PAPER • **OPEN ACCESS**

Neural network training with highly incomplete medical datasets

To cite this article: Yu-Wei Chang *et al* 2022 *Mach. Learn.: Sci. Technol.* **3** 035001

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
2 March 2022

REVISED
7 June 2022

ACCEPTED FOR PUBLICATION
22 June 2022

PUBLISHED
6 July 2022

PAPER

Neural network training with highly incomplete medical datasets

Yu-Wei Chang[1,7], Laura Natali[1,7], Oveis Jamialahmadi[2], Stefano Romeo[2,3,4], Joana B Pereira[5,6], Giovanni Volpe[1,*] and for the Alzheimer's Disease Neuroimaging Initiative[8]

[1] Department of Physics, University of Gothenburg, Gothenburg, Sweden
[2] Department of Molecular and Clinical Medicine, University of Gothenburg, Gothenburg, Sweden
[3] Clinical Nutrition Unit, Department of Medical and Surgical Sciences, University Magna Graecia, Catanzaro, Italy
[4] Cardiology Department, Sahlgrenska University Hospital, Gothenburg, Sweden
[5] Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Stockholm, Sweden
[6] Clinical Memory Research Unit, Department of Clinical Sciences, Lund University, Malmö, Sweden
[7] These authors contributed equally to this work.
[8] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu) (see supplementary materials). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: https://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgment_List.pdf.
[*] Author to whom any correspondence should be addressed.

E-mail: giovanni.volpe@physics.gu.se

Keywords: neural networks, Alzheimer disease, Covid-19, incomplete datasets

Supplementary material for this article is available online

Abstract

Neural network training and validation rely on the availability of large high-quality datasets. However, in many cases only incomplete datasets are available, particularly in health care applications, where each patient typically undergoes different clinical procedures or can drop out of a study. Since the data to train the neural networks need to be complete, most studies discard the incomplete datapoints, which reduces the size of the training data, or impute the missing features, which can lead to artifacts. Alas, both approaches are inadequate when a large portion of the data is missing. Here, we introduce GapNet, an alternative deep-learning training approach that can use highly incomplete datasets without overfitting or introducing artefacts. First, the dataset is split into subsets of samples containing all values for a certain cluster of features. Then, these subsets are used to train individual neural networks. Finally, this ensemble of neural networks is combined into a single neural network whose training is fine-tuned using all complete datapoints. Using two highly incomplete real-world medical datasets, we show that GapNet improves the identification of patients with underlying Alzheimer's disease pathology and of patients at risk of hospitalization due to Covid-19. Compared to commonly used imputation methods, this improvement suggests that GapNet can become a general tool to handle incomplete medical datasets.

1. Introduction

Supervised machine-learning models, such as the neural networks employed in deep learning, require to be trained and validated on large high-quality datasets [1]. In particular, these datasets need to be complete, i.e. each datapoint needs to have the values of all features, in order for them to be employed in standard neural-network training algorithms [2]. However, in many applications only incomplete datasets are available, i.e. each datapoint has values only for some of the relevant features [3]. For example, this often occurs in medical applications involving patient data, e.g. because various patients might undergo different clinical and diagnostic procedures at different times, with some patients even dropping out from a study [3]. Often, the dataset can contain medically relevant information despite being incomplete, motivating the interest on techniques that deal with missingness.

In order to deal with these missing data, there are two standard approaches. The first and most commonly employed one is to exclude the datapoints that do not have all relevant features [4]. However, data

exclusion reduces the statistical power of the dataset [5] and can introduce biases if the data are not missing completely at random [6]. The second and more complex approach is to impute the missing data. Various statistical imputation strategies have been proposed. The simplest one is arguably to substitute the missing values with their ensemble average [6]. In this study, we use as a benchmark five imputation techniques (see details in section 4.1). More sophisticated imputation strategies obtain better results employing, e.g. multilayer perceptrons, extreme gradient boosting machines, and support vector machines [7–9]. For example, a previous study [10] improved the classification of individuals in the datasets of the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative by employing a multiple recurrent graph convolutional network to impute the missing features (the brain volumes obtained from magnetic resonance imaging (MRI)). However, a major drawback of data imputation is that it can amplify biases in the available data or even introduce spurious correlations [4], especially when data are missing in great numbers or not at random [11]. Because of their drawbacks, both exclusion and imputation strategies can only deal with a limited amount of missing data. For instance, a study that employs similar imputation methods to impute 10%–50% of the data shows a significant increase in the error [12].

To address these limitations, here we introduce GapNet, an alternative neural-network training approach based on a hierarchical architecture that can make use of highly incomplete datasets. GapNet takes advantage of all available information without the need for imputation of missing data. As real-world test cases with highly incomplete datasets, we show that GapNet improves the identification of patients in the Alzheimer's disease (AD) continuum in the ADNI cohort and of patients at risk of hospitalization due to Covid-19 in the UK BioBank cohort. By distilling the information available in large incomplete datasets without having to reduce their size or to impute missing values, GapNet allows extracting valuable information from a wider range of datasets, being employable for many applications.

2. Results

2.1. GapNet working principle

To demonstrate the GapNet working principle, we first apply it on a simulated dataset, representing a two-class classification problem with F continuous input features.

2.1.1. The simulated dataset

We employ the simulated dataset Madelon [13], where the datapoints are clustered around the vertices of an F-dimensional hypercube and assigned to the class of the closest vertex (section 4.2). Figure 1(a) provides a schematic illustration for the case $F = 3$. Here, we consider a system with $F = 40$ features, so that $y = f(x_1, \ldots, x_{40})$, where $y \in \{0, 1\}$ is the class assignment for each datapoint. As shown in figure 1(b), the dataset consists of 1000 samples, where only 100 samples have all the 40 features (samples 451–550).

2.1.2. Vanilla approach

To establish a benchmark, we first consider a vanilla neural network approach, schematically shown in figure 1(c). We employ a dense neural network having an input layer with 40 nodes corresponding to the 40 features, two hidden layers with 80 nodes each with rectified linear unit (ReLU) activation, a dropout layer with a frequency of 0.5, and an output layer with a single node (sigmoidal activation). This vanilla neural network must be trained using the complete samples. Thus, we consider the dataset composed of complete cases and split it into five folds (four for training and one for testing). We train the neural network for 1000 epochs using the Adam optimizer [14] with binary cross-entropy loss.

2.1.3. Imputation approaches

To test the imputation methods, after the five-fold split, we concatenate the training dataset with the incomplete dataset, we impute the missing values using five imputation strategies (see details in section 4.1), and proceed to train the neural network as described above. Finally, the trained model is evaluated on the complete testing dataset.

2.1.4. GapNet stage I

The GapNet approach involves a two-stage training process, as schematically presented in figure 1(d). In training stage I (figure 1(d)), the input feature space is split into a set of non-overlapping clusters for which complete samples are available. We then build a group of identical neural network models, one for each cluster of the input data. In turn, we train a neural network using each of these clusters to predict the desired output. This stage can be seen as a means to extract the relevant information from all the clusters, which is conceptually similar to a recent preprint [15] for proposing a deep-clustering approach to maintain discriminablility in the embedding space. In the current example, we identify two clusters of features, the

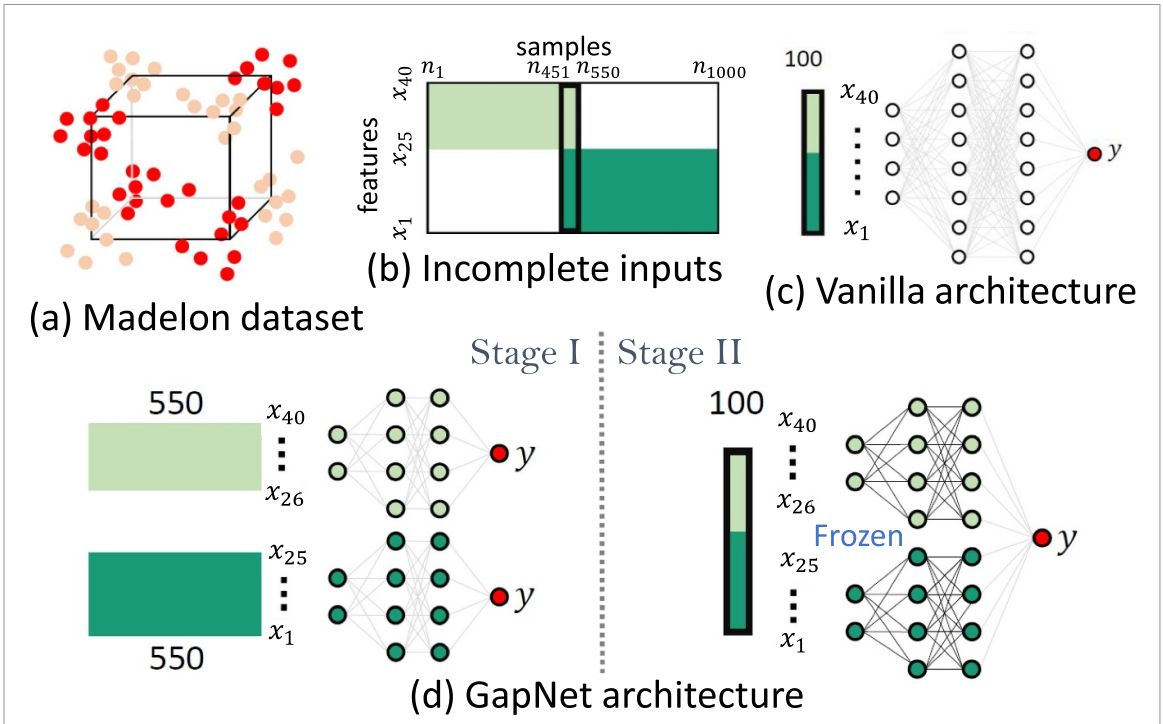


Figure 1. The GapNet approach to missing training data. (a) Schematic representation of Madelon [13], an artificial dataset used to test binary classification problems. (b) Simulated dataset with colored areas representing the available data, while white areas represent the missing data. (c) The baseline benchmark is provided by a vanilla neural network trained and tested on the complete samples. Additional benchmarking is provided by training neural networks with imputed data. (d) The GapNet approach: in training stage I, subsets containing all samples with certain features are used to train individual neural networks. In training stage II, their outputs are combined by an output node, whose connections are trained with all complete samples.

first one corresponds to features 1–25 (dark green area in figure 1(b)) and the second one to features 26 to 40 (light green area), each with 550 samples. Then, we train two dense neural networks to predict y using $x_1, \ldots, x_{25}$ and $x_{26}, \ldots, x_{40}$, respectively. Each neural network consists of an input layer (with 25 and 15 nodes, respectively, corresponding to each cluster of features), two hidden layers (with 50 and 30 nodes, respectively, with ReLU activation), a dropout layer (frequency 0.5), and finally an output layer with a single node (sigmoidal activation). We train these neural networks on the available samples (retaining 20 complete samples for testing, the same for both neural networks) for 1000 epochs (Adam optimizer [14], binary cross-entropy loss).

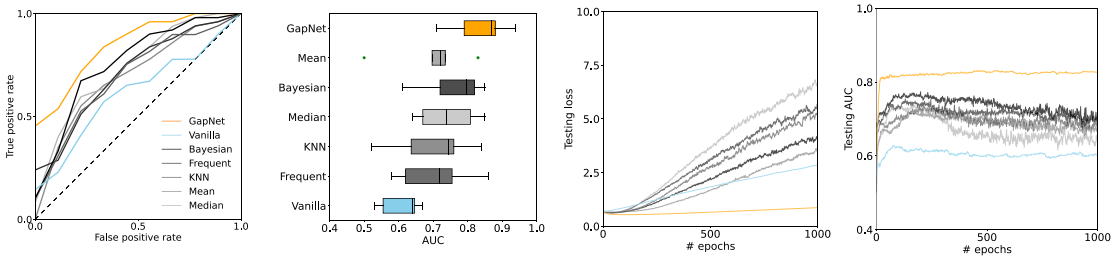
2.1.5. GapNet stage II

In training stage II (figure 1(d)), the input and hidden layers of the first-stage neural networks are combined into a single neural network, adding an output node (sigmoidal activation). The concatenated neural networks result in a model with an input for each one of the available features and a singular output. These new connections are trained for 1000 epochs (Adam optimizer [14], binary cross-entropy loss) using all available complete samples (retaining for testing the same 20 samples used for testing in training stage I). This concatenated neural networks can integrate the extracted information from separate clusters in stage I. We investigated also a variation of GapNet based on more stages of training (see details in supplementary materials, 'GapNet with intermediate training stage').

2.1.6. Comparison with the vanilla approach

In figure 2(a), we evaluate the performance using the receiver operator characteristic (ROC) curve, which plots the true positive rate versus the false positive rate as a function of the threshold. The area under the ROC curve (AUC) is then 0.5 for a random estimator and approaches 1 as the estimator performance improves. Figure 2(a) shows the ROC curve for GapNet (orange line) and vanilla neural network (blue line), obtained as the fivefold cross-validation average. The GapNet approach (AUC 0.838 $\pm$ 0.080) outperforms the vanilla neural network (AUC 0.608 $\pm$ 0.055). This difference is statistically significant with a p-value $p < 0.01$ ($z = 3.574$), tested using the Delong test [16].

Figure 2(b) shows box plots of the AUC values obtained from complete cases over fivefolds for the GapNet (orange) and the vanilla neural network (blue). GapNet results are significantly better than that of vanilla approach, demonstrating that GapNet can learn important information from the clustered network.



(a) Comparison of ROC curves (b) Boxplot of AU-ROC values (c) History of loss function (d) History of AU-ROC

Figure 2. GapNet improves the classification of simulated data. (a) The receiver operator characteristic (ROC) curves obtained with GapNet (orange, area under the curve (AUC) 0.838 $\pm$ 0.080) is significantly better than that of the vanilla neural network (blue, AUC 0.608 $\pm$ 0.055) and of other neural networks trained using imputed dataset. The solid lines represent the average ROC curves for the test datatset over the fivefold cross-validation. The black dashed line represents the ROC curve of a random classifier. The error bands are reported in supplementary figure S2 (first column). (b) Box plots of the AUC values obtained for the test dataset over the fivefold cross-validation for the GapNet (orange), the vanilla neural network (blue), and the impute-data-trained neural networks (gray lines). GapNet delivers the best result (large AUC). The black vertical lines mark the median value for each box plot, the green dots are the fliers, the boxes are the interquartile ranges, and the black horizontal lines are the whiskers. (c) The fivefold-averaged test loss for GapNet (orange) outperforms all other approaches (smaller loss consistently over 1000 epochs, while the losses for the other methods increase denoting overfitting). The error bands are reported in supplementary figure S2 (second column). (d) The fivefold-averaged testing AUC shows that GapNet (orange) produces the best predictive performance without overfitting. The error bands are reported in supplementary figure S2 (third column). The comparison for loss and AUC with the training dataset is shown in supplementary figure S1.

Supplementary table S1 shows that the sensitivity, specificity, accuracy, and precision computed by setting the predictions' threshold at 0.5 are improved by the GapNet approach.

A crucial aspect when selecting a neural network algorithm is its robustness to overfitting. For this reason, we monitor the test dataset over the training process. Figures 2(c) and (d) present the evolution of the test loss function and the test AUC over 1000 epochs (the evolution of the corresponding training is presented in supplementary figure S1). The test loss function for the vanilla neural network increases after the initial decrease and its AUC decreases after the initial increase, while GapNet results is robust against overfitting.

2.1.7. Comparison with imputation approaches

Figures 2(a) and (b) show that the GapNet approach outperforms imputed-data-trained neural networks in term of AUC. Nevertheless, the imputed-data-trained neural networks tend to outperform the simpler vanilla neural network. However, the test loss of the imputed-data-trained neural networks (gray lines in figure 2(c)) increases more during training than that of the vanilla neural network, denoting that using imputed data leads to more overfitting, even though their performance in terms of AUC remains better than the vanilla neural network (gray lines in figure 2(d)).

Overall, these proof-of-principle results on simulated data show the effectiveness and robustness of the GapNet approach to fully exploit incomplete datasets. Thus, the clear advantage is that GapNet can make use of all available data, without having to impute the missing data.

2.2. Identification of patients in the AD continuum in the ADNI cohort

As a first real-world GapNet application, we consider the identification of individuals with underlying amyloid pathology, which is one of the earliest pathological changes occurring in AD [18–21].

2.2.1. Incomplete data clustering

We use data from the ADNI cohort (section 4.3). In total, 869 individuals underwent 1465 neuroimaging visits including both baseline visits and subsequent longitudinal follow-up visits for some individuals. In each visit, one or more of the following three neuroimaging modalities were employed: structural MRI, amyloid positron emission tomography (amyloid-PET), and fluorodeoxyglucose-PET (FDG-PET), which are normally used to assess gray matter atrophy, amyloid deposition and glucose hypometabolism in AD, respectively (figure 3(a)). For MRI, we include the mean thickness of the 68 cortical regions of the Desikan atlas [22] and the volumes of 14 subcortical gray matter regions of the Aseg atlas [23]. For amyloid-PET, we include the mean amyloid standard uptake value ratio (SUVR) values from the same brain regions included in MRI. For FDG-PET, we include the SUVR of five composite brain regions [24]. The final number of features is 169, and the corresponding regions of interest (ROIs) are listed in supplementary table S2. MRI scans were acquired in 233 visits, amyloid-PET scans in 1045 visits, and FDG-PET scans in 1258. Only 120 visits (corresponding to 118 individuals, of which 40 cognitively normal subjects without amyloid pathology

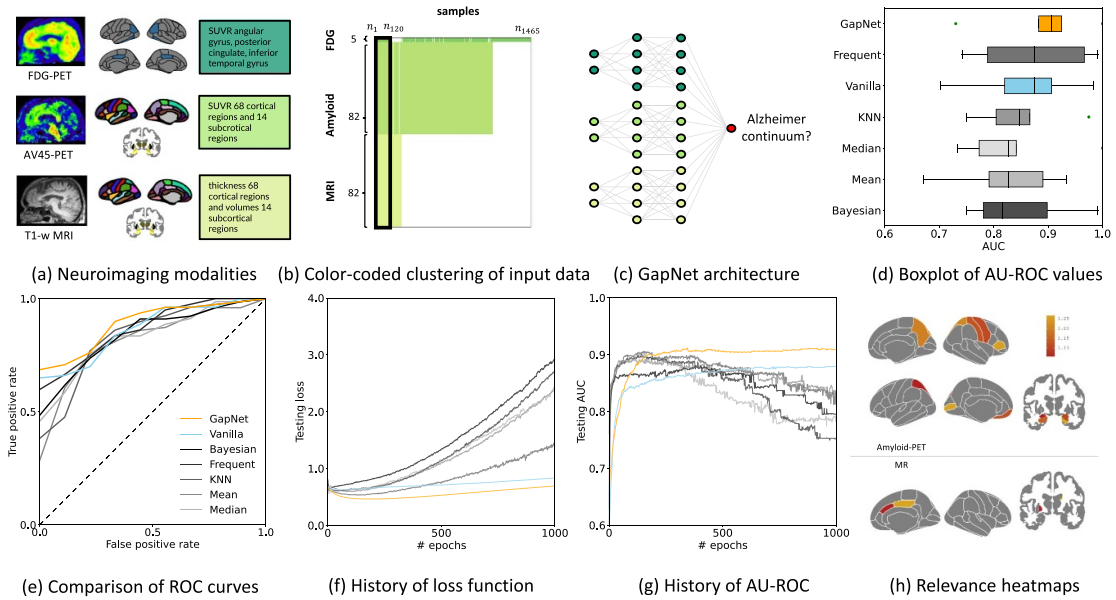


Figure 3. GapNet improves the identification of patients on the Alzheimer's disease continuum in the ADNI cohort. (a) Neuroimaging modalities (structural magnetic resonance (MRI), amyloid positron emission tomography (amyloid-PET), and fluorodeoxyglucose-PET (FDG-PET)) and corresponding brain regions of interest (ROIs). Mean values in each ROI are extracted by a standard Freesurfer processing pipeline [17]. (b) Schematic representation of the dataset. The dataset corresponds to 1465 visits in each of which one or more neuroimaging modalities are employed. Only 120 visits resulted in the acquisition of all three neuroimaging modalities (highlighted by the thick black line). (c) GapNet: first a neural network with two hidden layers is trained using each cluster of features (i.e. the brain region data for each neuroimaging modality). Then, all these neural networks are concatenated to return the final output. (d) Box plots of the test AUC values. GapNet outperforms the vanilla neural networks trained either on all individuals with complete features or on imputed dataset. The black mark represents the median, the black boxes are the interquartile ranges, and the black horizontal lines are the whiskers. (e) Receiver operating characteristic (ROC) curve showing that GapNet (orange, AUC 0.901 $\pm$ 0.075) outperforms the vanilla neural network (blue, AUC 0.86 $\pm$ 0.11) and the imputed-data-trained networks. The solid lines represent the averages over a fivefold cross-validation and the black dashed line represents the ROC curve of a random classifier. The error bands are reported in supplementary figure S4 (first column). (f) The average test loss and (g) the average test AUC show that GapNet (orange) outperforms all other approaches. The error bands are reported in supplementary figure S4 (second and third columns). The comparison with the loss and AUC for the training dataset is shown in supplementary figure S3. (h) Relevance heatmaps for the GapNet classification between cognitively normal individuals and patients in the AD continuum. The color-coded areas represent the regions with high classification importance. For amyloid-PET (upper panel), these regions include the left pericalcarine, right pars trianularis, bilateral superior parietal lobule, right precuneus, right postcentral lobule, left medial orbitofrontal cortex, right precentral lobule at the cortical level, as well as bilateral hippocampus and bilateral amygdala at the subcortical level. For MRI (lower panel), the regions are mainly from the right hemisphere, including the caudal anterior cingulate and posterior cingulate at the cortical level as well as the right caudate and left pallidum at the subcortical level.

and 78 subjects with amyloid pathology) resulted in the acquisition of all three neuroimaging modalities (figure 3(b)).

To apply the GapNet approach (figure 3(c)), we identify three clusters of features corresponding to the three imaging modalities. We evaluate the classification performance on the test sets in the fivefold cross-validation. We define the test set from 20% of the complete data ($N_{\text{test}} = 24$). We use the rest of the data for the training set ($N_{\text{train}} = 1441$ including the remaining 80% of the samples with complete data as well as all the incomplete samples). In training stage I, these three clusters of features are used to train three independent neural networks (input layer with 5, 82 and 82 nodes, respectively; two hidden layers with 10, 164, and 164 nodes, respectively, with ReLu activation; dropout layer with frequency 0.5; output layer with single neuron and sigmoidal activation). We train the MRI network on 209 samples, the amyloid-PET network on 1021 samples, and the FDG-PET network on 1234 samples (in all cases for 1000 epochs with binary cross-entropy loss function and Adam optimizer). These three networks are then combined into a single neural network in training stage II with a joint output node, and the new connections are retrained on the 96 complete training samples (1000 epochs, binary cross-entropy loss function, Adam optimizer).

2.2.2. Comparison with the vanilla approach

We compare the performance of GapNet and the vanilla neural network trained with complete cases in figures 3(d)–(g). In figure 3(d), the boxplots of AUC values are color-coded with GapNet in orange and the vanilla in blue. The results are sorted in descending order of median AUC values. The result of GapNet are better than those of the vanilla approach, indicating the combining information in the clustered network improves the classification task. Figure 3(e) shows that the ROC curve for the GapNet approach (orange,

AUC 0.901 $\pm$ 0.075) is superior than that of the vanilla neural network (blue, AUC 0.86 $\pm$ 0.11). This difference is statistically significant (Delong test [16], p-value $p < 0.05$, $z = 1.96$). Supplementary table S1 shows that the sensitivity, specificity, accuracy, and precision computed by setting the predictions' threshold at 0.5 are improved by the GapNet approach. Figures 3(f) and (g) show the test loss and AUC. It can be seen that GapNet performs better than the vanilla neural network and is not prone to overfitting. The evolution of the corresponding training is presented in supplementary figure S3.

2.2.3. Comparison with the imputed datasets

Figures 3(d) and (e) show that GapNet outperforms also the neural networks trained with imputed data. Interestingly, the AUC of the vanilla neural network are also better than those of most imputed-data-trained networks, probably due to the artifacts introduced during the imputation process.

The test loss (figure 3(f)) and the testing AUC (figure 3(g)) for the imputed-data-trained networks improve initially, but subsequently greatly overfit, eventually underperforming both GapNet and the vanilla neural network.

2.3. Prediction of Covid-19 hospitalization in the UK BioBank cohort

As a second example of a real-world application, we consider an incomplete dataset to predict hospitalization due to Covid-19. The dataset is based on the UK BioBank cohort, which contains information concerning Covid-19 test results, hospitalizations, and clinical examinations (section 4.4). The aim of this analysis is to discriminate patients at high risk of hospitalization due to severe Covid-19 symptoms from those at low risk of hospitalization, based on their previous medical records.

2.3.1. Incomplete data clustering

The cohort includes 4226 individuals and 32 different features with a varying number of records per feature, ranging from 1122 to 3776 values. The parameters include easily accessible testing information, such as red blood and white blood cell counts (see supplementary table S3 for the full list of features).

The missing data are irregularly distributed across the dataset, so we sort the features based on the number of missing values and we gather them into eight different clusters of four features each in order to reduce the amount of information loss. Figure 4(a) shows the input data color-coded based on the different clusters, from the largest cluster in dark green to the smallest in yellow. The colored areas represent the data, while the missing values are shown as white patches.

Discarding all patients with incomplete records leads to a reduced cohort of only 501 subjects, as indicated by the black line in figure 4(a). The schematic representation of the GapNet approach is provided in figure 4(b). We split the data into training and testing sets. The testing set includes about $20\%(N_{\text{test}} = 100)$ of the complete samples, while the training set includes the remaining complete data (401 samples) and all the incomplete data ($N_{\text{train}} = 4126$).

In training stage I, the eight clusters of features are used to train eight independent neural networks with input layer of four nodes. Each of these neural networks has two hidden layers with 25 nodes each (ReLu activation) and a dropout layer (frequency 0.5). The output layer consist of a single neuron (sigmoidal activation). We train the neural networks on each cluster of inputs for 1000 epochs using the binary cross-entropy loss function and Adam optimizer. These eight networks are then combined into a single neural network in training stage II with a joint output node, and the new connections are retrained on the complete samples (1000 epochs, binary cross-entropy loss function, Adam optimizer).

2.3.2. Comparison with the vanilla approach

We compare the performance of GapNet and the vanilla neural network in figure 4(c). The boxplots of AUC values are color-coded showing the GapNet in orange and the vanilla in blue, and the results are sorted in descending order of median AUC values. It is interesting to note that the GapNet results are significantly better that the vanilla approach, with an increase of 0.12 in the median AUC, demonstrating that combining the information in the clustered network improves the classification task.

Figure 4(d) shows that the ROC curve for the GapNet approach (orange, AUC 0.71 $\pm$ 0.06) is significantly better that of the vanilla neural network (blue, AUC 0.63 $\pm$ 0.05) as demonstrated by the Delong test [16] ($p = 0.0008$, $z = 3.4$). Also in this case, supplementary table S1 shows that the sensitivity, specificity, accuracy, and precision computed by setting the predictions' threshold at 0.5 are improved by the GapNet approach.

Finally, we compare the training process by looking at the evolution of the loss (figure 4(e)) and AUC (figure 4(f)) over 1000 epochs. The loss function for the vanilla neural network increases very rapidly after the initial decrease, while the GapNet results robust against overfitting also over a long training period. See also supplementary figure S5, for the comparison between training and testing loss function and AUC.

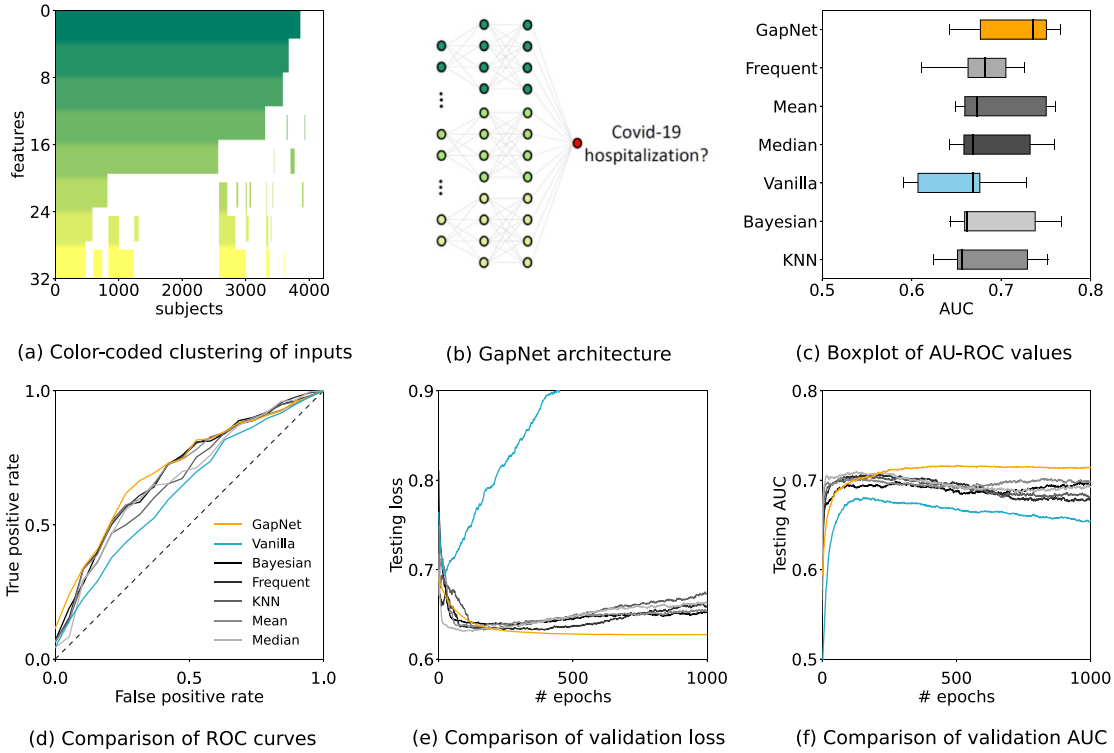


Figure 4. GapNet improves the identification of patients at risk of hospitalization due to Covid-19 in the UK BioBank cohort. (a) Eight clusters of features employed to predict hospitalization due to Covid-19 (sizes top to bottom 3776, 3627, 3552, 3322, 2652, 1122, 1188, 1493). The white areas represent the missing data. The 501 subjects with all features are delimited by the black line. (b) Network architecture employed for the GapNet approach. Neural networks with two hidden layers are trained for each cluster and then concatenated to obtain a single output. (c) Box plots showing the statistics of the AUC over the fivefold cross-validation. The black vertical lines mark the median value for each boxplot, the boxes are the interquartile ranges, and the black horizontal lines are the whiskers. The boxes are sorted in descending order of medians, and the performance of GapNet results better than all the other neural networks considered. (d) Receiver operating characteristic (ROC) curve showing that GapNet (orange, AUC 0.71 $\pm$ 0.07) outperforms the vanilla neural network (blue, AUC 0.64 $\pm$ 0.07) and all neural networks trained with imputed data. The solid lines represent the averages over the fivefold cross-validation and the black dashed line represents the ROC curve of a random classifier. The error bands are reported in supplementary figure S6 (first column). (e) The average test loss and (f) the average AUC show that the GapNet approach (orange) performs better that the vanilla neural network (blue) and the imputed-data-trained networks (gray lines). In particular, the GapNet approach is the only one that does not overfit. The error bands are reported in supplementary figure S6 (second and third columns). The comparison with the loss and AUC for the training dataset is shown in supplementary figure S5.

Moreover, after an initial stage, the AUC for the GapNet remains nearly constantly above 0.71, while the vanilla is below 0.7 throughout the training.

2.3.3. Comparison with the imputated datasets

Figure 4(c) shows that the GapNet approach outperforms also the imputed-data-trained networks in term of AUC. The neural networks trained with imputations at fixed values ('mean', 'median' and 'frequent') result intermediate between the GapNet approach and the vanilla neural network, while those trained with imputations with multiple Bayesian regression and k-nearest neighbor perform worse. A possible explanation is that the more advanced imputation methods enhance the biases already present in the dataset. This picture is generally confirmed by the ROC curves presented in figure 4(d).

Figures 4(e) and (f) show the averages of the test loss and of the test AUC, respectively. After an initially decreasing stage, the validation loss start to increase for all the imputation methods until values in the range 0.653–0.673, while the GapNet remains steadily around 0.627. The AUC follow similar trends, as all imputation methods start to decrease after the initial 300 epochs, ending up below 0.7, while GapNet remains stable around AUC = 0.71.

3. Discussion

The need to handle incomplete datasets is ubiquitous in all fields dealing with empirical data, such as medicine and engineering. While several imputation techniques have been developed, they always rely on (explicit or implicit) assumptions on the frequency and distribution of missing values, and often incur the

risk of introducing biases. Here, we have proposed an alternative approach that can be employed also when the missing data are very frequent and not missing at random. We have called this approach GapNet.

In the GapNet approach, the neural network undergoes two training stages, first training an ensemble of neural networks, each on a data subset with a complete cluster of features, and then combining these into a single estimator that is fine-tuned on the available complete samples. This estimator is better at predicting the results for complete data when compared both with a simple vanilla neural network approach trained only on complete data and trained on datasets imputed with five common imputation techniques. We have demonstrated the superior predictive ability of the GapNet approach on three examples, one corresponding to simulated data as a proof of principle, and the other two on real-word medical datasets of AD and Covid-19 patients.

These findings suggest that the improved predictions obtained by the GapNet approach are potentially generalizable to other datasets. We suggest employing GapNet can improve machine learning applications on real medical datasets [25–27], by increasing the dataset size adding incomplete feature and producing more reliable predictions when applied to new data.

3.1. ADNI

The major goal for the classification in the AD continuum is to establish a biomarker-based deep-learning model. In this respect, the GapNet approach outperforms the vanilla neural network as well as neural networks trained with imputed data, delivering robust results and detecting complex relationships when combining different imaging modalities. Another crucial evaluation is the importance analysis of features. In figure 3(h), we mapped the feature importance for the GapNet model by performing a permutation feature analysis [28] (see also supplementary figure S9). The results show that 16 out of the most important 20 features are derived from the amyloid-PET imaging modality, and the remaining 4 features are from MRI. Crucially, most of these features (or ROIs) have been reported to be impaired in AD by previous studies assessing patients at different disease stages. For amyloid-PET, the orbito-frontal and precuneus ROIs have been often identified in the early stages of amyloid pathology [29]; the subcortical ROIs (amygdala and hippocampus) are in line with the Aβ42 accumulation regions that have been reported in amyloid pathology during the early AD stages; and the other three cortical ROIs (pericalcarine, postcentral, and precentral) are consistent with the Aβ42 accumulation regions showing high SUVRs during the late AD stages [29]. For MRI, both the highlighted cortical ROIs (caudal anterior cingulate and posterior cingulate) and subcortical ROIs (caudate and pallidum) have also been reported in previous studies on informative regions across different stages of the AD continuum [30–35].

Overall, the most predictive features are mainly derived from the amyloid-PET modality, including parietal and frontal regions, which are typical sites of amyloid accumulation in AD [36]. These key findings on AD-related brain changes suggest that the GapNet estimator is capable of producing robust predictions for early and accurate AD diagnosis.

3.2. UK BioBank

A hot topic in COVID-19 research is to understand the effect of other comorbid diseases and conditions on the risk of developing severe Covid-19 symptoms [37–43]. This type of studies can help understanding, for instance, which individual's characteristics are associated with a higher risk of hospitalization, and who should be constantly monitored or prioritized in the vaccination process [44–46]. However, when analyzing patient data, the choice between discarding the incomplete values or imputing them, and the imputation technique employed [47] can lead to different results [48–50] depending also on the size of the cohort and the handling of missing data. Here, we provide a simple example to predict severe Covid-19 outcomes, with the aim of pointing out the advantages of using GapNet in this context: exploiting the incomplete values and avoiding biases or alterations of the original dataset. In the presented results, the incidence of the different features is consistent with previous findings in the literature. The most relevant clusters, in fact, include features such as systolic and diastolic blood pressure, red blood cell distribution width and serum creatinine levels (see details in supplementary materials, 'Feature importance analysis'), connected with known Covid-19 high-risk comorbidities [39, 41, 51, 52].

In conclusion, we have proposed GapNet, a conceptually effective model of neural network architectures, to produce more robust predictions in datasets with missing values, which have become increasingly common in research. We have shown how GapNet can detect complex nonlinear relationships between all the variables and is capable of learning and inferring from medical data with incomplete features. We have verified the effectiveness of GapNet in two real-world prominent datasets, the identification of patients in the AD continuum in the ADNI cohort, and the prediction of patients at risk of hospitalization due to Covid-19 in the UK BioBank cohort. We believe that GapNet is a preliminary step towards generic, scalable architectures that can investigate many real-world medical problems, or even tasks from many domains,

holding great potential for several future applications. One of the next steps will be to apply GapNet to more complex kinds of neural network architectures, such as recurrent and convolutional neural networks, as well as to apply it to more complex input data, such as time sequences and images.

4. Methods

4.1. Imputation methods

In this work, we consider five common imputation strategies to compare with GapNet and the vanilla approach. Three simple strategies [53] replace the missing features with a value depending on the statistical properties of the non-missing features; here, we use the mean, the median, and the most frequent features. The fourth imputation strategy, called multiple imputation by chained equations (MICE) [54], relies on a regression model (we use a Bayesian ridge regression) to recursively estimate the missing features. At each imputation step, one feature is used as the target of a regression model and another one as the input. The imputation consists of ten regression cycles, in each one of them, all features are regressed against all the others. We use a Bayesian ridge regression (labeled as 'Bayesian') as estimator for a MICE algorithm. The final imputation strategy considers the five nearest neighbors (in the Euclidean feature space) for each datapoint and uses the mean of the features of these datapoints to fill the missing features [55]. The distance between samples is estimated by euclidean distance between each pair of datapoints, discarding the missing values. The five smallest distances are selected and their mean completes the considered datapoint.

4.2. Simulated dataset

To verify the working principle of GapNet, we use a simulated dataset adapted from Madelon [13] implemented with scikit-learn [56] (figure 1(a)). We simulate a binary-classification dataset including 1000 samples with $F = 40$ features ($x_1, \ldots, x_{40}$). Of these features, 10 are informative features (x_3, x_9, x_{11}, x_{15}, x_{18}, x_{19}, x_{20}, x_{23}, x_{31}, x_{40}), 20 are linear combinations of the informative features (x_1, x_2, x_5, x_7, x_8, x_{10}, x_{12}, x_{13}, x_{14}, x_{17}, x_{22}, x_{24}, x_{26}, x_{28}, x_{30}, x_{32}, x_{34}, x_{36}, x_{37}, x_{39}), and 10 are uncorrelated random noise without information (x_4, x_6, x_{16}, x_{21}, x_{25}, x_{27}, x_{29}, x_{33}, x_{35}, x_{38}). These informative features are consistent with the ones individuated by GapNet (see details in supplementary materials, 'Feature importance analysis'). We introduce missingness in the dataset by removing the values of features x_1 to x_{25} from samples 1–450 and the values of features x_{25} to x_{40} for samples 551–1000, so that only 100 samples (samples 451–550) are complete (figure 1(b)).

4.3. ADNI cohort

The data used for this analysis were obtained from the ADNI database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. The individuals included in the current study were recruited as part of ADNI-GO and ADNI-2.

Amyloid pathology has been identified using a cut-off of $>976.6\,\mathrm{pg\,ml^{-1}}$ on cerebrospinal fluid (CSF) levels of Aβ42, following previously established procedures [57]. Subjects who are cognitively normal and have high CSF Aβ42 values are used as a healthy reference group, while subjects who are cognitively normal, have mild cognitive impairment, or AD dementia with low CSF Aβ42 values are included in the group with high risk of having AD (the AD continuum group).

4.3.1. Ethics approval and consent to participate

ADNI was reviewed and approved by all host study site institutional review boards and participants completed informed consent after receiving a comprehensive description of the ADNI. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. For up-to-date information, see http://adni.loni.usc.edu.

4.4. UK Biobank cohort

The dataset employed in the prediction of Covid-19 hospitalization is based on the general practitioners' records provided by the UK BioBank dedicated to Covid-related research [58].

We build the labels using the COVID-19 test result records and the hospital inpatient register data. The labels are assigned '0' for patients positive to Covid-19 who do not appear in the hospital records, and '1' for patients hospitalized due to Covid-19 and identified by the ICD-10 diagnostic code U07.1 [59]. The input

values come from merging the primary care data from the two largest providers, The Phoenix Partnership (TPP) [60] and Egton Medical Information System (EMIS) [61]. The data are cleaned to obtain the dataset: we select the patients positive to Covid-19, we filter only the medical coding systems SNOMED CT [62] and CTV3 [63], and we include records in the five years interval 1 January 2015 to 1 January 2020 (antecedent to the first reported case in the dataset). The selected values include many standard medical examinations, such as pressure measurements, body weight, and blood tests. At this point, we cut the less common codes to obtain a dataset of 501 subjects with complete features and we undersample the more represented category (the non-hospitalized subjects are nearly five times larger) to obtain a balanced dataset. The final number of attributes incorporated is 32, listed in table S3 together with their corresponding Read Code.

4.4.1. Ethics approval and consent to participate
UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (Approval Number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (37142) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001).

Data availability statement

The ADNI dataset is owned by the Alzheimer's Disease Neuroimaging Initiative (ADNI). Data are publicly and freely available from the Institutional Data Access / Ethics Committee (contact via http://adni.loni.usc.edu/data-samples/access-data/) upon sending a request that includes the proposed analysis and the named lead investigator.

The Covid-19 dataset generated by UK Biobank analysed during the current study are available via the UK Biobank data access process (see www.ukbiobank.ac.uk/register-apply/). The exact number of samples currently available in UK Biobank may differ slightly from those described in this paper.

No new data were created or analysed in this study.

Acknowledgments

We acknowledge support from the MSCA-ITN-ETN project *ActiveMatter* sponsored by the European Commission (Horizon 2020, Project Number 812780) and from the project ERC-CoG project MAPEI sponsored by the European Commission (Horizon 2020, Project No. 101001267). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense Award Number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research has been conducted using data from UK Biobank, a major biomedical database, under the following Application Number: 37142.

Conflict of interest

The authors declare no competing financial or non-financial interests.

Author contributions statement

Conceptualization: J B P, G V. Data curation: O J, S R, J B P. Formal analysis: L N, Y W C. Funding acquisition: J B P, G V. Investigation: L N, Y W C, J B P, G V. Methodology: L N, Y W C, J B P, G V. Project

administration: G V. Resources: J B P, S R. Programming: L N, Y W C, G V. Supervision: J B P, S R, G V. Validation: L N, Y W C, J B P, G V. Visualization: L N, Y W C, J B P, G V. Writing—original draft: L N, Y W C. Writing—review and editing: L N, O J, S R, Y W C, J B P, G V.

ORCID iDs

Yu-Wei Chang https://orcid.org/0000-0001-9598-2278
Giovanni Volpe https://orcid.org/0000-0001-5057-1846

References

[1] Yanase J and Triantaphyllou E 2019 A systematic survey of computer-aided diagnosis in medicine: past and present developments *Expert Syst. Appl.* **138** 112821
[2] Shilo S, Rossman H and Segal E 2020 Axes of a revolution: challenges and promises of big data in healthcare *Nat. Med.* **26** 29–38
[3] Little R J *et al* 2012 The prevention and treatment of missing data in clinical trials *New Engl. J. Med.* **367** 1355–60
[4] Jakobsen J C, Gluud C, Wetterslev J and Winkel P 2017 When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts *BMC Med. Res. Methodol.* **17** 162
[5] Ginkela J R V, Linting M, Rippe R C A and van der Voort A 2020 Rebutting existing misconceptions about multiple imputation as a method for handling missing data *Stat. Dev. Appl.* **102** 297–308
[6] Kang H 2013 The prevention and handling of the missing data *Korean J. Anesthesiol.* **64** 402–6
[7] Liu C-H, Tsai C-F, Sue K-L and Huang M-W 2020 The feature selection effect on missing value imputation of medical datasets *Appl. Sci.* **10** 2344
[8] Zhang X, Yan C, Gao C, Malin B A and Chen Y 2020 Predicting missing values in medical data via XGBoost regression *J. Healthc. Inform. Res.* **4** 383–94
[9] Huang M-W, Lin W-C, Chen C-W, Ke S-W, Tsai C-F and Eberle W 2016 Data preprocessing issues for incomplete medical datasets *Expert Syst.* **33** 432–8
[10] Vivar G *et al* 2020 Simultaneous imputation and disease classification in incomplete medical datasets using multigraph geometric matrix completion (MGMC) (arXiv:2005.06935 [cs, stat])
[11] Hughes R A, Heron J, Sterne J A C and Tilling K 2019 Accounting for missing data in statistical analyses: multiple imputation is not always the answer *Int. J. Epidemiol.* **48** 1294–304
[12] Jadhav A, Pramod D and Ramanathan K 2019 Comparison of performance of data imputation methods for numeric dataset *Appl. Artif. Intell.* **33** 913–33
[13] Guyon I, Gunn S, Ben-Hur A and Dror G 2004 Result analysis of the NIPS 2003 feature selection challenge *Advances in Neural Information Processing Systems* **17** 545–52
[14] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
[15] Saban O and Cukur T 2022 Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets (arXiv:2204.02275v1)
[16] DeLong E R, DeLong D M and Clarke-Pearson D L 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach *Biometrics* **44** 837–45
[17] Fischl B 2012 Freesurfer *NeuroImage* **62** 774–81
[18] Jack C R *et al* 2013 Amyloid-first and neurodegeneration-first profiles characterize incident amyloid pet positivity *Neurology* **81** 1732–40
[19] Aizenstein H J *et al* 2008 Frequent amyloid deposition without significant cognitive impairment among the elderly *Arch. Neurol.* **65** 1509–17
[20] Lim Y Y *et al* 2013 Rapid decline in episodic memory in healthy older adults with high amyloid-β *J. Alzheimer's Dis.* **33** 675–9
[21] Vlassenko A G, McCue L, Jasielec M S, Su Y, Gordon B A, Xiong C, Holtzman D M, Benzinger T L S, Morris J C, Fagan A M 2016 Imaging and cerebrospinal fluid biomarkers in early preclinical Alzheimer disease *Ann. Neurol.* **80** 379–87
[22] Desikan R S *et al* 2006 An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest *NeuroImage* **31** 968–80
[23] Fischl B *et al* 2002 Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain *Neuron* **33** 341–55
[24] Landau S M, Harvey D, Madison C M, Koeppe R A, Reiman E M, Foster N L, Weiner M W and Jagust W J 2011 Associations between cognitive, functional and FDG-PET measures of decline in AD and MCI *Neurobiol. Aging* **32** 1207–18
[25] Weiner M W *et al* 2017 The Alzheimer's disease neuroimaging initiative 3: continued innovation for clinical trial improvement *Alzheimer's Dementia* **13** 561–71
[26] Marek K *et al* 2018 The Parkinson's progression markers initiative (PPMI)—establishing a PD biomarker cohort *Ann. Clin. Transl. Neurol.* **5** 1460–77
[27] Allen N *et al* 2012 UK Biobank: current status and what it means for epidemiology *Health Policy Technol.* **1** 123–6
[28] Molnar C 2020 *Interpretable Machine Learning* (Morrisville, NC: Lulu.com)
[29] Palmqvist S, Mattsson N and Hansson O 2016 Cerebrospinal fluid analysis detects cerebral amyloid-β accumulation earlier than positron emission tomography *Brain* **139** 1226–36
[30] Kautzky A, Seiger R, Hahn A, Fischer P, Krampla W, Kasper S, Kovacs G G and Lanzenberger R 2018 Prediction of autopsy verified neuropathological change of Alzheimer's disease using machine learning and MRI *Front. Aging Neurosci.* **10** 406
[31] Jones B F, Barnes J, Uylings H B M, Fox N C, Frost C, Witter M P and Scheltens P 2005 Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric MRI study *Cereb. Cortex* **16** 1701–8
[32] Fennema-Notestine C, Hagler D J, McEvoy L K, Fleisher A S, Wu E H, Karow D S and Dale A M 2009 Structural MRI biomarkers for preclinical and mild Alzheimer's disease *Hum. Brain Mapp.* **30** 3238–53
[33] Davatzikos C, Bhatt P, Shaw L M, Batmanghelich K N and Trojanowski J Q 2011 Prediction of MCI to AD conversion, via MRI, CSF biomarkers and pattern classification *Neurobiol. Aging* **32** 2322.e19–27
[34] Madsen S K, Ho A J, Hua X, Saharan P S, Toga A W, Jack C R, Weiner M W and Thompson P M 2010 3D maps localize caudate nucleus atrophy in 400 Alzheimer's disease, mild cognitive impairment and healthy elderly subjects *Neurobiol. Aging* **31** 1312–25

[35] Rallabandi V P S, Tulpule K and Gattu M *et al* 2020 Automatic classification of cognitively normal, mild cognitive impairment and Alzheimer's disease using structural MRI analysis *Inform. Med. Unlocked* **18** 100305
[36] Grothe M J, Barthel H, Sepulcre J, Dyrba M, Sabri O and Teipel S J 2017 *In vivo* staging of regional amyloid deposition *Neurology* **89** 2031–8
[37] Foy B H *et al* 2020 Association of red blood cell distribution width with mortality risk in hospitalized adults with SARS-CoV-2 infection *JAMA Netw. Open* **3** e2022058
[38] Henry B M, Benoit J L, Benoit S, Pulvino C, Berger B A, de Olivera M H S, Crutchfield C A and Lippi G 2020 Red blood cell distribution width (RDW) predicts COVID-19 severity: a prospective, observational study from the cincinnati SARS-CoV-2 emergency department cohort *Diagnostics* **10** 618
[39] Wang C *et al* 2020 Red cell distribution width (RDW): a prognostic indicator of severe COVID-19 *Ann. Transl. Med.* **8** 1230
[40] Pakos I S *et al* 2020 Characteristics of peripheral blood differential counts in hospitalized patients with COVID-19 *Eur. J. Haematol.* **105** 773–8
[41] D'Marco L, Puchades M J, Romero-Parra M, Gimenez-Civera E, Soler M J, Ortiz A and Gorriz J L 2020 Coronavirus disease 2019 in chronic kidney disease *Clin. Kidney J.* **13** 297–306
[42] Hu X, Chen D, Wu L, He G and Ye W 2020 Declined serum high density lipoprotein cholesterol is associated with the severity of COVID-19 infection *Clinica Chim. Acta* **510** 105–10
[43] Radenkovic D, Chawla S, Pirro M, Sahebkar A and Banach M 2020 Cholesterol in relation to COVID-19: should we care about it? *J. Clin. Med.* **9** 1909
[44] Hassan-Smith Z, Hanif W and Khunti K 2020 Who should be prioritised for COVID-19 vaccines? *Lancet* **396** 1732–3
[45] Cook T and Roberts J 2021 Impact of vaccination by priority group on UK deaths, hospital admissions and intensive care admissions from COVID-19 *Anaesthesia* **76** 608–16
[46] Hezam I M, Nayeem M K, Foul A and Alrasheedi A F 2021 Covid-19 vaccine: a neutrosophic MCDM approach for determining the priority groups *Results Phys.* **20** 103654
[47] Zhang C *et al* 2020 A novel scoring system for prediction of disease severity in COVID-19 *Front. Cell. Infection Microbiol.* **10** 318
[48] Zeng F, Li L, Zeng J, Deng Y, Huang H, Chen B and Deng G 2020 Can we predict the severity of coronavirus disease 2019 with a routine blood test? *Pol. Arch. Intern. Med.* **130** 400–6
[49] Bastug A *et al* 2020 Clinical and laboratory features of COVID-19: predictors of severe prognosis *Int. Immunopharmacol.* **88** 106950
[50] Elliott J, Bodinier B, Whitaker M, Delpierre C, Vermeulen R, Tzoulaki I, Elliott P and Chadeau-Hyam M 2021 Covid-19 mortality in the UK Biobank cohort: revisiting and evaluating risk factors *Eur. J. Epidemiol.* **36** 299–309
[51] Gallo Marin B *et al* 2021 Predictors of COVID-19 severity: a literature review *Rev. Med. Virol.* **31** 1–10
[52] Lippi G, Wong J and Henry B M 2020 Hypertension and its severity or mortality in coronavirus disease 2019 (COVID-19): a pooled analysis *Pol. Arch. Intern. Med.* **130** 304–9
[53] Donders A R T, Van Der Heijden G J M G, Stijnen T and Moons K G M 2006 Review: a gentle introduction to imputation of missing values *J. Clin. Epidemiol.* **59** 1087–91
[54] Azur M J, Stuart E A, Frangakis C and Leaf P J 2011 Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatric Res.* **20** 40–49
[55] Batista G E A P A and Monard M C 2003 An analysis of four missing data treatment methods for supervised learning *Appl. Artif. Intell.* **17** 519–33
[56] Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30
[57] Hansson O *et al* 2018 CSF biomarkers of Alzheimer's disease concord with amyloid-β PET and predict clinical progression: a study of fully automated immunoassays in BioFINDER and ADNI cohorts *Alzheimers. Dement.* **14** 1470–81
[58] UK BioBank 2020 (available at: www.ukbiobank.ac.uk/) (Accessed September 2021)
[59] World Health Organization 2020 Emergency use icd codes for COVID-19 disease outbreak (available at: www.who.int/standards/classifications/classification-of-diseases/emergency-use-icd-codes-for-COVID-19-disease-outbreak)
[60] UK Biobank online resource centre 2020 GP clinical event records (TPP source)
[61] UK Biobank online resource centre 2020 GP clinical event records (EMIS source)
[62] Systemized Nomenclature of Medicine Clinical Terms (SNOMED) International 2002 (available at: www.snomed.org/)
[63] Clinical Terms Version 3 2018 Data coding 7128 (available at: www.ukbiobank.ac.uk/)