Long-Short Term Memory Networks for Electric Source Imaging with Distributed Dipole Models

Hecker, Lukas[1,2,3,4,5]; Rupprecht, Rebekka[6]; Tebartz van Elst, Ludger[1,5]; Kornmeier, Jürgen[1,3,4,5]

[1] Department of Psychiatry and Psychotherapy, University of Freiburg Medical Center, Freiburg, Germany.
[2] Department of Psychosomatic Medicine and Psychotherapy, University of Freiburg Medical Center, Freiburg, Germany.
[3] Perception and Cognition Lab, Institute for Frontier Areas of Psychology and Mental Health, Freiburg, Germany.
[4] Faculty of Biology, University of Freiburg, Freiburg, Germany.
[5] Faculty of Medicine, University of Freiburg, Freiburg, Germany.
[6] Now working with SICK AG, Waldkirch, Germany.

*Corresponding author: lukas.hecker@uniklinik-freiburg.de

- **Funding:** The project was funded by the European campus (eucor) seeding money
- **Conflict of interest:** The authors declare that there is no conflict of interest regarding the publication of this article.

1 Abstract

Magneto- and electroencephalography (M/EEG) are widespread techniques to measure neural activity *in vivo* at a high temporal resolution but relatively low spatial resolution. Locating the sources underlying the M/EEG poses an inverse problem, which is itself ill-posed. In recent years, a new class of source imaging methods was developed based on artificial neural networks. We present a long-short term memory (LSTM) network to solve the M/EEG inverse problem. It integrates several aspects essential for qualitative inverse solutions: Low computational cost, exploitation of both spatial and temporal information, input flexibility and robustness to noise. Using simulation data, the LSTM shows higher accuracy on multiple metrics and for varying numbers of neural sources, compared to classical inverse solutions but also compared to our alternative architectures without integration of temporal information. It successfully integrates the temporal context given by sequential data, which is particularly useful with noisy data. Real data of a median nerve stimulation paradigm was used to show that the LSTM predicts plausible sources that are in concordance with classical inverse solutions. The performance of the LSTM regarding its robustness to noise renders it a promising and easy-to-apply inverse solution to be considered in future source localization studies and for clinical applications.

2 Introduction

Magneto- and electroencephalography (M/EEG) are non-invasive tools to measure human brain activity with high temporal resolution. The ease of use and temporal precision has made it one of the most used techniques in human neurophysiology. M/EEG suffers, however, from a considerably low spatial resolution. Inferring the active brain regions (henceforth called *sources*) given a signal measured on the scalp is not possible without adding further constraints, posing what is called the *inverse problem* (Grech et al., 2008; He, Sohrabpour, Brown, & Liu, 2018; Michel & Brunet, 2019; R. D. Pascual-Marqui, 1999). The reason for this is the underdetermined equation one has to solve: There are many more neurons (or ensembles thereof) in the brain than there are EEG electrodes. Furthermore, *volume conduction*, i.e., the summation of currents of various sources in the brain, also renders signals ambiguous and thus lowers the spatial resolution of the EEG further. Solving the inverse problem is useful for various reasons, e.g., to localize EEG responses in basic research (Kornmeier, Friedel, Hecker, Schmidt, & Wittmann, 2019) or aberrant activity as in pre-surgical epilepsy diagnostics (Mégevand, Hamid, Dümpelmann, & Heers, 2019).

In order to solve the inverse problem, one has to introduce some constraint on the solution. One such constraint is the *minimum norm*, used in various ways in a group of inverse solutions referred to as Minimum Norm Estimates (MNE; Hämäläinen & Ilmoniemi, 1994; Ioannides, Bolton, & Clarke, 1990). These inverse solutions aim to find the source configuration that minimizes the required power to generate a given potential at the scalp electrodes. Low Resolution Electromagnetic Tomography (LORETA) and its iterations are famous proponents of the MNE-family that forces sources to be smoothly distributed through the spatial Laplacian operator (R. Pascual-Marqui, Michel, & Lehmann, 1994; R. D. Pascual-Marqui, 2007). Beamformers constitute a new category of inverse solutions that assumes sources to be uncorrelated and suppresses noise (Van Veen, van Drongelen, Yuchtman, & Suzuki, 1997). While MNE-based inverse algorithms and Beamformers are fairly simple-to-apply and effective in finding sources at the correct locations, they are known to misjudge the size of active brain sites. Another family of inverse solutions uses the Bayesian inference scheme with great success, e.g., multiple sparse priors (Friston et al., 2008) and maximum entropy on the mean (Amblard, Lapalme, & Lina, 2004; Chowdhury, Lina, Kobayashi, & Grova, 2013; Wipf & Nagarajan, 2009). Although these inverse algorithms have shown to, not only find the location but also estimate the correct size of sources, they consume a considerable amount of computing power during inference and are thus suitable when time is not an issue. For more time critical applications like neurofeedback, however, the processing time of the aforementioned Bayesian approaches are not feasible.

Artificial neural networks (ANNs) have made their first appearance in 1943 and saw an important leap in development with the perceptron (McCulloch & Pitts, 1943; Rosenblatt, 1958). A couple of decades later, ANNs found their first application in solving the inverse problem of the EEG based on single or few dipoles (e.g., Abeyratne, Zhang, & Saratchandran, 2001; Robert, Gaudy, & Limoge, 2002). With the large increase of computing power

and data resources in the past two decades, ANNs have gained in popularity and are now used successfully in a variety of tasks, e.g., image classification (Krizhevsky, Sutskever, & Hinton, 2012) and classification of single trial EEG (Schirrmeister et al., 2017). With this leap in technology, ANNs are now being reconsidered to solve the inverse problem in M/EEG and various research groups are starting to develop and refine architectures (see Awan, Saleem, & Kiran, 2019; Fedorov, Koshev, & Dylov, 2020; Razorenova et al., 2020; Zorzos, Kakkos, Ventouras, & Matsopoulos, 2021 for reviews). Deep ANNs were considered for inverse problems in other domains, too (Jin, McCann, Froustey, & Unser, 2017).

In our recent study, we proposed a simple convolutional neural network (CNN) termed ConvDip (Hecker, Rupprecht, van Elst, & Kornmeier, 2020) that solves the EEG inverse problem on single time points. As input, ConvDip takes a single time point of EEG and outputs the corresponding estimated source vector, thereby creating inverse solutions independent of temporal context. Pantazis and Adler (2021) developed two model architectures, a multi-layer perceptron (MLP) -type architecture and a CNN-type architecture. The MLP model was constructed similarly to ConvDip, the difference being that there was no first convolutional layer and an alternative output structure. The output vector of the last layer only contained the position (x-, y- and z-coordinate) of one, two or three sources. Thereby, the number of expected sources must be known in advance. Despite promising results and their simple design, these models discard a considerable amount of information by ignoring the temporal dimension and reduce the spatial pattern to merely a predefined number of dipoles, which discards the complex and partially coherent functional source configurations which are to be expected in the brain (Destexhe, Contreras, & Steriade, 1999; Leopold, Murayama, & Logothetis, 2003).

In the first paper on ANN-based inverse solutions over the past years, Cui et al. (2019) introduced a long short-term memory (LSTM) network trained on simulated EEG data which was designed to find the location and time-course of single-dipole sources. Since the model transforms one spatiotemporal input into a spatiotemporal output of the same length, it can be called a sequence-to-sequence (seq2seq) model. This was an important idea, resulting in a leap in ANN-based inverse solutions. However, the simulation framework was a vast simplification of real M/EEG activity and the results can be only regarded as a proof-of-concept for the feasibility of LSTMs in solving the inverse problem. Single-dipole sources pose an oversimplification of realistic cerebral activity patterns which can be assumed to have the following properties: They are distributed and (to a varying degree) clustered into active sites of varying size.

Sun, Sohrabpour, Ye, and He (2020) followed a different approach and designed a spatiotemporal CNN called *SIFNet*. The model transforms an EEG sequence of fixed length (500 samples) to an output of summed source activity. Unlike the architecture of Cui et al. (2019), a sequence-to-one (seq2one) model was implemented. Albeit limited by the lack of a temporal dimension of the output vector, the model showed promising results compared to standardized LORETA (sLORETA). Particularly, it showed great robustness to varying levels of noise in the input data, which appears to be one key advantage of ANN-based inverse solutions over classical approaches. Furthermore, in scenarios where the temporal

dimension of source activity can be discarded (e.g., when localizing a segment of epileptiform activity) this approach may prove especially useful.

Another spatiotemporal model developed by Pantazis and Adler (2021) harnesses temporal context using two-dimensional spatiotemporal filters and four fully-connected layers. Similar to the author's earlier MLP architecture described above, the output vector represented dipole positions. Therefore, this architecture also yields seq2one transformations.

An interesting approach was published by Huang et al. (2020), who designed the Data-Synthesis-Based Denoising Autoencoder (DST-DAE). The architecture is based on an autoencoder network with convolutional (encoding) and deconvolutional (decoding) layers. The architecture requires a fixed-size temporal dimension, which was set to 40 time slices. Although the DST-DAE was trained with a single and two source patches, the model showed promising results to reliably estimate both source positions and their sizes at varying levels of noise.

Given the high temporal resolution of the M/EEG in the range of 1 millisecond and estimates about the durations of neural processing steps in the range between $10ms$ and $80ms$, the temporal dimension (i.e., EEG data from neighboring time points as input) can contain important information for source localization estimates. One approach to add memory to any inverse solution was proposed by Dinh, Samuelsson, Hunold, and Hämäläinen (2019), who incorporated temporal context into dynamic statistical parametric mapping (dSPM, Dale & Sereno, 1993) using an LSTM network and Markov chains. While this method constraints each prediction using prior data, it does not change the implicit or explicit priors given by the inverse solver (e.g., dSPM). One drawback of this method is that temporal context can only be used from past but not from subsequent samples. Furthermore, the model needs a few samples to build a sufficient history, which means that the first few time points do not benefit from any temporal context. This makes sense for online source analysis where only previous time points are available. However, the method misses out on information during offline analyses, where both preceding and subsequent time points are available as context.

In order to bring ANN-based inverse solutions out of the infancy state there are some issues that need to be tackled. Each of the existing ANN-based inverse solvers has at least one of the following shortcomings: (1) Large number of parameters, (2) physiologically implausible source outputs (i.e., single dipole position/s), (3) inability to exploit temporal information flexibly.

In this work we present an extension of the initial ConvDip, a bidirectional LSTM network which solves the M/EEG inverse problem for distributed dipole models in a seq2seq manner. Inspired by Godard, Matzen, and Uyttendaele (2018), we created a list of criteria an ANN-based algorithm for solving the inverse problem should optimally fulfil:

1. **Accessibility:** The ANN should be computable on common computers with the number of parameters not exceeding 500,000.

2. **Flexible input:** The ANN should be able to work both on single time instances and time series data (e.g., event-related potentials) of varying length.

3. **Flexible priors:** The assumptions on the number, size and shape of sources as well as assumptions on temporal structure of time courses should be chosen carefully to facilitate adaptive inverse solutions.

4. **Robustness to noise:** The ANN should be capable of handling signals with variable SNRs appropriately, especially when applied to online EEG data.

Note, that these criteria don't just affect the design of the network but also the properties of the simulated training data. In order to fulfil criterion (1) we have tried to keep the number of trainable parameters as low as possible. Furthermore, we analyzed the sparsity-performance trade-off. In conjunction with criterion (2) we decided to use a sequence-to-sequence (seq2seq) LSTM architecture which accepts inputs of different numbers of time points (see Fig. 1 C). Criterion (3) is fulfilled if the simulations are maximally diverse. This regards parameters like the number of source patches, their amplitudes, sizes, shapes as well as their temporal progression. We aimed to define particularly large value ranges of these parameters in our simulations, which are explained in more detail in the chapter *Source and EEG Simulation* and outlined in Table 1. Criterion (4) is handled in conjunction with criterion (3) since the ANNs' ability to correctly estimate sources from noisy signals to some degree depends on the noisiness of the data it is confronted with. We therefore simulated trials with a wide range of signal-to-noise ratios (SNRs). To our best knowledge, there is currently no ANN-based algorithm for the M/EEG inverse problem that satisfies all of these criteria.

3 Methods

3.1 Forward Model

We used an anatomical template brain "fsaverage" (Fischl, Sereno, Tootell, & Dale, 1999) by the Freesurfer image analysis suite[1]. EEG simulations were carried out using a precomputed three shell boundary element method (BEM; Fuchs, Kastner, Wagner, Hawes, & Ebersole, 2002) forward solution as provided by mne-python (v20.3, Gramfort et al., 2013). Each shell (brain, skull & scalp tissue) was composed of 5120 vertices. The conductivity was set to $0.3S/m^2$ for brain and scalp tissue, and $0.06S/m^2$ for the skull.

In order to spare computational resources, we chose a source model with low spatial resolution of $p = 1284$ dipoles with icosahedral spacing. For the EEG electrodes we used the Easycap-M1 layout consisting of $q = 61$ electrodes of the 10-20 system. Using the forward model and the parameters described, we calculated a leadfield $K \in \mathbb{R}^{q \times p}$, that assigns a gain value to each electrode for every source dipole in the brain. Therefore, it is often referred to as *gain matrix*. A fixed orientation of dipoles perpendicular to the cortical surface of the cortex was used (Michel & Brunet, 2019).

[1] http://surfer.nmr.mgh.harvard.edu/

Parameter	Parameter Range	Domain
Number of sources	[1, 10]	spatial
Diameter	[5, 40] mm	spatial
Max. dipole Moment	[1, 10] nAm	spatial
Source shape	{gaussian, flat}	spatial
Source time course frequency spectrum	$[1/f^{1.5}, 1/f^{0.5}]$	temporal
EEG noise frequency spectrum	$[1/f^{1.5}, 1/f^{0.5}]$	temporal
EEG signal-to-noise ratio (trial-avg)	[2, 10]	temporal

Table 1: **Simulation Parameters**

3.2 Source and EEG Simulation

Our goal with the design of the simulation was to make as few assumptions as possible with respect to spatial and temporal properties. The parameters in our simulation are given in Table 1. For each simulated sample, the number of active sources was chosen randomly between 1 and 10. Furthermore, for each active source patch we randomly chose its diameter, maximum dipole moment, shape and the frequency spectrum of its time course. Diameters were chosen between 5 mm and 40 mm. The use of extended sources (as opposed to single dipoles) is motivated by numerous studies showing that neural sites of activity are coherent in space and time (Destexhe et al., 1999; Leopold et al., 2003). The shape parameter refers to the dipole moment attenuation with increasing distance to the seed location. "Flat" refers to a uniform dipole moment across all members within the chosen source patch diameter, whereas "Gaussian" refers to a Gaussian drop-off of dipole moment with respect to distance to the seeding dipole. The number of time points for each sample was chosen randomly between 1 and 1000 to train the LSTM to use different sequence lengths. The variable-length time series were generated randomly for each source patch with a frequency spectrum defined by $1/f^{\beta}$, whereas β was randomly chosen between 0.5 and 1.5 for each source individually. This approach yields highly variable time courses from aberrant (e.g., as in noisy raw EEG recordings) to smooth (e.g., as in event-related potentials). The source simulation yielded a source matrix $Y \in \mathbb{R}^{p \times t}$ with $p = 1284$ dipoles and $t = \{1, 2, ..., 1000\}$ time points. This source matrix was multiplied with the leadfield, which yields the simulated and noise-free EEG $M \in \mathbb{R}^{q \times t}$:

$$M = K \cdot J + \zeta, \tag{1}$$

Noise of a random frequency spectrum ($\frac{1}{f^{\beta}}$, $0.5 < \beta < 1.5$) was then added to the EEG, scaled such that a randomly chosen SNR (within the range given in Tab. 1) was achieved in relation to the averaged global field power (GFP) of signal and noise.

We simulated 10,000 samples of spatiotemporal source activity and corresponding EEG. Prior to training, each EEG sample and time step was re-referenced to common average. Furthermore, each time slice of EEG was normalized by using a custom implementation of the robust scaling procedure as described by the scikit-learn python packages (Pedregosa

et al., 2011). Sources were re-scaled by division of their absolute maximum, which deviates from the normalization procedure of the EEG due to the sparsity of the source simulations.

3.3 Neural Networks

We will describe and test three different ANN architectures: (1) The ConvDip architecture as presented in Hecker et al. (2020). (2) A fully-connected (FC) network and (3) the LSTM network. The neural networks were built and trained using Python and the machine learning libraries Tensorflow 2.5.0 (Abadi et al., 2016) and Keras 2.5.0 (Chollet et al., 2015). We designed each ANN to have $\approx 500,000$ parameters which we deem is a good balance between computational complexity and model capacity. Mathematical notation of the three architectures are given in Appendix A.

3.3.1 ConvDip

ConvDip was originally designed to receive EEG data from a single time point of EEG data (M_t). Instead of feeding the raw EEG as input it was first interpolated into a two-dimensional 7×11 scalp map. For the present work we adjusted this interpolation procedure by using a larger scalp map of size 9×9. This interpolated scalp map serves as input into the ConvDip model. The ConvDip architecture was described already in Hecker et al. (2020) and will hence only be briefly summarized. We adjusted the ConvDip architecture to fulfil our requirement of having $\approx 500,000$ parameters (Fig. 1, A). The architecture was built using two convolutional layers with kernel sizes of 3×3 and 8 filters each. The convolutional layers were then followed by three fully-connected ("FC") layers of 250 units (i.e., neurons) and an output layer with the size of the source model (1284 units). Each intermediary layer used rectified linear units (ReLU, Nair & Hinton, 2010) activation functions and was followed by a dropout layer with a rate of 20%. No activation function was applied to the output layer. The total number of parameters sums up to $498,698$.

3.3.2 Fully-Connected Network

The FC network is a simple multilayer perceptron with 2 hidden layers with $k = 300$ neurons each (Fig. 1, B) and an output layer of the size of the source model (1284 neurons). ReLU activation functions were applied at each hidden layer and dropout layers with a drop rate of 20% followed. The total number of parameters sums up to $495,384$.

3.3.3 Long Short-Term Memory Network

The LSTM network has a more complex architecture since its computing units are not simple neurons but so-called LSTM cells. These cells allow the network to retain (and forget) information from previous samples and thus foster predictions that are dependent on the temporal context. This sets the model apart from the previously described architectures. An LSTM cell consists of four different components: The input gate, forget gate, cell state and output gate. Multiple LSTM cells can comprise one LSTM layer with the advantage

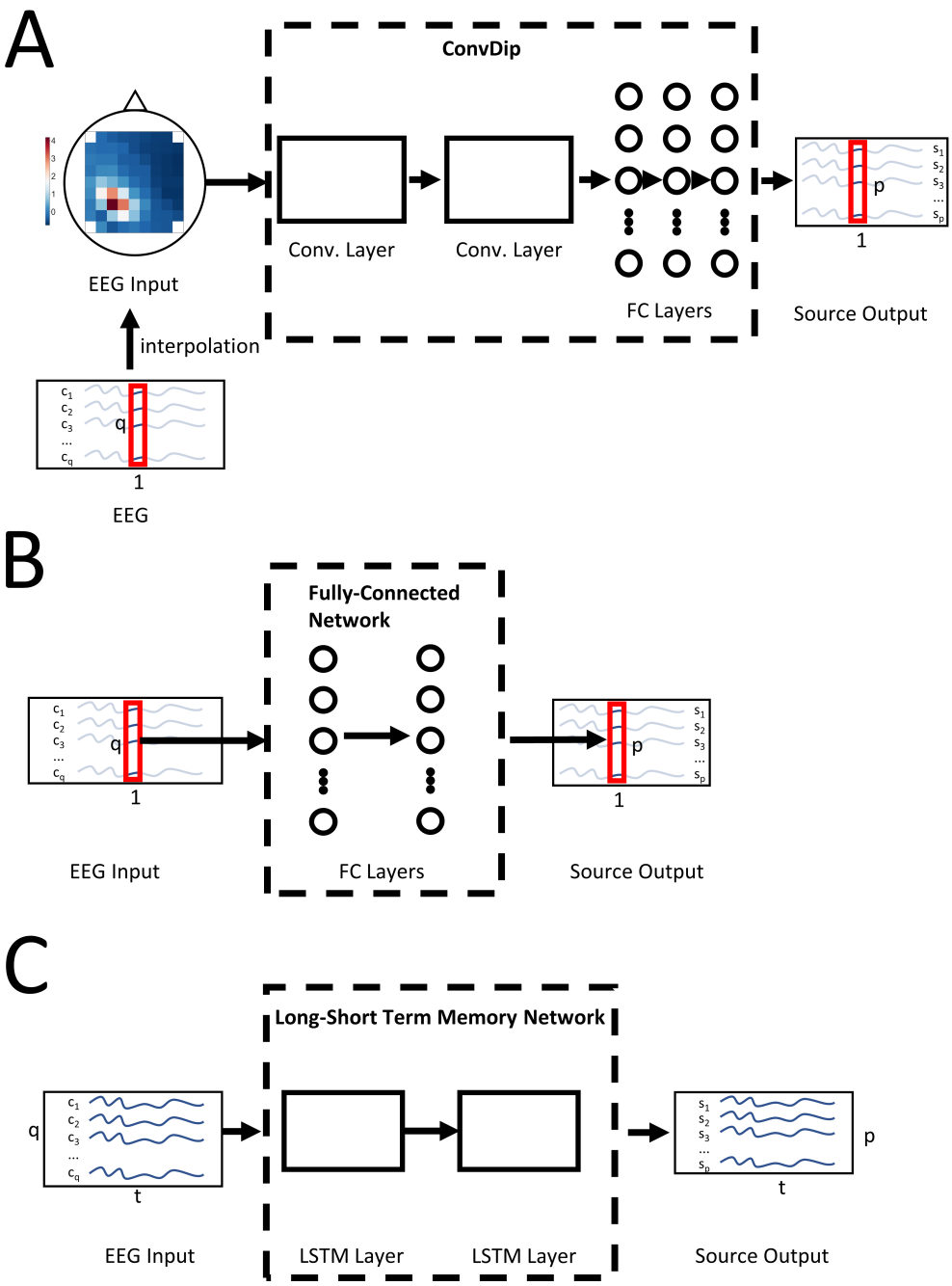

Figure 1: **Neural network architectures.** A: ConvDip, B: The fully-connected (FC) network. C: The LSTM network. q: number of electrodes, p: number of dipoles, t: time points, c_i: channel at index i, s_i dipole at index i

of capturing diverse temporal patterns in the data. The forward-pass of an LSTM network is described in detail in Appendix A.

The input of the LSTM network is a 2D EEG matrix of shape channels $\times$ time points, whereby the number of channels (q) is fixed and the number of time points (t) is arbitrary. The output of the network $Y \in \mathbb{R}^{p \times q}$ with p dipoles and t time points.

In an effort to facilitate the incorporation of the whole temporal context we have used an implementation of the bidirectional LSTM, which can not only utilize past time points but also subsequent time points (Schuster & Paliwal, 1997). Note, that this disqualifies the model for real-time neurofeedback application. Our LSTM architecture consisted of two bidirectional LSTM layers à 170 units (85 per direction) each (Fig. 1, C). Dropout was used at each hidden layer with a drop rate of 20%. The tanh and sigmoid activation functions were used for standard and recurrent activation, respectively. The output layer consisted of a fully-connected layer just like the two previous architectures, through which each time step of the previous layer was fed to yield the sequential output. The total number of parameters was $493,604$.

3.3.4 Training Procedure

The resulting normalized simulated samples were split into a training (90%) and validation (10%) set randomly in order to monitor convergence and prevent overfitting. A maximum of 150 epochs of training was set. Early stopping of the training procedure was done after two epochs without reduction of the validation loss. The loss was calculated by *cosine-similarity* between predicted and target source vectors, a distance metric that penalizes errors in source patterns but not in absolute source amplitudes. As an optimizer we chose adaptive momentum estimation (ADAM, Kingma & Ba, 2014) with a learning rate of 0.001. In order to avoid the problem of exploding gradients we clipped all gradients to a norm of 1. The batch size was set to 8.

Due to the scaling of the training data and the scale-agnostic loss function, it follows that the ANNs will predict sources that are not in the original scaling of [1,10] nano ampere meter (nAm). To rescale the sources, we simply rescaled the predicted source such that the resulting predicted EEG was of equal GFP as the original input. See Appendix A for more details.

3.4 Evaluation

In order to assess the quality of inverse solutions we created an additional set of simulations henceforth referred to as *evaluation set*. The simulation parameters were kept equal to those used for the training set, however just 1,000 samples were simulated to reduce computation time. We then predicted the sources of each of these samples using the three ANN-based inverse solutions and the three methods of comparison: eLORETA, MNE and LCMV Beamformer. In order to assess the quality of these inverse solutions we decided for the following metrics:

- **Area under the ROC curve.** We calculated the area under the receiver operating characteristic curve (ROC-AUC) as described by Grova et al. (2006) and in our earlier work (Hecker et al., 2020). AUC_{far} is a metric that captures how well an inverse solution (1) finds the sources at the correct locations and (2) avoids false positives. AUC_{close} also captures (1) how well sources are correctly localized but also (3) how well the size was estimated. Therefore, AUC_{close} captures the dispersion of an inverse solution.

- **Mean localization error.** The Euclidean distance between the locations of the predicted source maximum and the target source maximum is a common measure of inverse solution accuracy as it captures the ability to localize the source center accurately. MLE was calculated between the positions of the maxima in the predicted source ($\hat{j}$) and the ground truth source (j). This metric is only suitable for calculating MLE when a single source patch is present.

  For multiple sources, we adapted the following procedure. First, we identified local maxima in both the true source vector j and the estimated source vector $\hat{j}$. A voxel constituted a maximum if its value was larger than all of its neighboring voxels. This yielded many local maxima, which had to be filtered in order to achieve reasonable results. Filtering involved the elimination of all maxima whose neighbors were not sufficiently active ($< 10\%$ of the maximum). This takes care of false positive maxima that do not constitute an active cluster of voxels. Then, we removed these maxima that had a larger maximum voxel within a radius of 30 mm. These procedures result in a list of coordinates of maxima for both j and $\hat{j}$. We then calculated the pairwise Euclidean distances between the two lists of coordinates of maxima. For each true source, we identified the closest estimated source and calculated the MLE by averaging these minimum distances. We further labelled those estimated sources that were $\geq 30mm$ away from the next true maximum as ghost sources. True maxima that did have an estimated source within a radius of 30 mm were labelled as found sources, whereas those true maxima that did not have an estimated maximum within a radius of 30 mm were labelled as missed sources. Finally, we calculated the percentage of found sources, i.e., the ratio of the number of correctly identified sources and the number of all true sources.

- **Normalized Mean Squared Error.** Calculation of the normalized MSE (nMSE) was done by normalizing both the true source vector j and the predicted source vector $\hat{j}$ before calculating the MSE. By normalizing the vectors, we get rid of possible offsets and capture the overall differences in patterns more closely.

$$nMSE = \frac{1}{p}\sum_{i=1}^{p}(j_i - \hat{j}_i)^2 \tag{2}$$

3.5 Conventional Inverse Solutions

We used the implementation provided by the mne-python package (Gramfort et al., 2013, 2014) to calculate the eLORETA, MNE and LCMV inverse solutions with default parameters.

Inverse solutions were calculated for each sample individually. Noise covariance matrices were calculated based on a noise baseline segment which we added prior to each simulated sample (i.e., trial). This segment was generated using subsegments of noise randomly drawn from the noise which was added to the EEG. For LCMV, we furthermore calculated the data covariance matrix for the whole segment excluding the baseline. LCMV inverse solutions were calculated based on a forward model with free dipole orientations since this yielded better results during our testing phase.

3.6 Application to Real Data

In order to evaluate the performance of the inverse algorithms with real data we decided to calculate the sources of the somatosensory response in the brainstorm (Tadel, Baillet, Mosher, Pantazis, & Leahy, 2011) median nerve stimulation (MNS) data set[2]. The data contains MEG data of a single participant receiving 200 pulses of MNS on the left and right hand. MEG was recorded with a CTF 275 system at the Montreal Neurological Institute by Bock, Florin, Tadel & Baillet. The data set was made public by brainstorm, and we have received permission for its usage in the present publication.

For simplicity, we selected only trials in which the left median nerve was stimulated and removed trials containing artifacts. This left us with 51 clean trials of MEG data. We then calculated the event-related fields (ERF) by averaging and applied baseline correction by subtracting the averaged signal during the 100 ms prior to stimulus onset.

All mentioned ANNs were re-built for the MEG sensor layout and the resulting forward model and had to be retrained with a new simulation set of 10,000 trials. The architectures and simulation parameters were kept the same as described above.

4 Results

4.1 Comparison among ANNs

As a first step we will evaluate the effect of the three ANN architectures on the evaluation set performance. This is assessed using the original loss function *cosine distance*. Fig. 2 depicts the cosine distance for each ANN architecture. Each model architecture was newly created with different initial weights and retrained ten times. Cosine error was calculated for each time step in each sample and averaged. This yields a distribution of errors per architecture. One-way ANOVA revealed significant effects of model architecture on cosine distance ($F(2, 27))1425.76, p = 4.21 \cdot 10^{-28}, \eta_p^2 = 0.99$).

[2]See https://neuroimage.usc.edu/brainstorm/DatasetMedianNerveCtf for a description

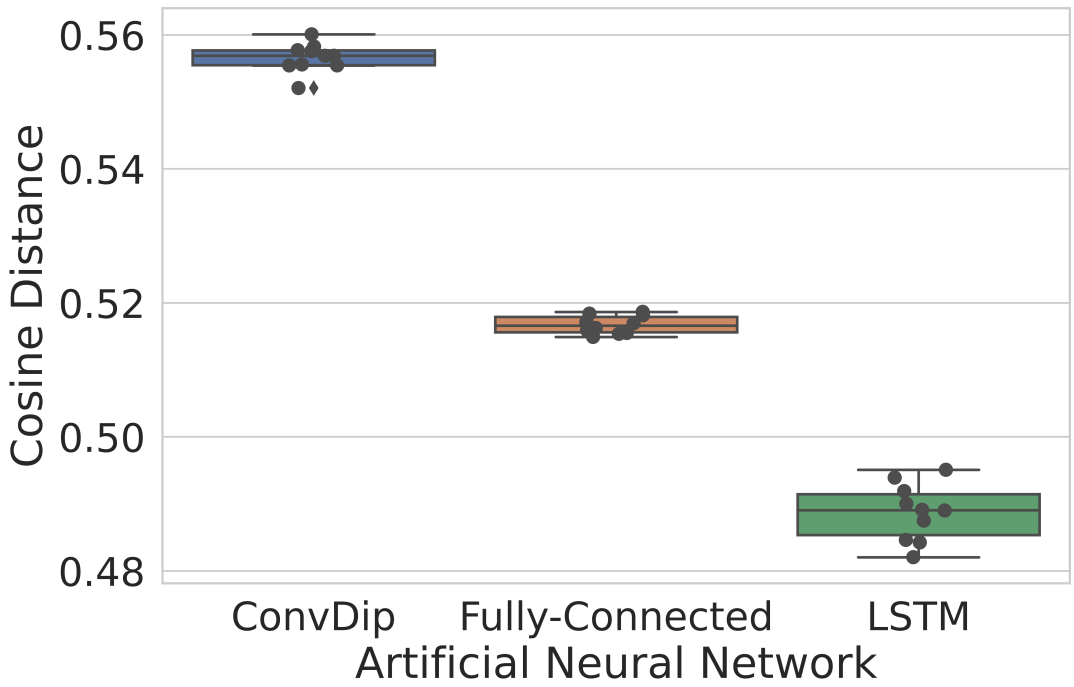

Figure 2: **Cosine Distance for Multiple Models on the Evaluation Set** The performance of all ANN-based inverse solutions is depicted. The cosine distance was used as a loss function during training and thus reflects the training success and degree of generalization on the evaluation set. Ten models were newly created and trained with the same data set for each architecture in order to show the variance caused by random weight initialization. The cosine distance spans from perfectly correlated ($d_cos = 0$) over orthogonal ($d_cos = 1$) and anti-correlated ($d_cos = 2$).

A large leap in performance from the original ConvDip to the Fully-Connected architecture with an error reduction of 7.71% that was highly significant based on the Games-Howell Post-Hoc test ($p < 10^{-16}, d = 22.27$). The change from the Fully-Connected to the LSTM architecture results in a further error reduction of 5.72% ($p = 2.16 \cdot 10^{-9}, d = 8.88$). Notably, the distributions in Fig. 2 do not overlap, which emphasizes the importance of the individual architectures on model performance.

4.2 Utilization of Temporal Context

We investigated the ability of the LSTM to exploit temporal information in the EEG to produce improved inverse solutions. For a set of 10,000 simulated samples containing 200 time points each, we iteratively removed time points, i.e., information, for the LSTM to incor-

porate into the prediction. In each iteration the prediction performance was re-calculated for the final sample, which yielded an average performance as a function of the amount of temporal context. Results are shown in Fig. 3. By definition, the fully-connected ANN and ConvDip are not designed to utilize temporal information which is reflected in the invariant performance over time. The LSTM quickly outperforms both the FC network and ConvDip with only little temporal information. The LSTM network clearly shows improved predictive performance with increased temporal context. All performance metrics displayed show the highest improvement between 0 and 20 time points preceding the target sample. Context of more than 25 time points does not appear to further decrease *mean localization error* (MLE). However, *normalized mean squared error* (nMSE) and *area under the receiver operating characteristic curve* (AUC), measures of global (dis-)similarity, seem to improve steadily with increased temporal context, albeit with diminishing returns after about 25 time points.

4.3 General Performance Evaluation

Figure 4 depicts exemplary inverse solutions to an EEG containing activity from a single source patch. By visual inspection it is apparent that the ANN-based inverse solutions (top row) reproduce the sparsity of the ground truth with little blur, whereas the classical inverse solutions (bottom row) produce higher blur. This observation is reflected in global (dis-)similarity metrics, namely the nMSE and the AUC for both single sources (see Fig. 5 & Tab. 2) and multiple sources (Fig. 6 & Tab. 3). Furthermore, the LSTM network shows the best performance in nMSE and AUC compared to the FC network and ConvDip, which probably stems from the usage of temporal context as presented earlier (Fig. 3).

The margin of the ANN-based inverse solutions is quite remarkable, especially regarding nMSE. This is in part a consequence of the similarity between the training data and the evaluation data as they were constructed using the same simulation parameters (Tab. 1). The high performance of the ANN-based inverse solutions is thus a sign that the priors which we introduced via the training data set were learned successfully. Particularly, we want to highlight the very high AUC (98.66%) for the LSTM in the case of single source patches. The implications will be critically discussed in the Discussion section.

Among the ANN-based inverse solutions, the MLE was lowest for the LSTM network. The analysis shown in Fig. 3 **A** provides evidence that this advantage is at least in part due to the incorporation of temporal information in the LSTM. Interestingly, ConvDip shows a slightly higher MLE compared to the fully-connected network. eLORETA shows the lowest median MLE when single sources are present. The mean localization error of eLORETA inverse solutions were, however, not significantly lower compared to those produced by the LSTM network ($t = 1.19, p = 0.23, d = 0.18$). This is partially in concordance with our previous publication Hecker et al., 2020, in which eLORETA already showed competitive localization errors in the case of single source patches. This admittedly small advantage, however, vanishes for simulations that contain various numbers of source patches. In this case, the LSTM network shows the lowest localization error (Fig. 6 **A**).

A statistical analysis of these presented effects is depicted in Tab. 4. It becomes ap-

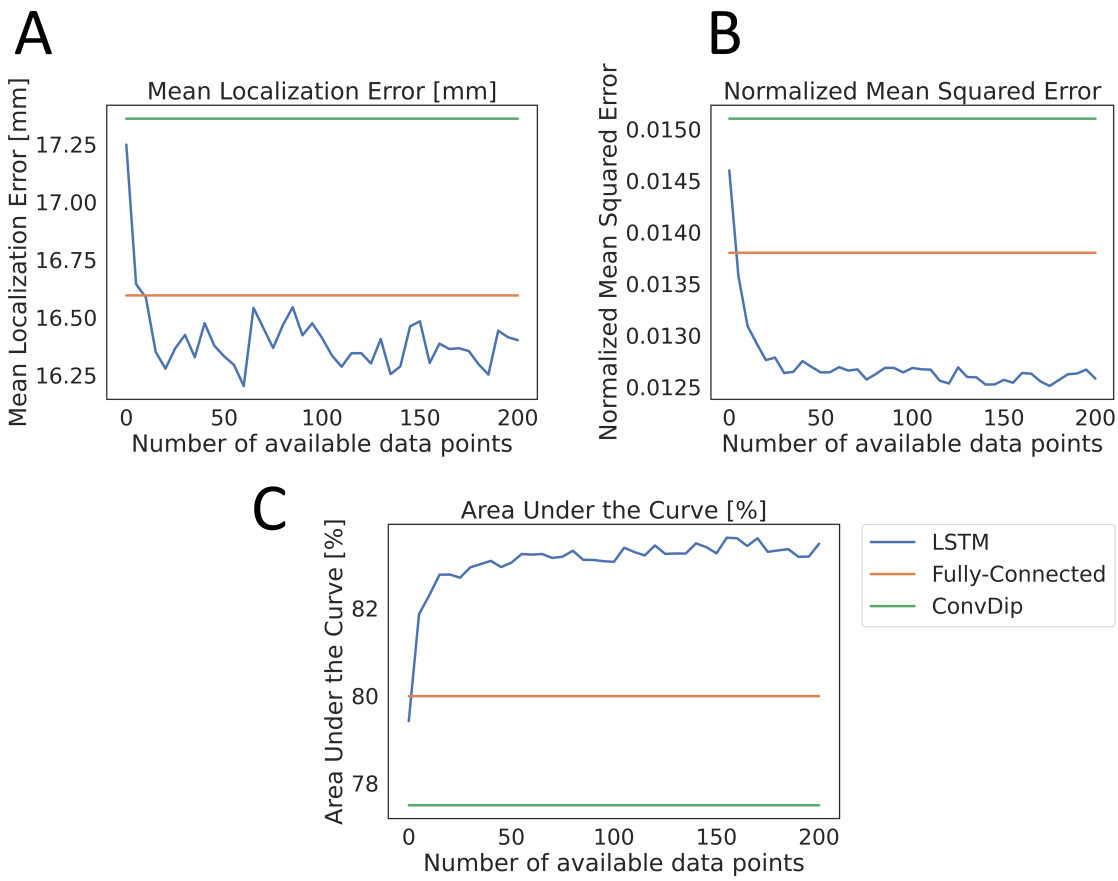

Figure 3: **Temporal Information and Performance** The performance of ANN-based inverse solutions as a function of temporal context. The performance metrics depicted are (A) mean localization error (MLE), (B) normalized mean squared error (nMSE) and (C) area under the receiver operator curve (AUC). While the fully-connected ANN and ConvDip are not able to process temporal information the long-short term memory (LSTM) network clearly shows improved performance with increased temporal information.

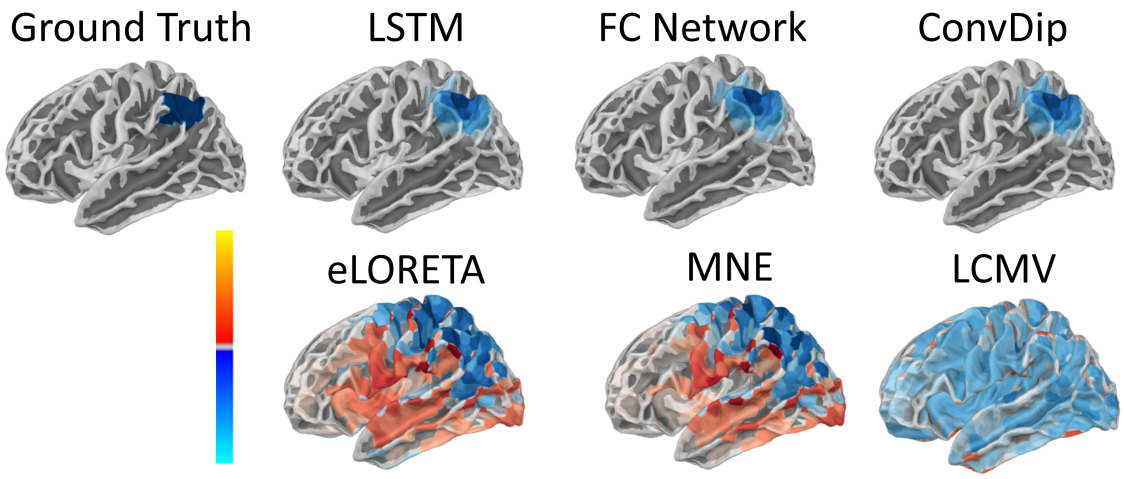

Figure 4: **Inverse Solutions of Simulated EEG Containing a Single Source Patch** Ground truth and predictions of various inverse solutions. Coloring of brain plots is equal between inverse solutions. The lowest absolute 5% of activity are omitted through transparency. The ANN-based inverse solutions (top row) correctly estimated the location and size of the source patch, whereas the classical approaches (bottom row) produced more blurry predictions. LSTM: Long-short term memory (LSTM) network; FC: Fully-connected network

Method	MLE (SD) [mm]	AUC (SD) [%]	nMSE (SD)
LSTM	11.95 (4.85)	**98.66 (5.17)**	**0.0073 (0.0029)**
Fully-Connected	13.27 (5.31)	95.94 (8.1)	0.0105 (0.0041)
ConvDip	14.44 (5.23)	95.73 (6.26)	0.0107 (0.0040)
eLORETA	**11.04 (8.06)**	88.17 (15.04)	0.0495 (0.0634)
MNE	21.97 (7.64)	82.09 (15.98)	0.0407 (0.0251)
LCMV	17.76 (10.2)	64.03 (21.89)	0.0285 (0.0217)

Table 2: **Evaluation Metrics for Single Sources** Medians and standard deviations (SD) of the mean localization error (MLE), area under the ROC curve (AUC) and normalized mean squared error (nMSE). The best performing inverse solutions are marked in bold font.

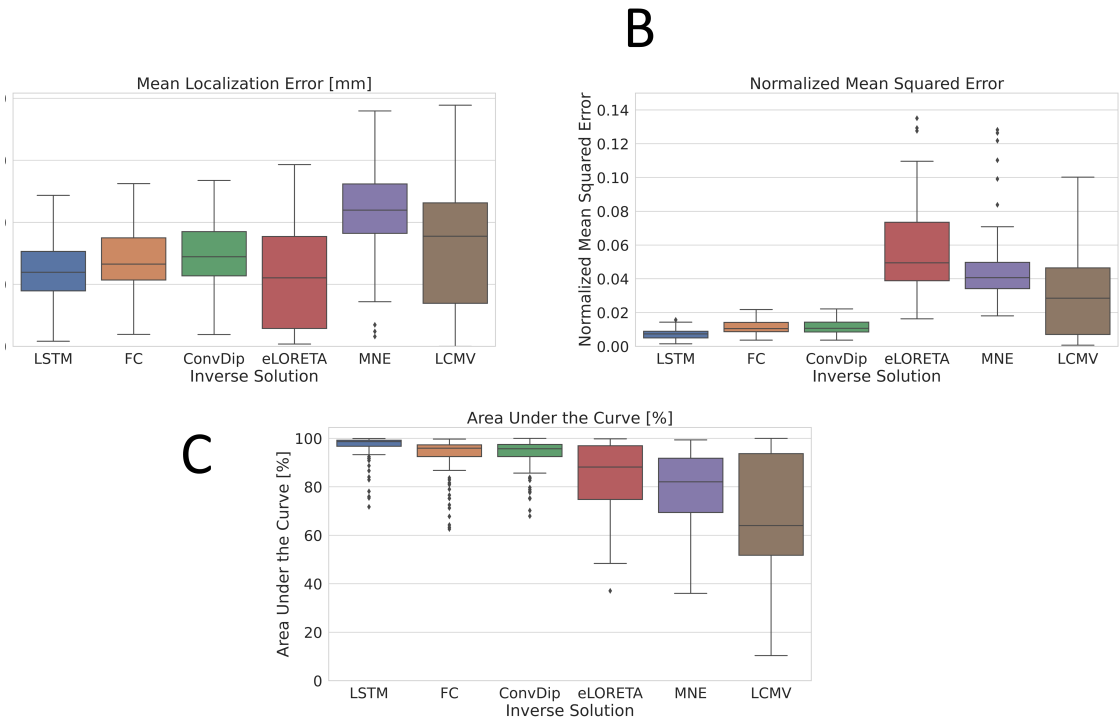

Figure 5: **Performance with Single Source Patches** (A) Mean localization error (MLE) between true and predicted sources in mm. (B) Normalized mean squared error (nMSE). (C) Area under the curve (AUC) in %. The LSTM produced the lowest errors compared to all other inverse solutions, except for the slightly lower MLE produced by eLORETA.

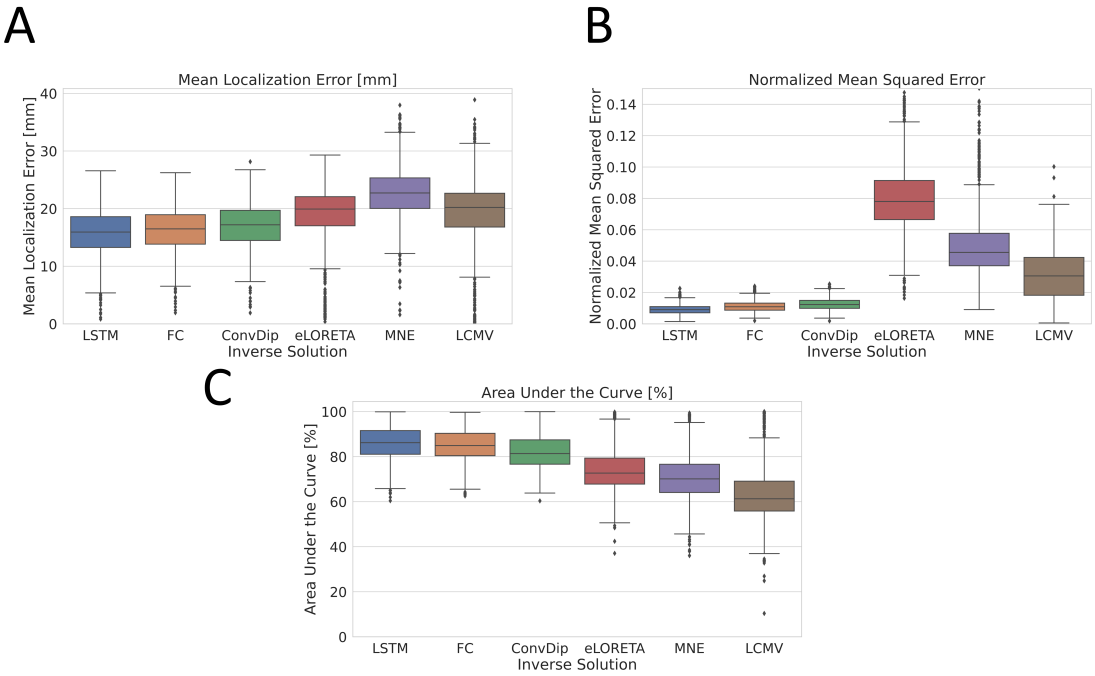

Figure 6: **Performance with Multiple Source Patches** (A) Mean localization error (MLE) in mm between true and predicted sources. (B) Normalized mean squared error (nMSE). (C) Area under the curve (AUC) in %

Method	MLE (SD) [mm]	AUC (SD) [%]	nMSE (SD)
LSTM	**15.94 (4.00)**	**86.23 (7.6)**	**0.0091 (0.0031)**
Fully-Connected	16.50 (3.74)	84.93 (7.09)	0.0109 (0.0035)
ConvDip	17.21 (3.82)	81.39 (7.83)	0.0124 (0.0037)
eLORETA	19.94 (5.35)	72.72 (9.80)	0.0781 (0.054)
MNE	22.73 (4.54)	70.15 (10.40)	0.0456 (0.0216)
LCMV	20.22 (6.00)	61.31 (12.56)	0.0306 (0.0165)

Table 3: **Evaluation Metrics for Multiple Sources** Medians and standard deviations (SD) of the mean localization error (MLE), area under the ROC curve (AUC) and normalized mean squared error (nMSE). The best performing inverse solutions are marked in bold font.

parent that the LSTM has a small leap over the Fully-Connected network with respect to Mean Localization Error and AUC, whereas an improvement in nMSE was medium-sized and highly significant. Compared to the classical approaches, the LSTM consistently brings significant improvements reflected by medium (0.64) to very high effect sizes (2.68). Strongest effects were found for the global (dis-)similarity metrics, indicating that improvements stem mostly from the correct estimation of the size of sources than the position of the center of the source patches.

4.4 Performance and Signal-to-Noise Ratio

One of the differences between the ANN-based approaches and classical approaches is that the latter use covariance matrices to estimate the noise in the EEG. Ultimately, this leads to more regularized inverse solutions that have a reduced risk to mistake noise for signal. In theory, this can be a considerable advantage of classical approaches over for ANN-based inverse solutions that don't incorporate temporal aspects of the EEG data.

We analyzed the influence of SNR on the inverse solution quality using the evaluation set introduced above. Results are depicted in Fig. 8.

Across source analysis methods and evaluation measures we found a non-surprising overall pattern of better performance with increasing SNR. Interestingly, however, the performance difference between source analysis measures is larger than the difference within the individual methods across SNR values. We also found a clear advantage of the ANN-based inverse solutions compared to the classical approaches to the inverse problems (particularly LCMV Beamformer and eLORETA) across all evaluation metrics (Fig. 8 A, B & C). Focusing only on these classical approaches performances of the different methods vary across evaluation measures. For example, MNE performs worst on the MLE measure (Fig. 8 A), LCVM performs worst on the AUC measure (Fig. 8 B) and eLORETA performs worst on the nMSE measure (Fig. 8 C).

The ANN based solutions provide a more coherent picture. LSTM shows the best across the four evaluation measures. Further, the difference between LSTM and the other two ANN solutions becomes consistently smaller, the larger the SNR is. Thus, LSTM seems

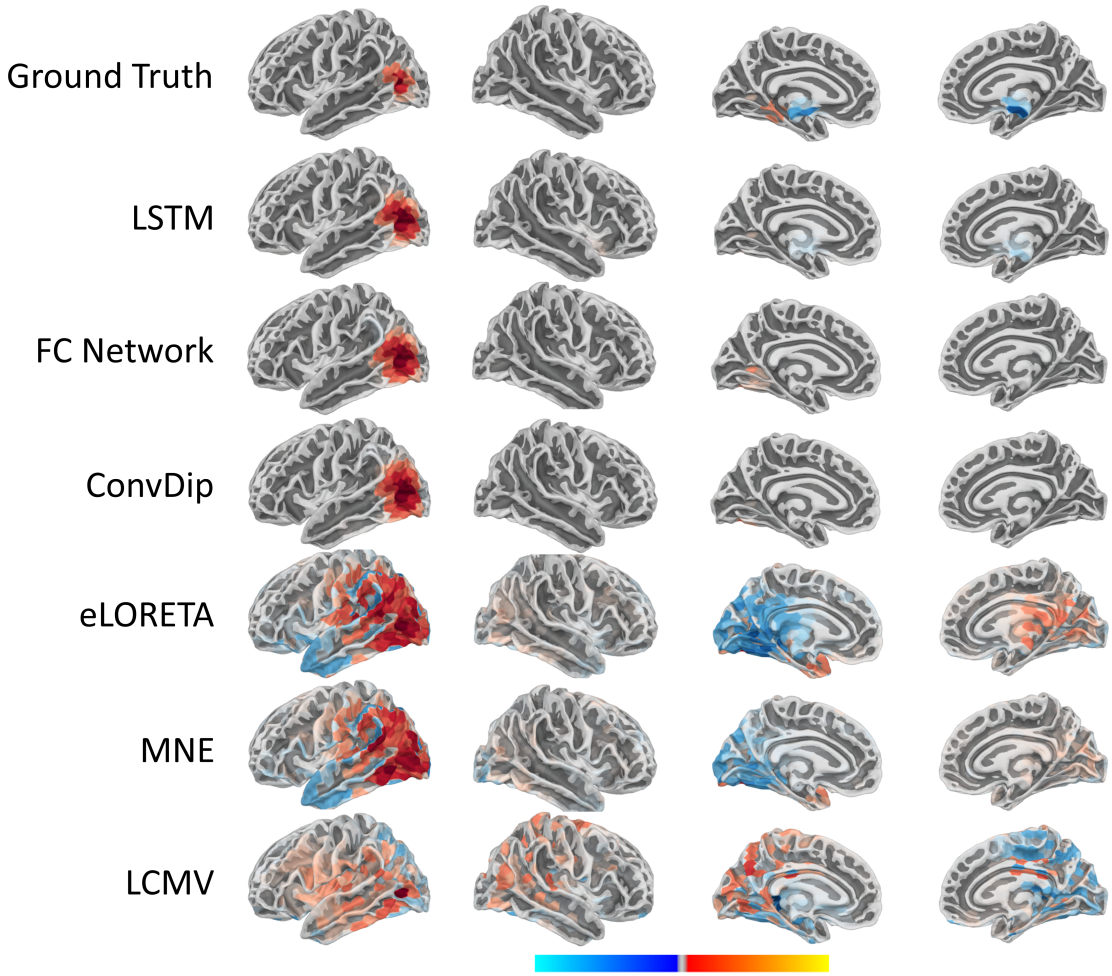

Figure 7: **Inverse Solutions of Simulated EEG Containing Three Source Patches** Ground truth and predictions of various inverse solutions. The lowest absolute 5% of activity are omitted through transparency.

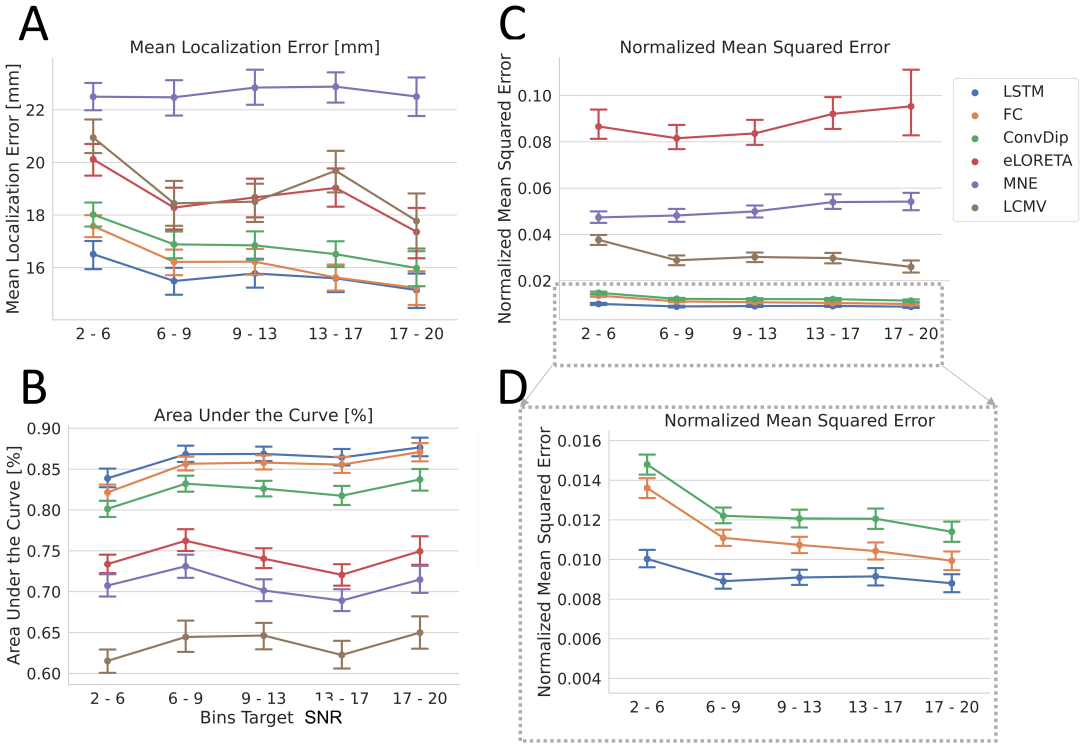

Figure 8: **Performance for different signal-to-noise ratios (SNR)** (A) Mean localization error (MLE) in mm, (B) Area under the curve (AUC) in %, (C) Normalized mean squared error (nMSE), (D) Zoom of nMSE to visualize the errors of ANN-based inverse solutions.

Contrast	Mean Localization Error			Area Under the Curve			Normalized Mean Squared Error		
LSTM vs.	T	p	d	T	p	d	T	p	d
FC	-2.83	5.4e-02	-0.13	3.37	1.0e-02	0.15	-13.72	3.1e-13	-0.61
ConvDip	-6.61	7.4e-10	-0.30	11.72	0.0e+00	0.52	-21.93	1.3e-12	-0.98
eLORETA	-14.30	1.8e-13	-0.64	31.02	2.5e-13	1.39	-45.78	3.8e-13	-2.05
MNE	-36.08	0.0e+00	-1.61	37.83	2.0e-13	1.69	-59.89	0.0e+00	-2.68
LCMV	-14.87	6.3e-13	-0.67	48.97	4.0e-13	2.19	-40.47	0.0e+00	-1.81

Table 4: **Statistics between LSTM and other Source Imaging Methods** Post-Hoc tests result using the Games-Howell test. Positive T-values and Cohens d values indicate larger values for the LSTM compared to the method of contrast.

to have the highest advantage in cases where the EEG has low SNR, indicating that it uses temporal information for noise regularization.

4.5 Validation on Real Data

Our analyses have shown that the ANN-based methods produce reasonable inverse solutions given simulated EEG data. However, it remains unclear whether this translates to real M/EEG data from humans. Furthermore, we have not yet presented evidence that the ANN-based approaches also work analogously with MEG data. We have therefore calculated inverse solutions to the brainstorm median nerve stimulation (MNS) data set of one single participant.

The participants' average response to 51 stimulations of the left hand is depicted in Fig. 9 A. The peak latency of interest was defined at 0.03s after stimulation onset. The topographic MEG distribution at that latency is depicted in Fig. 9 B and inverse solutions at that time point are depicted in Fig. 9 C.

It is apparent that all inverse algorithms find a very similar spatial pattern overall which is in concordance with the expected left-hemispheric primary somatosensory response. Upon closer inspection, slightly higher sparsity of the shown ANN-based inverse algorithms (top row) becomes visible. In order to compare the various inverse solutions we have calculated the position of the maximum voxels and found that all inverse solutions except for ConvDip find the maximum activity at the same location ($x = 41.69, y = -31.89, z = 55.27$). ConvDip finds a maximum voxel 9.7mm medial to the other inverse solutions ($x = 33.13, y = -35.74, z = 53.00$).

We furthermore analyzed the similarity between inverse solutions with respect to their spatial pattern using the AUC metric in the following way: One inverse solution serves as ground truth on which the other inverse solution is measured against. Note that this method is not commutable and interchanging ground truth method and test method can reveal different AUC values. For example, given eLORETA served as ground truth, the LSTM would match it with an AUC of 78%. On the contrary, if the LSTM served as ground truth, eLORETA would match it with an AUC of 96%. The discrepancy between the two outcomes can be explained simply by the difference in sparsity of the two inverse solutions.

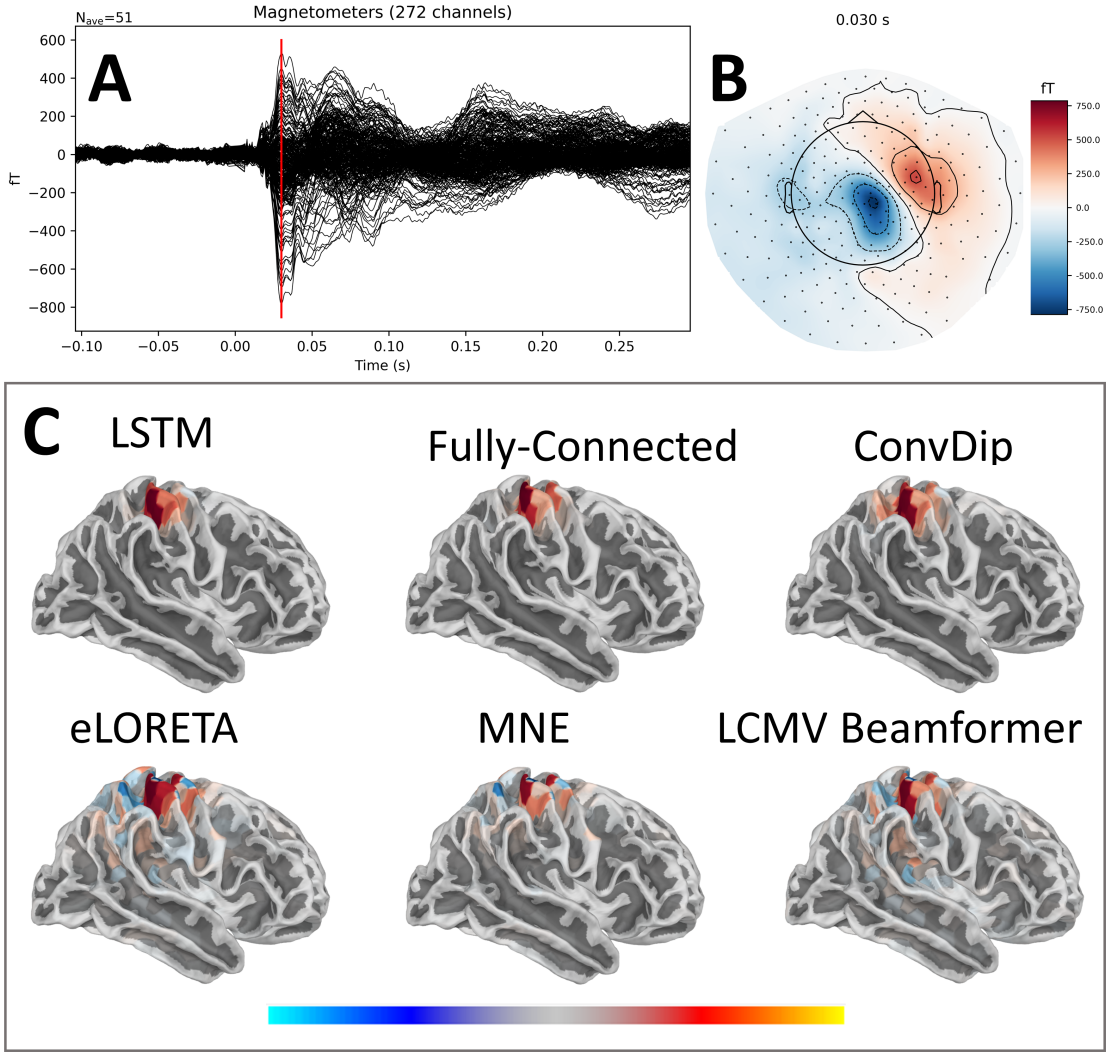

Figure 9: **Median Nerve Stimulation MEG Dataset** (A) shows the MEG ERF for each channel. The red vertical line marks the time point of interest at 0.03s after stimulus onset. (B) The topographic distribution of magnetic fields at the time point of interest. (C) Various inverse solutions at the time point of interest. Top row: ANN-based inverse solutions. Bottom row: conventional inverse solutions.

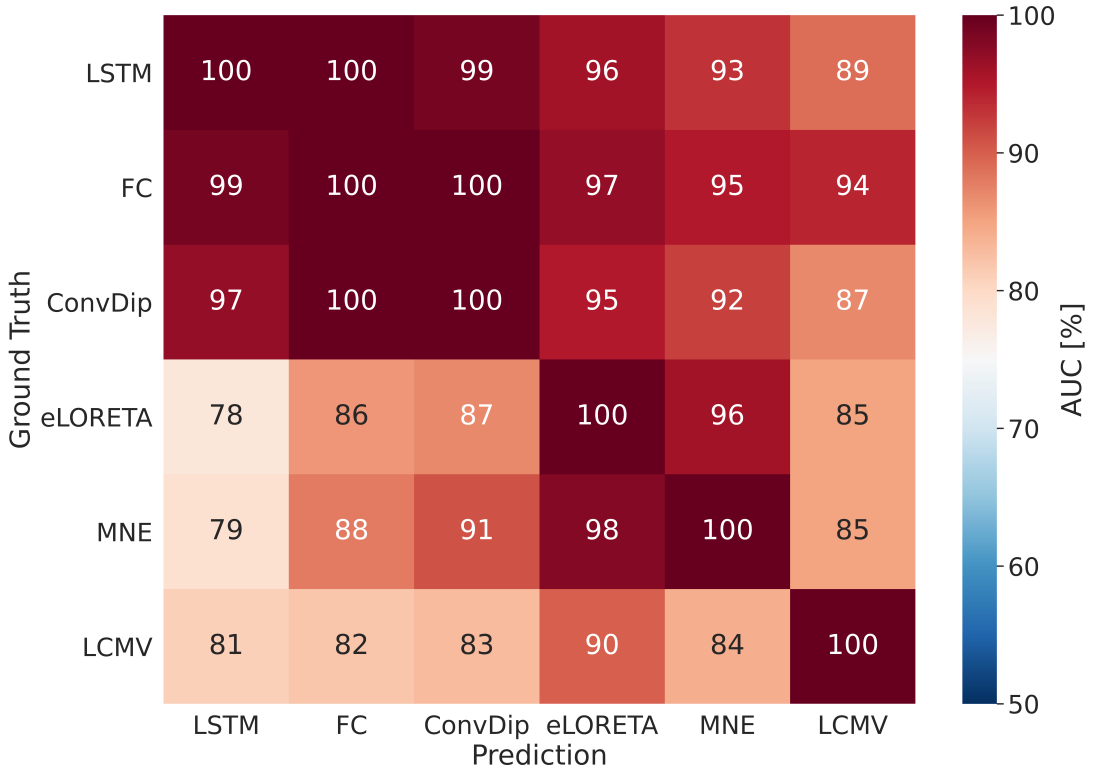

Figure 10: **Similarity Analysis of Inverse Solutions for the MNS data set.** The matrix reports pairwise AUC between each pair of inverse solutions to the MNS data set. AUC is a measure of similarity between a ground truth (rows) and a prediction (columns). Note, that the asymmetry in the AUC analysis design results in non-commutativity of the AUC calculations and a matrix with no symmetry along the first diagonal.

The results of this similarity analysis are shown in Fig. 10 and indicate that ANN-based inverse solutions have high concordance (97% - 100%, Fig. 10, top-left quadrant) among each other. Similarly, there is also relatively high concordance of 83% - 98% (Fig. 10, bottom-right quadrant) between the classical inverse solutions.

The concordance between ANN-based inverse solutions and the classical approaches is lower ($78\% - 96\%$, bottom-left and top-right quadrants). Beyond the tiny differences between methods, it is reassuring that there is generally a high similarity between all tested inverse solutions. Particularly, the sources estimated by ANN-based methods are covered almost entirely ($> 87\%$, top right quadrant) by classical inverse solutions.

5 Discussion

The EEG and MEG are widely used techniques to measure neurophysiological signals *in vivo*. Increasing their spatial resolution by solving the inverse problem, i.e. finding the neural sources given the signals measured on the scalp, is of major interest both for basic research and in the clinical context (e.g., localization of seizure onset zones in epilepsy). ANN-based methods are gaining attention in the past years and appear a viable option to solve the inverse problem of the M/EEG (Cui et al., 2019; Dinh et al., 2019; Hecker et al., 2020; Huang et al., 2020; Pantazis & Adler, 2021; Sun et al., 2020; Tankelevich, 2019).

In order to bring ANN-based inverse solutions out of the infancy state there are some issues that need to be tackled. Each of the existing ANN-based inverse solvers has at least one of the following shortcomings: (1) Large number of parameters, (2) physiologically implausible source outputs dimensions (i.e., single dipole position/s), (3) inability to exploit temporal information flexibly. In the present study we introduced an LSTM network which offers to solve these problems. The LSTM network was evaluated alongside two other ANN architectures and three classical source imaging methods using simulated source and EEG data. Furthermore, we applied all methods to an MEG data set containing event-related fields to median nerve stimulation.

We have demonstrated that the LSTM network was able to exploit temporal information. Most improvement in inverse solution accuracy was reached when about 25 time points were incorporated. More samples did not lead to lower MLE, but continuously improved nMSE and AUC, albeit with diminishing returns. It is possible that our proposed architecture does not have enough capacity (i.e., number of parameters) to utilize the temporal information to full extent beyond these 25 time points. This raises the general question about the optimal length of a time window that can be fed into the network. The general goal of source analyses is to describe neural sources responsible for certain brain states/mental states at a certain moment in time. A number of studies have provided evidence that brain states and/or mental states have a temporal expansion, which is in the range of tens of milliseconds (Atmanspacher, Bach, Filk, & Kornmeier, 2008; Atmanspacher & Filk, 2013; Koenig et al., 2002; Kornmeier & Bach, 2004, 2005, 2012; Kornmeier, Bach, & Atmanspacher, 2004; Kornmeier et al., 2019; Kornmeier, Friedel, Wittmann, & Atmanspacher, 2017; Wernery et al., 2015). Time windows in this range may serve as a reference for future studies with a specific focus on the optimal time window length for source analysis.

A rather practical aspect is, to keep the number of parameters low such that ANNs can be trained on typical hardware available to researchers and clinicians. On the other hand, there are cases in which larger computational efforts are justified, e.g., during presurgical procedures in epilepsy. In this case it may be reasonable to build ANNs with a larger number of parameters. Future research should use automated machine learning (Auto-ML) and investigate optimal ANN architectures with varying size restrictions (Hutter, Kotthoff, & Vanschoren, 2019).

In comparison to classical inverse solutions, we found that ANN-based inverse solutions yielded clearly better average predictions in almost every case we tested. The exception was

a scenario with a single source. In this case eLORETA yielded comparable or even lower MLE as the ANN methods, which confirms earlier findings (Hecker et al., 2020). Overall, best performance was seen for the LSTM network, which is most probably a direct result of the incorporation of temporal information. Compared to ConvDip, we found a 6 % improvement in AUC, 39 % in nMSE and a 9 % improvement in MLE. This improvement can be at least in part attributed to the LSTMs ability to include temporal information into its prediction, rendering it a more powerful tool for electrical source imaging compared to ConvDip.

The results depicted in Fig. 5 & 6 are a confirmation that the priors, which were introduced in the training set (cf. Tab. 1), were learned successfully by all ANNs. It is to no surprise that classical approaches are at a disadvantage when confronted with comparatively sparse sources. From our perspective, the root of this problem is the lack of ground truth data. There is evidence that brain-electric activity is to some degree coherent in space and time (Destexhe et al., 1999; Leopold & Logothetis, 1999; Spiridon, Fischl, & Kanwisher, 2006) and that a cortex area of about $\approx 2.54cm^2$ must be synchronously active in order to produce a signal that can be measured on the scalp using EEG (Nunez & Srinivasan, 2006). These findings question the theoretical applicability of single-dipole models. However, a lot about the spatial distribution of brain activity remains unknown. Thus, any approximation of a potential ground truth depends heavily on our knowledge about brain function, which is to considerable amount based on non-invasive measurements with all their limitations.

This poses a big problem for data-driven inverse solutions: What kind of data should be used to train the ANN models? How many sources of what shape and size do we need to simulate? Until we have an answer to these questions, we should simulate neural activity as broadly as physiologically plausible, wherefore we chose e.g., source diameters to be in a wide range between $5mm$ (size of a pea) and $4cm$ (size of a golf ball). Possibly, a review of the existing literature on the distribution of brain-electric activity is the next step necessary to qualify data-driven inverse solutions as the powerful tool they promise to be.

Ultimately, we have applied all methods to real MEG recordings during MNS with very promising results. Different inverse solvers identified largely similar patterns at the expected contralateral dorsal primary somatosensory cortex (Fig. 9, Sutherland & Tang, 2006). Their similarity across source analysis methods is a very positive sign that our proposed LSTM network trained on simulations works as intended with real data.

5.1 Limitations

Despite promising results obtained for the LSTM architecture, we find that the center of activity is oftentimes estimated more reliably using eLORETA, especially when only single source patches are present in the data (see the MLE analysis with single sources above). Possibly, future changes to the ANN-based inverse solutions should aim to develop a more suitable loss function which penalizes this shortcoming. One future option to improve out LSTM may be to replace the punctual voxel-by-voxel based loss functions like mean squared error and cosine similarity by metrics that take spatial relations between voxels

into consideration, e.g., the earth-mover distance (also known as the Wasserstein metric or the Kantorovich–Rubinstein metric, Kantorovich, 1960). This would especially help to foster sparse inverse solutions.

A further disadvantage of ANN-based neural networks is the comparatively long training time compared to classical solutions. Until computer hardware evolves to a level where everyone can afford to compute the ANNs themselves we should aim to find ways to shorten training times. One possible solution could be publicly available pre-trained models and data sets.

We jave gained some early evidence that the length of the time series entering the LSTM may be a critical parameter. The decision about the optimal length for ANNs with memory, like the present LSTM, needs to be based on plausible from the literature on brain states and their temporal extent. However, the LSTMs ability to handle a given data set length may also depend on the capacity of the LSTM. Optimizing these two parameters entails potential for further improvement of the ANNs performance.

In this work we conducted evaluation mainly on simulated data and only exemplarily on real data. A thorough evaluation on biological M/EEG data needs to be conducted in order to further increase confidence in ANN-based inverse solutions. The a priori hard-to-solve problem with the evaluation of real M/EEG recordings, however, is the missing ground truth. An intermediate step to approach this problem may be the use of EEG data from epilepsy patients with implanted electrodes undergoing brain stimulation (e.g., Mikulan et al., 2020), which was already successfully used to evaluate various inverse solutions (Pascarella et al., 2021).

In summary, we show that the ANN-based inverse solutions are in the vast majority of cases at an advantage over classical source analysis methods. Furthermore, we showed that the LSTM network makes use of temporal information, which leads to better predictions from noisy EEG and overall more accurate predictions compared to ANN solutions ignoring the time domain. This is particularly the case in situations with overall low SNRs. As a proof-of-principle, we finally demonstrated, that ANN-based inverse solutions, trained with simulated data, produce reasonable results when applied to real MEG data. Moreover, the remarkable similarity of the ANN-based results with the results from the classical approaches is further confirmation for this novel source analysis approach and renders it a viable choice, be it applications in epileptology, neurofeedback or ERP studies.

6 Data Availability

The simulations and neural network models that were presented in this work can be recreated using the esinet package for python. The open source code of the package can be found here: https://github.com/LukeTheHecker/esinet.

7 Acknowledgements

We would like to thank Nvidia for sponsoring this work with a Titan V graphical computing unit as part of their academic seeding program. Furthermore, we thank all the contributors of the brainstorm MNS data set and particularly Francois Tadel for the permission to use the data in this publication.

8 Appendix A

8.1 Forward Pass in ConvDip

The first layer of ConvDip is a convolutional layer. In order to prepare the EEG data $M \in \mathbb{R}^{t,q}$ at any given time point $\mathbf{m} \in \mathbb{R}^q$ ($q = 61$ electrodes) for the convolutional layer it was interpolated to a 9×9 2D image as commonly done for scalp map visualizations. Please refer to Fig. 1 for a visualization. The resulting interpolated EEG $\tilde{\mathbf{m}} \in \mathbb{R}^{9\times 9}$.

The first layer contained 8 filter kernels $F_i, i = \{1, ...8\}$ of size 3×3. The weights of these filter kernels are initialized randomly (Glorot & Bengio, 2010) and optimized during the training period. In the forward pass, the 2D-interpolated EEG data at a single time point $\tilde{\mathbf{m}}$ is convolved with each filter kernel F_i, resulting in 8 feature maps $G_i^1 \in \mathbb{R}^{7\times 7}$:

$$G_i = \tilde{\mathbf{m}} * F_i \tag{3}$$

The feature maps $G_i^1, i \in \{1, ...8\}$ are stacked to a tensor $\tilde{G^1} \in \mathbb{R}^{7\times 7\times 8}$. Then, a second convolutional layer was followed with 8 filter kernels of size 3×3. The resulting feature maps of the second convolutional layer $G_i^2 \in \mathbb{R}^{5\times 5}$ were then flattened to a vector $\tilde{\mathbf{g}} \in \mathbb{R}^{200}$ and fed forward to three fully-connected layers à 250 neurons. The forward pass through the fully-connected layers is defined as follows:

$$\begin{aligned} \mathbf{y}_1 &= ReLU(\tilde{\mathbf{g}} \cdot W_1^T + \mathbf{b}_1) \\ \mathbf{y}_2 &= ReLU(\mathbf{y}_1 \cdot W_2^T + \mathbf{b}_2) \\ \mathbf{y}_3 &= ReLU(\mathbf{y}_2 \cdot W_3^T + \mathbf{b}_3) \end{aligned} \tag{4}$$

where:

$ReLU()$ = Rectified Linear Unit (activation function)
$W \in \mathbb{R}^{q\times k}$ = Weight matrix
$\mathbf{b} \in \mathbb{R}^{1\times k}$ = Bias vector
$\mathbf{y} \in \mathbb{R}^{1\times k}$ = Output vector of hidden layers

The linear output layer consisted of $p = 1284$ neurons, which correspond to the dipole locations in the source model.

8.2 Forward Pass in the Fully Connected Neural Network

The Fully Connected ANN is similar to ConvDip without the need for the 2D interpolation of the EEG data. It therefore does not utilize any convolutional layer.

The full forward pass of the Fully Connected Neural Network is described as follows:

$$\begin{aligned} Y_1 &= ReLU(\mathbf{m} \cdot W_1^T + \mathbf{b}_1) \\ Y_2 &= ReLU(Y_1 \cdot W_2^T + \mathbf{b}_2) \end{aligned} \tag{5}$$

where:

$M \in \mathbb{R}^{t \times q}$ = EEG matrix
$\mathbf{m} \in \mathbb{R}^{q}$ = EEG input at single time point
$k = 300$ = Number of neurons in hidden layers
$q = 61$ = Number of EEG channels
$p = 1284$ = Number of dipoles in source model

The output layer is equivalent to the output layer of the ConvDip model (linear activation, 1284 neurons).

8.3 Forward Pass in the LSTM Network

Each LSTM cell contains three separate gates, namely the forget gate (f), the input gate (i) and an output gate (o) (Hochreiter & Schmidhuber, 1997). Each gate can be viewed as its own neural network with its weight matrices U and W. σ denotes the sigmoid function and tanh the hyperbolic tangent function. The candidate values for the new cell state are denoted in $\bar{\mathbf{c}}_\mathbf{t}$.

The output of each gate is defined as follows:

$$\begin{aligned}
\mathbf{f_t} &= \sigma(W_f \cdot \mathbf{x_t} + U_f \cdot \mathbf{h_{t-1}} + b_f) \\
\mathbf{i_t} &= \sigma(W_i \cdot \mathbf{x_t} + U_i \cdot \mathbf{h_{t-1}} + b_i) \\
\mathbf{o_t} &= \sigma(W_o \cdot \mathbf{x_t} + U_o \cdot \mathbf{h_{t-1}} + b_o) \\
\bar{\mathbf{c}}_\mathbf{t} &= \tanh(W_c \cdot \mathbf{x_t} + U_c \cdot \mathbf{h_{t-1}} + b_c)
\end{aligned} \tag{6}$$

Where X_t is the input vector at time point t and $\mathbf{h}_{t-1}$ is the cell's hidden state from the previous time point. W and U are weight matrices that are optimized during the training. Bias vectors are denoted by b.

The hidden state and cell state are updated as follows:

$$\begin{aligned}
\mathbf{c_t} &= \mathbf{f_t} \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \bar{\mathbf{c}}_\mathbf{t} \\
\mathbf{h_t} &= \mathbf{o_t} \circ \tanh(\mathbf{c_t})
\end{aligned} \tag{7}$$

The $\circ$ operator denotes the Hadamard (i.e., element-wise) product.

Depending on the design of the LSTM model different variables may serve as the output of the LSTM cell. In our case, the hidden state at each time point was used. For seq2one models it is common to use only the final hidden state as output. In special cases, the cell state can be used as output, too.

The output layer of the LSTM model is a single linear fully-connected layer, through which the temporal outputs of the LSTM layers are passed sequentially.

Note, that in the present work, bi-directional LSTM cells were used. A bi-directional LSTM cell consists of two LSTM cells, one of which receives the time steps in reverse orders. This makes it possible to capture both past and future time points and thus improve predictions.

8.4 Scaling of Predicted Sources

The scaling of the predicted sources is done such that the predicted EEG $\tilde{\hat{M}}$ has equal global field power (GFP) as the true EEG M. Therefore, first the predicted EEG is calculated using the leadfield.

$$\tilde{\hat{M}} = K\hat{\tilde{Y}} \tag{8}$$

where:

$K \in \mathbb{R}^{q \times p}$ = Leadfield
$\hat{\tilde{\mathbf{y}}} \in \mathbb{R}^{p}$ = Unscaled predicted source at single time point
$\hat{\tilde{\mathbf{m}}} \in \mathbb{R}^{q}$ = Unscaled predicted EEG at single time point

Then, the unscaled predicted source $\hat{\tilde{\mathbf{y}}}$ is scaled as follows:

$$\hat{\mathbf{y}} = \hat{\tilde{\mathbf{y}}} \cdot \frac{std(\mathbf{m})}{std(\hat{\tilde{\mathbf{m}}})} \tag{9}$$

where:

$\hat{\mathbf{y}} \in \mathbb{R}^{p}$ = Scaled predicted source
$\mathbf{m} \in \mathbb{R}^{q}$ = True EEG and single time point
$std()$ = Standard deviation

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Abeyratne, U. R., Zhang, G., & Saratchandran, P. (2001). EEG source localization: A comparative study of classical and neural network methods. *International journal of neural systems*, *11*(04), 349–359.

Amblard, C., Lapalme, E., & Lina, J.-M. (2004, March). Biomagnetic source detection by maximum entropy and graphical models. *IEEE Transactions on Biomedical Engineering*, *51*(3), 427–442. doi: 10.1109/TBME.2003.820999

Atmanspacher, H., Bach, M., Filk, T., & Kornmeier, J. (2008). Cognitive time scales in a Necker-Zeno model for bistable perception. *The Open Cybernetics & Systemics Journal*, *2*(1).

Atmanspacher, H., & Filk, T. (2013). The Necker–Zeno model for bistable perception. *Topics in cognitive science*, *5*(4), 800–817.

Awan, F. G., Saleem, O., & Kiran, A. (2019). Recent trends and advances in solving the inverse problem for EEG source localization. *Inverse Problems in Science and Engineering*, *27*(11), 1521–1536.

Chollet, F., et al. (2015). Keras.

Chowdhury, R. A., Lina, J. M., Kobayashi, E., & Grova, C. (2013, February). MEG Source Localization of Spatially Extended Generators of Epileptic Activity: Comparing Entropic and Hierarchical Bayesian Approaches. *PLOS ONE*, *8*(2), e55969. doi: 10.1371/journal.pone.0055969

Cui, S., Duan, L., Gong, B., Qiao, Y., Xu, F., Chen, J., & Wang, C. (2019). EEG source localization using spatio-temporal neural network. *China Communications*, *16*(7), 131–143.

Dale, A. M., & Sereno, M. I. (1993). Improved localizadon of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: A linear approach. *Journal of cognitive neuroscience*, *5*(2), 162–176.

Destexhe, A., Contreras, D., & Steriade, M. (1999, June). Spatiotemporal Analysis of Local Field Potentials and Unit Discharges in Cat Cerebral Cortex during Natural Wake and Sleep States. *The Journal of Neuroscience*, *19*(11), 4595–4608. doi: 10.1523/JNEUROSCI.19-11-04595.1999

Dinh, C., Samuelsson, J. G., Hunold, A., & Hämäläinen, M. S. (2019). Contextual Minimum-Norm Estimates (CMNE): A Deep Learning Method for Source Estimation in Neuronal Networks. , 14.

Fedorov, M., Koshev, N., & Dylov, D. V. (2020). Deep Learning for Non-invasive Cortical Potential Imaging. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings* (Vol. 12449, p. 45). Springer Nature.

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*(4), 272–284. doi: 10.1002/(sici)1097-0193(1999)8:4⟨272::aid-hbm10⟩3.0.co;2-4

Friston, K., Harrison, L., Daunizeau, J., Kiebel, S., Phillips, C., Trujillo-Barreto, N., ... Mattout, J. (2008). Multiple sparse priors for the M/EEG inverse problem. *NeuroImage*, *39*(3), 1104–1120.

Fuchs, M., Kastner, J., Wagner, M., Hawes, S., & Ebersole, J. S. (2002). A standardized boundary element method volume conductor model. *Clinical Neurophysiology*, *113*(5), 702–712.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256). JMLR Workshop and Conference Proceedings.

Godard, C., Matzen, K., & Uyttendaele, M. (2018). Deep Burst Denoising. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11219, pp. 560–577). Cham: Springer International Publishing. doi: 10.1007/978-3-030-01267-0_33

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Parkkonen, L. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, *7*, 267.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460.

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., ... Vanrumste, B. (2008). Review on solving the inverse problem in EEG source analysis. *Journal of neuroengineering and rehabilitation*, *5*(1), 25.

Grova, C., Daunizeau, J., Lina, J.-M., Bénar, C. G., Benali, H., & Gotman, J. (2006). Evaluation of EEG localization methods using realistic simulations of interictal spikes. *Neuroimage*, *29*(3), 734–753.

Hämäläinen, M. S., & Ilmoniemi, R. J. (1994). Interpreting magnetic fields of the brain: Minimum norm estimates. *Medical & biological engineering & computing*, *32*(1), 35–42.

He, B., Sohrabpour, A., Brown, E., & Liu, Z. (2018). Electrophysiological Source Imaging: A Noninvasive Window to Brain Dynamics. *Annual Review of Biomedical Engineering*, *20*(1), 171–196. doi: 10.1146/annurev-bioeng-062117-120853

Hecker, L., Rupprecht, R., van Elst, L. T., & Kornmeier, J. (2020). ConvDip: A convolutional neural network for better M/EEG Source Imaging. *bioRxiv*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, G., Yu, Z. L., Wu, W., Liu, K., Gu, Z., Qi, F., ... Liang, J. (2020). Electromagnetic Source Imaging via a Data-Synthesis-Based Denoising Autoencoder. *arXiv preprint arXiv:2010.12876*.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

Ioannides, A. A., Bolton, J. P. R., & Clarke, C. J. S. (1990). Continuous probabilistic solutions to the biomagnetic inverse problem. *Inverse Problems*, *6*(4), 523.

Jin, K. H., McCann, M. T., Froustey, E., & Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, *26*(9), 4509–4522.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management science*, *6*(4), 366–422.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koenig, T., Prichep, L., Lehmann, D., Sosa, P. V., Braeker, E., Kleinlogel, H., ... John, E. R. (2002). Millisecond by millisecond, year by year: Normative EEG microstates and developmental stages. *Neuroimage*, *16*(1), 41–48.

Kornmeier, J., & Bach, M. (2004). Early neural activity in Necker-cube reversal: Evidence for low-level processing of a gestalt phenomenon. *Psychophysiology*, *41*(1), 1–8. doi: 10.1046/j.1469-8986.2003.00126.x

Kornmeier, J., & Bach, M. (2005). The Necker cube—an ambiguous figure disambiguated in early visual processing. *Vision research*, *45*(8), 955–960.

Kornmeier, J., & Bach, M. (2012). Ambiguous Figures – What Happens in the Brain When Perception Changes But Not the Stimulus. *Frontiers in Human Neuroscience*, *6*. doi: 10.3389/fnhum.2012.00051

Kornmeier, J., Bach, M., & Atmanspacher, H. (2004). Correlates of perceptive instabilities in event-related potentials. *International Journal of Bifurcation and Chaos*, *14*(02), 727–736.

Kornmeier, J., Friedel, E., Hecker, L., Schmidt, S., & Wittmann, M. (2019). What happens in the brain of meditators when perception changes but not the stimulus? *PLoS One*, *14*(10), e0223843.

Kornmeier, J., Friedel, E., Wittmann, M., & Atmanspacher, H. (2017). EEG correlates of cognitive time scales in the Necker-Zeno model for bistable perception. *Consciousness and cognition*, *53*, 136–150.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in cognitive sciences*, *3*(7), 254–264.

Leopold, D. A., Murayama, Y., & Logothetis, N. K. (2003, April). Very Slow Activity Fluctuations in Monkey Visual Cortex: Implications for Functional Brain Imaging. *Cerebral Cortex*, *13*(4), 422–433. doi: 10.1093/cercor/13.4.422

McCulloch, W. S., & Pitts, W. (1943, December). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133. doi: 10.1007/BF02478259

Mégevand, P., Hamid, L., Dümpelmann, M., & Heers, M. (2019). New horizons in clinical electric source imaging. *Zeitschrift für Epileptologie*, *32*(3), 187–193.

Michel, C. M., & Brunet, D. (2019). EEG Source Imaging: A Practical Review of the Analysis Steps. *Frontiers in Neurology*, *10*. doi: 10.3389/fneur.2019.00325

Mikulan, E., Russo, S., Parmigiani, S., Sarasso, S., Zauli, F. M., Rubino, A., ... Pigorini, A. (2020, April). Simultaneous human intracerebral stimulation and HD-EEG, ground-truth for source localization methods. *Scientific Data*, *7*(1), 127. doi: 10.1038/s41597-020-0467-x

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807–814).

Nunez, P. L., & Srinivasan, R. (2006). *Electric fields of the brain: The neurophysics of EEG*. Oxford University Press, USA.

Pantazis, D., & Adler, A. (2021). MEG Source Localization Via Deep Learning. *Sensors*, *21*(13), 4278.

Pascarella, A., Mikulan, E., Sciacchitano, F., Sarasso, S., Rubino, A., Sartori, I., ... Nobili, L. (2021). An in-vivo validation of ESI methods with focal sources. *bioRxiv*.

Pascual-Marqui, R., Michel, C. M., & Lehmann, D. (1994). Low-resolution electromagnetic tomography–a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, *18*, 49–65.

Pascual-Marqui, R. D. (1999). Review of methods for solving the EEG inverse problem. *International journal of bioelectromagnetism*, *1*(1), 75–86.

Pascual-Marqui, R. D. (2007). Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: Exact, zero error localization. *arXiv preprint arXiv:0710.3341*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Razorenova, A., Yavich, N., Malovichko, M., Fedorov, M., Koshev, N., & Dylov, D. V. (2020). Deep Learning for Non-Invasive Cortical Potential Imaging. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology* (pp. 45–55). Springer.

Robert, C., Gaudy, J.-F., & Limoge, A. (2002). Electroencephalogram processing using neural networks. *Clinical Neurophysiology*, *113*(5), 694–701.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, *45*(11), 2673–2681.

Spiridon, M., Fischl, B., & Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. *Human brain mapping*, *27*(1), 77–89.

Sun, R., Sohrabpour, A., Ye, S., & He, B. (2020, May). *SIFNet: Electromagnetic Source Imaging Framework Using Deep Neural Networks* (Preprint). Bioengineering. doi: 10.1101/2020.05.11.089185

Sutherland, M. T., & Tang, A. C. (2006, December). Reliable detection of bilateral activation in human primary somatosensory cortex by unilateral median nerve stimulation. *NeuroImage*, *33*(4), 1042–1054. doi: 10.1016/j.neuroimage.2006.08.015

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011). Brainstorm: A user-friendly application for MEG/EEG analysis. *Computational intelligence and neuroscience*, *2011*.

Tankelevich, R. (2019, February). Inverse problem's solution using deep learning: An EEG-based study of brain activity. Part 1 - rel. 1.0.

Van Veen, B. D., van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997, September). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE transactions on bio-medical engineering*, *44*(9), 867–880. doi: 10.1109/10.623056

Wernery, J., Atmanspacher, H., Kornmeier, J., Candia, V., Folkers, G., & Wittmann, M. (2015). Temporal processing in bistable perception of the Necker cube. *Perception*, *44*(2), 157–168.

Wipf, D., & Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, *44*(3), 947–966.

Zorzos, I., Kakkos, I., Ventouras, E. M., & Matsopoulos, G. K. (2021). Advances in Electrical Source Imaging: A Review of the Current Approaches, Applications and Challenges. *Signals*, *2*(3), 378–391.