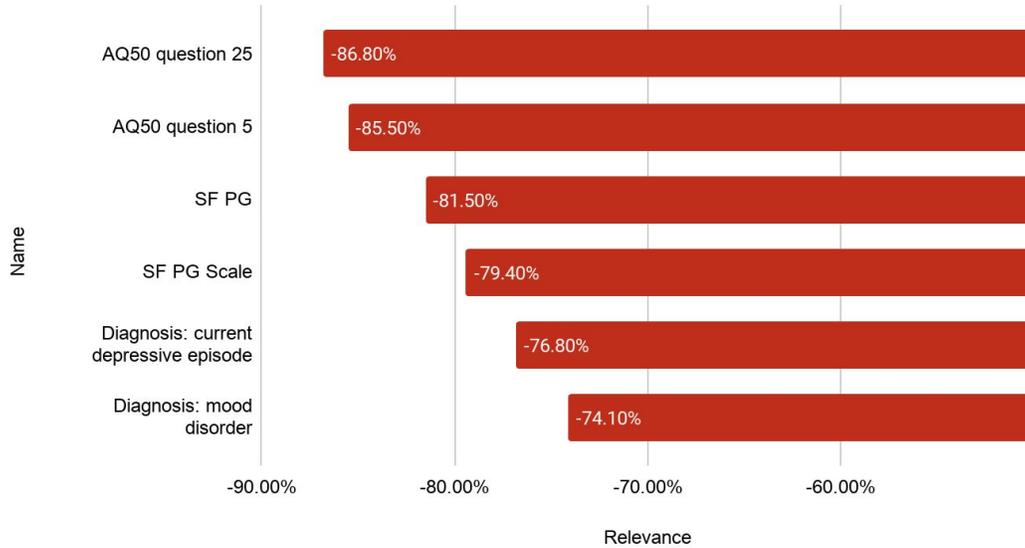


Figure 9: Important factors that reduce the risk to negative predictions in model 9. They can be interpreted as (in that order): being able to deal with a broken routine (-86.80%), noticing sounds that others do not (-85.50%), psychiatric health (-81.50%), psychiatric health point scale (-79.40%), currently experiencing a depressive episode (-76.80%), suffering from a mood disorder (-74.10%)

Relevance of factors to positive predictions (risk reducing)

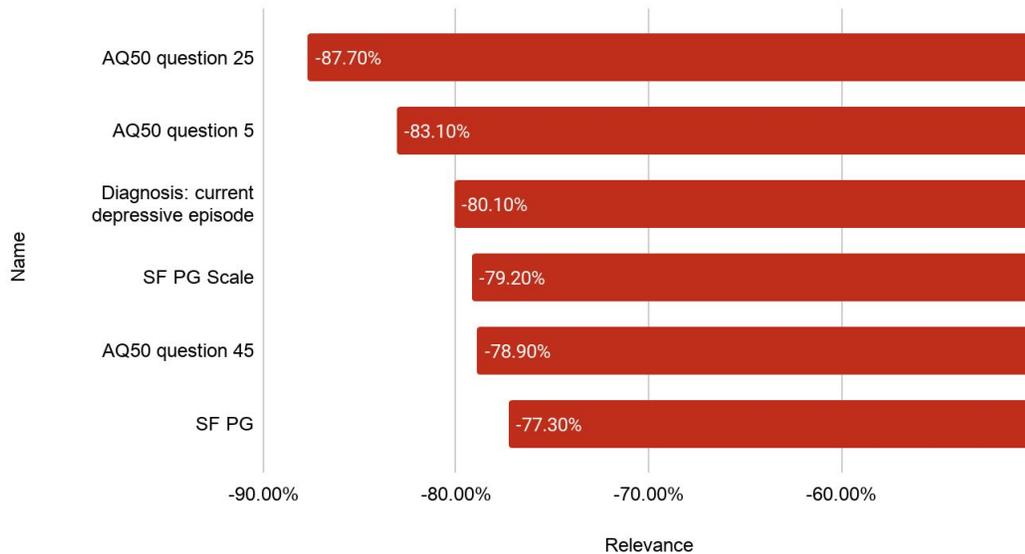


Figure 10: Important factors that reduce the risk to positive predictions in model 9. They can be interpreted as (in that order): being able to deal with a broken routine (-87.70%), noticing sounds that others do not (-83.10%), currently experiencing a depressive episode (-80.10%), psychiatric health point scale (-79.20%), having trouble to understand the goals of others (-78.90%), psychiatric health (-77.30%)

### 4.2.3.2 Second best performing model - No.7

Relevance of factors to negative predictions (risk reducing)

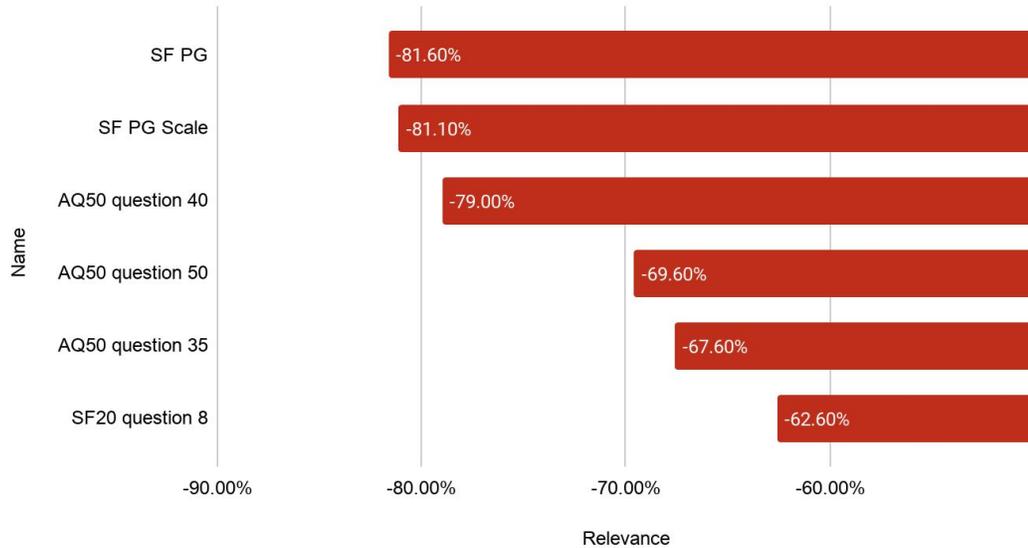


Figure 11: Important factors that reduce the risk to negative predictions in model 7. They can be interpreted as (in that order): psychiatric health (-81.60%), psychiatric health point scale (-81.10%), having enjoyed games that involve pretending as a kid (-79.00%), finds it easy to play games that involve pretending with kids (-69.60%), usually being the last one to get a joke (-67.60%), having had physical pain recently (-62.60%)

Relevance of factors to positive predictions (risk reducing)

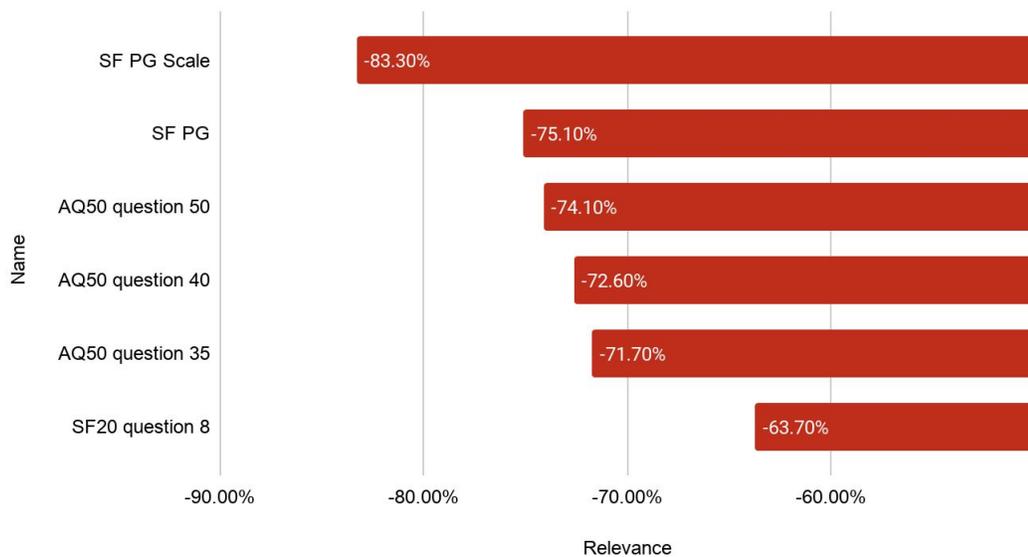


Figure 12: Important factors that reduce the risk to positive predictions in model 7. They can be interpreted as (in that order): psychiatric health point scale (-83.30%), psychiatric health (-75.10%), finds it easy to play games that involve pretending with kids (-74.10%), having enjoyed games that involve pretending as a kid (-72.60%), usually being the last one to get a joke (-71.70%), having had physical pain recently (-63.70%)

### 4.3 Additional Model

The plan was to store all 10 created models for reasons of reproducibility. Unfortunately, all models were stored under the same name which led to them overwriting each other. Therefore, only one, corrupt, model was stored which can not be utilized.

To provide at least one model along with this paper, another model was trained. That model was trained on the same code as all other included models, using the same training and test data as in models 9. That ensures that the data is split the same way. It can be seen in the results of model 10, which shares that property and still is the worst performing model, that this is no guarantee for a high performance. However, the additional model reaches almost the same accuracy as the best performing one (84.18%) and should be interesting to consider.

While the performance is similar to the best performing model, it relies on different factors. While the two best performing models from the initial run heavily rely on combined scores such as the total scores of the OQ and IDS questionnaire, this model considers specific sub-questions of those scores much more relevant.

It seems to focus on what reduces risk, rather than what increases it, as the two most relevant parameters are ones that reduce risk. Specifically, question 18 of the AQ50 is considered the most relevant predictor. It states "While I am talking it is hard for others to throw in a word." The second most important parameter is the general mental health (PG) score from the SF, reducing the risk with 83.8% relevance. Only then parameters that raise the risk follow. Interestingly the 3rd most relevant

parameter is question 8 from CAARS (asking if one experiences temper tantrums), raising the risk with 76.5% relevance. According to this model a high level of education also reduces the risk with a relevance of 70.1%.

## **5 Conclusion**

The goal of this research was to answer the following question:

How accurately can a neural network that was trained on data from common psychological tests identify subjects that suffer from suicidal ideation, and what are the most important factors ?

### **5.1 Performance**

The best performing model had an accuracy of 85.3%, showing that this accuracy is achievable. While this is already better than the baseline model which would have an expected accuracy of 64.4%, it is probable that this can be improved by means outlined in the discussion. The positive predictive value of the best performing model was 81.538%, its negative predictive value was 87.5%.

### **5.2 Important Factors**

Age has been identified as a relevant factor to suicidal ideation by previous research. Not all age groups are equally as likely to have suicidal thoughts (GGD, 2017), so is suicidal ideation more prevalent in the age-group of 19 to 34 years than in the group of 35 to 45 year olds. Suicidal ideation is also more common in males than in females. It has been found that suicide occurs more often in the male population than in the female population (Eurostat, 2020b; Hedegaard et al., 2018) and it has been shown that males also experience suicidal ideation more often (GGD, 2017).

Age and gender were thus expected to be relevant in predicting suicidal ideation. However, they were not at all included in the relevant factors of either model.

### **5.2.1 Risk Factors**

The questionnaires that contain almost all the most important parameters are the Outcome Questionnaire (OQ) and Inventory of Depressive Symptomatology (IDS). The most important parameter to an elevated risk are the OQ total score and two of its subscales; one encoding general symptomatic distress (SD) and the other one reflecting issues with interpersonal relations (IR). The IDS total score is also considered very relevant by both models, with the individual questions 2 and 10 being notably more relevant than the others. Question 2 is about issues with sleeping through a night while question 10 is about being sad.

Furthermore, the questions 16, 18 and 19 of the Short Form survey (SF), seem relevant. While question 16 is about being “so sad that nothing can cheer one up”, questions 18 and 19 are concerned with being in generally good health.

The only question that is relevant to an elevated risk from a different survey is question 2 of CAARS, a questionnaire designed to look into ADHD. That question is about being busy all the time as if ‘driven by an external motor’.

### **5.2.2 Protective Factors**

While some factors differ between the models there are some things they have in common. The risk-reducing variables prominently feature the subscore of the SF

concerned with mental health (PG). It is represented within the top 5 relevant protective factors by both, its scaled outcome and point scale in 3 out of 4 cases.

Essentially meaning that this variable is counted double.

Most other questions featured in any model stem from the AQ, a questionnaire designed to look into autism. That questionnaire is not commonly used to diagnose or look into suicidal ideation. However, it looks like those questions do provide additional insight. Specifically, the following questions were featured: 5, 25, 35, 40, 50. Question 5 is about noticing sounds that others do not which could be interpreted as being attentive to one's surroundings. Question 25 is about dealing with an interrupted routine, while question 35 is about usually being the last one to get a joke. Question 40 asks if someone enjoyed games that involve pretending things when they were a kid and question 50 asks if the person finds it easy to play games that involve pretending things with kids. Questions 35, 40 and 50 can be summed up as 'being imaginative'.

Three more of the top five factors have been identified in both models, all of which were unexpected. Two of them are diagnoses: having a mood disorder and currently experiencing a depressive episode. The third is question 8 from the SF20 which concerns having had physical pain recently. This is especially surprising, as one might expect bodily pain to be related to symptomatic distress as described in the OQ (OQ SD), an identified risk-factor. However, the OQ SD captures a number of variables, and physical pain could still somehow reduce the risk of experiencing suicidal ideation when considered alone.

## 6 Discussion

While not all the models were equally accurate (75.1% - 85.3%), there were several models that performed similarly well; the two best models were only 2.1% points apart (85.3% and 83.1%). Still, their judgement seems to be based on different information. This might be due to overlap between questionnaires or for other reasons. Optimally these models would rely on different structures in the data, in which case ensemble methods would be a promising way to increase performance. Either way, both explanations for decisions should be seen as similarly justified.

At first glance, this accuracy is comparable to the previous research into the important factors to suicidal ideation by Galiatsatos et al. (2015). However, that research was done using Bayesian networks which were not tested on a separate training set. The accuracy measured there only reflects how well the networks trained in that research were able to classify the data that was used in their training. As that study focussed on finding important factors, that did not present an issue. When comparing that accuracy to the one achieved here however, it does. It has to be expected that the performance of that model would be worse when classifying samples from a different dataset. In contrast, within this research a part of the data was held back during training and exclusively used for testing later.

The identified important parameters have to be taken with some caution. While they do seem to be good predictors for the network, we do not understand why for specific cases they seem to work well. It is unreasonable to assume that all people that suffer from suicidal ideation do this for the same reasons and all factors are

always equally important. The analysis is not fine-grained enough to look at those nuances and differences between the underlying specific patterns. Even if that was possible, the learning process of a neural network is based solely on correlation, so found patterns do not necessarily imply causative relationships.

However, the identified risk-factors are related to mental health and seem to match what was expected. While the protective factors were much less expected, and should be further investigated.

The reported accuracy has been compared to a base-model, but the significance of the findings has not been verified by a permutation analysis or anything alike. That leaves the size of the possibility that the data in the test-set was skewed in favour of a higher classification accuracy unknown.

The measured accuracy might have also been influenced by a possible misclassification of 12.2% of the input samples due to a bug in the code. However, it seems that the likelihood of that is low. More details can be found in section [3.1.3 Participants](#). Even if one was to assume that all the samples were misclassified, noisy data would most likely reduce the performance of the model, indicating that it would actually work better than was shown here.

## 7 Limitations and Ethical implications

It is important to keep in mind that the dataset is relatively small and not a representative sample of the society. Most people in the dataset have some sort of condition that might change the way they react to different influences. Despite the included control group it can not simply be assumed that the findings generalize to the general population. If the model was trained on data that does not reflect the use case it might, just like other programs that deal with any kind of data, suffer from an issue colloquially described as “garbage in garbage out” (Sanders & Saxe, 2017). This basically means that if one gives inadequate data to the model one can not expect it to work properly or even remotely in the intended way. This also holds for biased data.

Automated solutions are often seen as a way to counter human biases. However, human biases can be and usually are ingrained into datasets and designs made by humans. That way machine learning may learn and reinforce human bias. That can potentially lead to unfair treatment, which is a concern especially if that bias concerns protected groups<sup>1</sup>(Schönberger, 2019). It is important to avoid bias, as discrimination must be avoided. By validating the reasons the network gives for the prediction, any bias could theoretically be identified, but only under one condition: The human user needs to manually identify it as such, which can be less trivial than it sounds. Consider the following:

---

<sup>1</sup> Such as race, age, sexual orientation, ability, or belief.

The model could consistently predict people with a specific sexual orientation to suffer from suicidal ideation. That could be the result of discrimination against people of that group, leading to difficulties in their lives that could stimulate suicidal ideation. It could also be a bias against that group which was taken over from the dataset. One would be a valid reason, while the other must be avoided.

If the model outputs the sexual orientation as an explicit reason, then one has to find out if that was justified. Did the dataset contain a bias which influenced the model or if it was just that the underlying reasons were not included. If the model does not state the sexual orientation as a reason, it might still implicitly rely on it. Neural networks are extremely good at capturing hidden patterns and have been shown to determine things like sexual orientation from data that humans can not (as well) utilize for that same purpose (Wang & Kosinski, 2018). Thus even if the reasons seem unrelated to a bias, they may contain it and even though the model is more transparent than common neural networks it still includes black-box elements that may behave unexpectedly. Therefore it is important that the output and explanations of the network are used with caution.

No matter how many samples the network is trained on, if only relatively few of those samples belong to a specific group that may lead to issues. The factors relevant to that specific group may differ from the factors that are relevant to the majority of samples. If such factors exist that are specifically and only relevant to such a group, they would not be learned as well as the more common factors. Measures to tackle such imbalances exist and have been implemented in this research. So was the sensitivity of the model raised by oversampling the minority class (see section [3.4.4 Tackling Class Imbalance](#)). However, no measures were taken to compensate for any bias possibly present in the dataset.

Even if the model does suffer from a bias against a protected group, it is still possible that it is not as good at identifying certain cases or conditions as it is in general. Careful consideration is necessary to ensure that the quality of care does not worsen for people with atypical conditions. That also includes considering that clinicians might have trouble handling the output of such a predictive system properly. The system will not always be correct, and it would be dangerous if anybody started relying on the (statistics based) analysis of a neural network too much. To maintain meaningful human control any application of this model has to take this into account.

It should also be noted that it is still under debate to what extent people can accidentally commit suicide. Such a thing could, for example, happen by unintentionally going further than intended with non lethal self-harm. In such cases, detecting suicidal ideation may not be an effective way of preventing the suicide.

## 8 Future Research

One could investigate if a more regular deep neural network in combination with post-hoc explanation methods could reveal different information about the factors. However, it seems reasonable to first try to improve the performance of this model. This could be done by fine-tuning parameters such as the number of epochs during training or the size of the layers. It is also an option to combine several of these models in an ensemble approach or utilizing different data.

Either way, a bigger dataset could help improve the performance. If one was to find a label-invariant transformation that could be applied to the data, be it from the same dataset or another, data augmentation would be an option. The transformation could easily be integrated in the code utilized in this project.

With or without those adaptations, it would be interesting to see if and how the performance changes with alterations to which data is used. One could train the model on only combined scores of the questionnaires or remove those combined scores, leaving only singular questions. That would certainly change the explanations, possibly revealing new important factors.

Likewise, one could investigate how the model performs on a more limited subset of questions in general or on a dataset with people that represent the general population better. Even looking into different topics all together is possible utilizing this basic setup.

## **9 Possible Applications**

The concept shown here can be used as it is when treating patients that suffer from suicidal ideation. Patients in psychiatric care settings are usually asked to fill in at least some questions contained in the dataset. If those answers were fed to such a model and the prediction is correct, the explanations could give insight into why the patient is experiencing suicidal ideation, aiding in treatment.

Also, it should be possible to find proxies for the identified risk-factors in other settings. Training a similar model on those may enable the implementation of early warning systems in schools and other settings where entrusted persons can be informed and try to reach out to the person at risk.

## **10 Acknowledgements**

I would like to thank my supervisors Dr. Pim Haselager, Dr. Rose Collard and Dr. Peter Mulders for their invaluable feedback and insight. Without the combination of their perspectives this would not have been the same. Without the critical questions of Meilina Reksoprodjo however, probably nobody would have been able to understand what I made of that insight; thank you! Dr. Petra Muckel made sure that people are not only able to understand me, but also feel the comfort of proper formatting and style, which I am very grateful for. Had I lost motivation though none of this would have mattered though. I want to thank all my friends and family for their continued support and especially my boyfriend Julian for helping me to have the energy for this. Thank you!

### 11 References

- Agarap, A. F. (2018). Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.
- Bachmann, S. (2018). Epidemiology of suicide and the psychiatric perspective. *International journal of environmental research and public health*, 15(7), 1425. <https://doi.org/10.3390/ijerph15071425>
- Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed research international*, 2015.
- Bruffaerts, R., Demyttenaere, K., Hwang, I., Chiu, W., Sampson, N., Kessler, R., . . . Nock, M. (2011). Treatment of suicidal people around the world. *British Journal of Psychiatry*, 199(1), 64-70. <https://doi.org/10.1192/bjp.bp.110.084129>
- Collard, R. (2021). Measuring Integrated Novel Dimensions in Neurodevelopmental and Stress-related Mental Disorders. Manuscript submitted for publication.
- Eurostat (2020a, June 19). Causes of death - standardised death rate. Retrieved September 25, 2020, from

[https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Causes\\_of\\_death\\_%E2%80%94\\_standardised\\_death\\_rate,\\_2017\\_\(per\\_100\\_000\\_inhabitants\)\\_Health20.png](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Causes_of_death_%E2%80%94_standardised_death_rate,_2017_(per_100_000_inhabitants)_Health20.png)

Eurostat (2020b, June 02). Causes of death - standardised death rate by residence.

Retrieved from

[https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK\\_DS-417853\\_QID\\_-1C267FA\\_UID\\_-3F171EB0&layout=SEX,L,X,0;GEO,L,Y,0;UNIT,L,Z,0;TIME,C,Z,1;AGE,L,Z,2;ICD10,L,Z,3;INDICATORS,C,Z,4;&zSelection=DS-417853AGE,TOTAL;DS-417853ICD10,X60-X84\\_Y870;DS-417853UNIT,RT;DS-417853INDICATORS,OBS\\_FLAG;DS-417853TIME,2017;&rankName1=ICD10\\_1\\_2\\_-1\\_2&rankName2=TIME\\_1\\_0\\_-1\\_2&rankName3=UNIT\\_1\\_2\\_-1\\_2&rankName4=AGE\\_1\\_2\\_-1\\_2&rankName5=INDICATORS\\_1\\_2\\_-1\\_2&rankName6=SEX\\_1\\_2\\_0\\_0&rankName7=GEO\\_1\\_2\\_0\\_1&rStp=&cStp=&rDCh=&cDCh=&rDM=true&cDM=true&footnes=false&empty=false&wai=false&time\\_mode=NONE&time\\_most\\_recent=false&lang=EN&cfo=%23%23%23%2C%23%23%23.%23%23%23](https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-417853_QID_-1C267FA_UID_-3F171EB0&layout=SEX,L,X,0;GEO,L,Y,0;UNIT,L,Z,0;TIME,C,Z,1;AGE,L,Z,2;ICD10,L,Z,3;INDICATORS,C,Z,4;&zSelection=DS-417853AGE,TOTAL;DS-417853ICD10,X60-X84_Y870;DS-417853UNIT,RT;DS-417853INDICATORS,OBS_FLAG;DS-417853TIME,2017;&rankName1=ICD10_1_2_-1_2&rankName2=TIME_1_0_-1_2&rankName3=UNIT_1_2_-1_2&rankName4=AGE_1_2_-1_2&rankName5=INDICATORS_1_2_-1_2&rankName6=SEX_1_2_0_0&rankName7=GEO_1_2_0_1&rStp=&cStp=&rDCh=&cDCh=&rDM=true&cDM=true&footnes=false&empty=false&wai=false&time_mode=NONE&time_most_recent=false&lang=EN&cfo=%23%23%23%2C%23%23%23.%23%23%23)

Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint arXiv:2101.03961

Galiatsatos, D., Konstantopoulou, G., Anastassopoulos, G., Nerantzaki, M., Assimakopoulos, K., & Lymberopoulos, D. (2015, September). Classification of the most significant psychological symptoms in mental patients with depression using Bayesian network. In Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS) (pp. 1-8).

<https://doi.org/10.1145/2797143.2797159>

GGD Brabant-Zuidoost (2017). Serieuze suïcidegedachten onder volwassenen.

Retrieved January 23, 2021, from

<https://www.ggdbzo.nl/ggdkompas/Documents/Infographic%20su%C3%AFci%20gedachten%20volwassenen.pdf>

Goodfellow, I., Bengio Y., & Courville A. (2016). Deep Learning. MIT Press, 172-175.

<http://www.deeplearningbook.org>

Harmer, B., Lee, S., Duong, T., & Saadabadi, A. (2020). Suicidal Ideation. StatPearls.

Hedegaard, H., Curtin, S. C., & Warner, M. (2018). Suicide rates in the United States continue to increase (pp. 1-8). Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Hussain A., Elbaghdadi, O., Bardarov I., Hoenes C. (2020, January 31). Self Explaining Neural Networks: A Review with Extensions.

<https://amanhussain.com/publication/self-explaining-neural-networks/>

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

<https://doi.org/10.1038/nature14539>

Melis, D. A., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems* (pp. 7775-7784).

Nock, M., Borges, G., Bromet, E., Alonso, J., Angermeyer, M., Beautrais, A., . . . Williams, D. (2008). Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *British Journal of Psychiatry*, 192(2), 98-105.

<https://doi.org/10.1192/bjp.bp.107.040113>

oW\_♦(2019). What makes binary cross entropy a better choice for binary classification than other loss functions?

Retrieved January 28, 2021, from

<https://datascience.stackexchange.com/questions/53400/what-makes-binary-cross-entropy-a-better-choice-for-binary-classification-than-o>

Radboudumc (n.d.). MIND-SET. Retrieved September 25, 2020, from

<https://www.radboudumc.nl/trials/mind-set>

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020).

A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ digital medicine*, 3(1), 1-12.

<https://doi.org/10.1038/s41746-020-0287-6>

Runeson, B., Odeberg, J., Pettersson, A., Edbom, T., Jildevik Adamsson, I., &

Waern, M. (2017). Instruments for the assessment of suicide risk: a systematic review evaluating the certainty of the evidence. *PLoS one*, 12(7), e0180292.

<https://doi.org/10.1371/journal.pone.0180292>

Sanders, H., & Saxe, J. (2017). Garbage in, garbage out: How purport-edly great ML models can be screwed up by bad data. *Proceedings of Blackhat 2017*.

Schönberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*, 27(2), 171-203.

<https://doi.org/10.1093/ijlit/eaz004>

Shatte, A., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9), 1426–1448.

<https://doi.org/10.1017/S0033291719000151>

Tran, T., & Kavuluru, R. (2017). Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks. *Journal of biomedical informatics*, 75, S138-S148.

<https://doi.org/10.1016/j.jbi.2017.06.010>

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2), 246. <https://doi.org/10.1037/pspa0000098>

WHO (2019a). Suicide.

Retrieved January 23, 2021, from

<https://www.who.int/teams/mental-health-and-substance-use/suicide-data#:~:text=Close%20to%20800%20000%20people,prevent%20suicide%20and%20suicide%20attempts>

WHO (2019b). Suicide.

Retrieved January 23, 2021, from

<https://www.who.int/news-room/fact-sheets/detail/suicide>

Zheng, H., Fernandes, E., & Prakash, A. (2019). Analyzing the Interpretability  
Robustness of Self-Explaining Models.

## 12 Appendix

### 12.1 MIND-SET Publications

A list of all publications that resulted from MIND-Set, kindly provided by Rose Collard.

No.	Title and authors	(Target) journal	Date first registration	Status (Date)	No.
1	Alexithymia mediates the relationship between childhood trauma and emotion regulation difficulties in psychiatric patients. Everaert D., Schene, A.H., Schellekens, A.F.A., Collard, R.M., van Eijndhoven, P., van Oostrom, I., Vrijzen, J.N.	COTR	25-04-2018	In-prep: 30-06-2020	
2	Systematic Review of Affective Cognitive Biases in Autism Spectrum Disorder: Towards an Understanding of the Prevalent Comorbidity with Depression. Bergman, M.A., Schene, A.H., Constance Th.W.M., Vrijzen, J.N., Kan, C., van Oostrom, I.		4-5-2018	4-5-2018	
3	Affective attentional biases in Autism Spectrum Disorder and/or Major Depressive Disorder: an eye-tracking study. Bergman, M.A.,		4-5-2018	4-5-2018	

	Constance Th.W.M., Vrijzen, J.N., Rinck, M.M., Schene, A.H.				
4	Attentional biases in Neurodevelopmental & Mood Disorders: a Network approach Bergman, M.A., Constance Th.W.M., Vrijzen, J.N., Rinck, M.M., Brolsma, S.C.A. Schene, A.H.		4-5-2018	4-5-2018	
5	Continued stress from a network perspective (analysis on the resting state scan after stress induction, compared to the resting state scan after a neutral control condition). J. van Oort, I. Tendolkar, A. Schene, G. Fernandez, P. van Eijndhoven		04-06-2018	Accepted	1
6	Challenging the negative learning bias hypothesis of depression: reversal learning in a naturalistic psychiatric sample. S.C.A. Brolsma, J.N. Vrijzen, E. Vassena, M. Rostami Kandroodi, M.A. Bergman, P. van Eijndhoven, R.M. Collard, H.E.M. den Ouden, A.H. Schene, R. Cools	Psychological Medicine	16-10-2018	Accepted	2
7	Negative learning bias in depression revisited: Enhanced neural response to surprising reward across psychiatric disorders.	Biological Psychiatry: Cognitive Neuroscience and Neuroimaging	16-10-2018	Accepted	4

	S.C.A. Brolsma*, E. Vassena*, J.N. Vrijzen, G. Sescousse, P. van Eijndhoven, R.M. Collard, A.H. Schene, R. Cools				
8	BOLD activity in visual association areas is modulated by unexpected outcomes. S.C.A. Brolsma, J.N. Vrijzen, E. Vassena, A.H. Schene, R. Cools		16-10-2018	External research group will continue	
9	Relation between childhood adversity and the volume of the hippocampus and amygdala J. van Oort, I. Tendolkar, A.H. Schene, P. van Eijndhoven		13-11-2018	In prep	
10	Is an attentional deficit (ADHD) the "cure" for negative attentional bias in depressed patients? Schuthof, C., Tendolkar, I., Collard, R.M., Eijndhoven, P., Schene, A.H., Vrijzen, J.N.		01-06-2019	In prep: 08-12-2020	
11	The Importance of Perseverative Cognition for Both Mental and Somatic Disorders in a Naturalistic Psychiatric Patient Sample. Appel, J., Schene, A.H., Eijndhoven, P., Collard, R.M., Tendolkar, I., Vrijzen, J.N.		01-06-2019	Submitted: 08-12-2020	
12	Negative memory bias as a transdiagnostic cognitive marker for depression symptom severity. Duyser,	Journal of Affective Disorders	24-07-2019	Accepted	3

	F.A., Van Eijndhoven, P.F.P., Bergman, M.A., Collard, R.M., Schene, A.H., Tendolkar, I., Vrijisen, J.N.				
13	Amygdala reactivity as neural correlate of negative memory bias in a naturalistic psychiatric patient sample. Duyser, F.A., Vrijisen, J.N., Van Oort, J., Collard, R.M., Schene, A.H., Tendolkar, I., Van Eijndhoven, P.F.	Biological Psychiatry: Cognitive Neuroscience and Neuroimaging	24-07-2019	In prep: 08-12-2020	
14	Anhedonia as a Transdiagnostic Symptom in symptom clusters of Depression, Anxiety Sensitivity, ADHD and Autistic Traits: A Network Approach. Guineau, M., Ikani, N., Rinck, M., Collard, R.M., Van Eindhoven, P.F.P., Schene, A.H., Becker, E., Vrijisen, J.N.		30-06-2020	In prep: 08-12-2020	
15	Transdiagnostic brain-behavioral mapping using sparse multiple canonical correlational analysis (msCCA) regression Peter Mulders, Andre Marquand, Philip van Eijndhoven, Indira Tendolkar, Aart Schene		01-05-2020	Analysis started	

16	Neural correlates of repetitive negative thinking (relation of RNT (measured with the PTQ) with resting state functional connectivity and with stress induced changes in connectivity). J. van Oort, I. Tendolkar, R. Collard, D. Geurts, A.H. Schene, P. van Eijndhoven		30-06-2020	In prep	
17	Relation of psychiatric symptoms with resting state connectivity and stress induced changes in connectivity (a Linked ICA analysis). J. van Oort, I. Tendolkar, A.H. Schene, P. van Eijndhoven		30-06-2020	Analysis started	
18	Pooled SRET project: an examination of the task to find optimal outcome variables/best predictors for psychopathology. Duyser, F.A., Van Eijndhoven, P.F., Schene, A.H., Tendolkar, I., Vrijssen, J.N.		30-06-2020	Cleaning up data started (November 2020)	

## 12.2 Full Code and Results

The full code, including results from all runs can be found here:

<https://gitlab.com/deislukas/senn-identifies-si>

### **13 Addendum**

- 1) How many models are there in total?
- 2) Are the results still reproducible given that the models were lost?

#### **13.1 How many models are there in total?**

There are eleven models in total. The eleventh model is excluded from the general analysis as it was decided that there was no reason to deviate from the planned procedure for this project. The plan was to create ten models to analyze, so the analysis should be about the ten models created for this purpose. The secondary reason is that the eleventh model was created after the analysis had already progressed fairly far and time constraints had to be considered.

The only reason that the eleventh model is mentioned at all, is that it differs significantly from the other models that were analyzed, which may put the results of this research into a different perspective.

#### **13.2 Are the results still reproducible given that the models were lost?**

If one wanted to reproduce the results critically, it would have probably been best to train new models either way. Otherwise one would not independently reproduce the results, but rather retest the created models, reproducing only a part of what was done.

The implementation of all models is the same, the same code was run several times. Consequently, the only thing that changes between the runs is how the dataset

is split (as that is done randomly) and then how the network changes in response (the pre-processing layers introduce some random variance when adapting to the data).

If one would like to extend this work it should be said that it is probably possible to train the model in a more optimized way, while any adaptation requires training a new model either way. Whether one would like to reproduce the outcomes of this research or extend upon them, they require two things:

- 1) The code, which is available online as described in section [12.2 Full Code and Results](#). With minor adaptations to the environment one is working in, it should be easy to utilize it.
- 2) Data. For information on the data that was used to conduct this research, please refer to section [2.4 MIND-SET](#).

Consequently, it is possible to reproduce the results of this research as well as to expand upon them.