




OPEN

Multimodal graph attention network for COVID-19 outcome prediction

Matthias Keicher^{1,6}, Hendrik Burwinkel^{1,6}, David Bani-Harouni^{1,6}, Magdalini Paschali^{1,4}, Tobias Czempiel¹, Egon Burian^{2,3}, Marcus R. Makowski², Rickmer Braren², Nassir Navab¹ & Thomas Wendler^{1,5}

When dealing with a newly emerging disease such as COVID-19, the impact of patient- and disease-specific factors (e.g., body weight or known co-morbidities) on the immediate course of the disease is largely unknown. An accurate prediction of the most likely individual disease progression can improve the planning of limited resources and finding the optimal treatment for patients. In the case of COVID-19, the need for intensive care unit (ICU) admission of pneumonia patients can often only be determined on short notice by acute indicators such as vital signs (e.g., breathing rate, blood oxygen levels), whereas statistical analysis and decision support systems that integrate all of the available data could enable an earlier prognosis. To this end, we propose a holistic, multimodal graph-based approach combining imaging and non-imaging information. Specifically, we introduce a multimodal similarity metric to build a population graph that shows a clustering of patients. For each patient in the graph, we extract radiomic features from a segmentation network that also serves as a latent image feature encoder. Together with clinical patient data like vital signs, demographics, and lab results, these modalities are combined into a multimodal representation of each patient. This feature extraction is trained end-to-end with an image-based Graph Attention Network to process the population graph and predict the COVID-19 patient outcomes: admission to ICU, need for ventilation, and mortality. To combine multiple modalities, radiomic features are extracted from chest CTs using a segmentation neural network. Results on a dataset collected in Klinikum rechts der Isar in Munich, Germany and the publicly available iCTCF dataset show that our approach outperforms single modality and non-graph baselines. Moreover, our clustering and graph attention increases understanding of the patient relationships within the population graph and provides insight into the network's decision-making process.

Reflecting on the coronavirus disease 2019 (COVID-19) pandemic¹, the first wave, in particular, brought unprecedented challenges to the healthcare system. The exponential surge in cases overwhelmed intensive care units (ICUs), presenting scenes that had never been witnessed in the age of modern medicine^{2,3}. During such a state of emergency, optimizing the allocation of hospital resources, e.g., ICU beds, mechanical ventilators, or personnel, becomes crucial. An essential aspect of effective patient management is correctly assessing treatment necessity and potential outcomes. When there is only a limited understanding of a previously unknown disease paired with highly multimodal data, as in the case of a novel pandemic, performing such an assessment and prediction of patient outcomes is very challenging. The resulting sudden overload of care facilities, in combination with the high complexity of the obtained data structure, motivates the need for assistance systems for fast outcome prediction and triaging based on available patient information. At the start of the COVID-19 pandemic, upon a patient's hospital admittance, a multitude of parameters—such as sex, age, body weight, symptoms, co-morbidities, blood

¹Computer Aided Medical Procedures and Augmented Reality, School of Computation, Information and Technology, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany. ²Department of Diagnostic and Interventional Radiology, School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany. ³Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Technical University of Munich, Ismaninger Str. 22, 81675 Munich, Germany. ⁴Department of Radiology, Stanford University School of Medicine, Stanford, CA 94304, USA. ⁵Department of Diagnostic and Interventional Radiology and Neuroradiology, Clinical Computational Medical Imaging Research, University Hospital Augsburg, Stenglinstr. 2, 86156 Augsburg, Germany. ⁶These authors contributed equally: Matthias Keicher, Hendrik Burwinkel and David Bani-Harouni. ✉email: matthias.keicher@tum.de

cell counts, inflammatory parameters, biochemical values, cytokine profiles, among others—were obtained and documented⁴. These parameters—“tabular data” in the following—and radiological images, including radiographs or X-ray computed tomography (CT) images, were available within the first hours after a new patient arrived at the hospital. This deems the two data sources ideal for early triaging and outcome prediction. Traditional disease outcome prognosis performed by clinicians is based, however, also on anamnestic information and clinical experiences. The general logic of a physician’s decision-making process partly relies on the information embedded in similar patients where the outcome and the connection of this information to currently treated patients are known⁵. Such population relationships are particularly useful when little to no disease-specific epidemiological information, as well as deep medical expertise, is available, as might be the case in potential upcoming health crises. Following this method of reasoning, we propose a decision support system that performs multimodal data analysis to create a population graph that clusters patients which is then used in a graph neural network with an attention mechanism to refine the patient outcome prediction by taking into account similar patients. The used similarity metric, attention mechanism, and generated pathology segmentations provide added insight into the decision-making process. This becomes possible as the weighting of clinical features and the most influential patients used in the prediction process can be directly observed. Our contributions are as follows:

- We introduce U-GAT, an end-to-end, graph-based method for leveraging medical images, extracted radiomics, and clinical data for predicting patient outcomes. In this work, we use multimodal data to predict COVID-19 patient outcomes, namely ICU admission, need for ventilation, and mortality. This method is generalizable and can easily be adapted to different types of anatomies, modalities, and clinical tasks.
- Our model uses a multitasking approach, where segmentation and classification are learned simultaneously. A U-Net⁶ is used to segment the healthy and pathological regions of the lung in chest CTs. From these segmentations, we extract scalar values, in the following called “radiomics”, e.g., the percentage of healthy or pathological lung tissue volume, and subsequently perform a joint feature fusion of image, radiomic, and clinical features. This combined feature vector is refined in our Graph Attention Network (GAT)⁷ by leveraging similar patients to perform the final outcome prediction.
- We present an interpretable, multimodal patient similarity metric for graph construction and effective batch selection.
- We introduce a novel equidistant image sampling method allowing for end-to-end training of volumetric image feature extraction in a graph convolutional setting with multiple patients per batch graph. At test time, we make use of all available slices.
- We thoroughly evaluate our novel approach on a newly acquired dataset collected in Klinikum rechts der Isar in Munich during the first COVID-19 wave of 2020 as well as an external and publicly available dataset and showcase our model’s ability to predict patient-specific disease outcomes. The dataset from Klinikum rechts der Isar contains expert annotations of a diverse range of COVID-19 pathologies and is available for research purposes upon request.
- While we validate our method on COVID-19, it is disease-agnostic and the insights about modeling multimodal data without prior experience with patient trajectories can be easily adapted to new contexts of novel disease outbreaks.

Related work

Fusing imaging and tabular data

Within the field of multi-modal learning, within recent years, different works have been published. One interesting approach to interweave features from multiple modalities was introduced by Perez et al. for visual reasoning tasks⁸. A Feature-wise Linear Modulation (FiLM) layer affinely transformed the output of a Convolutional Neural Network (CNN) with a learned scaling and shifting factor using the text of the input question. Dynamic Affine Feature Map Transform (DAFT)⁹ extended FiLM to combine the features of 3D brain T1-weighted MRI scans and non-imaging biomarkers for Alzheimer’s prediction. DAFT affinely transformed the imaging features extracted by a 3D Fully CNN by a learned scaling and shifting factor using nine non-imaging features, such as age, sex, and genetic factors. A multi-headed cross-attention block has been recently proposed to fuse imaging and tabular data for skin lesion classification using a transformer architecture¹⁰ showing marginal improvement over joint fusion.

Taleb et al.¹¹ introduced ContIG, a self-supervised pre-training approach trained on 500k individuals from the UK Biobank¹² combining retinal fundus images with genetic information tested on different classification and segmentation downstream tasks. A contrastive loss based on cosine similarity was utilized to decrease the distance of the embeddings of the multimodal features of one patient. Moreover, Duanmu et al.¹³ combined breast MRI scans and clinical biomarkers to predict chemotherapy response. A network trained on the non-imaging data learned scalar weights that were multiplied with the intermediate results from the imaging network to generate feature maps containing interactive information between imaging and tabular data.

Inspired by the holistic decision-making approach taken by experienced physicians and medical boards, which involves integrating knowledge from diverse fields of expertise¹⁴, there is a growing interest in developing similar machine learning systems. Huang et al.¹⁴ outlined three methods for integrating features in deep learning models for radiology: merging extracted image features with non-imaging features (early fusion), combining features with a joint end-to-end (image) feature extraction (joint fusion), and consolidating predictions made by independent models (late fusion). Our method employs joint fusion. In contrast to early and late fusion, joint fusion processes the different modalities separately but integrates them during intermediate stages, allowing for inter-modal interactions and joint model training. Backpropagating the loss function to the feature extraction

allows the model to optimize the feature extraction based on the final output or prediction error, ensuring a more synergistic learning process. In the following, we review the fusion methods in the context of COVID-19.

Early fusion

For the COVID-19 detection and the prediction of patient outcome, most of the proposed methods integrating both imaging and non-imaging data apply early fusion of features^{4,15–20}. Chassagnon et al.²¹ demonstrated the importance of combining a wide range of non-imaging and extracted imaging features for the outcome prognosis of COVID-19 patients in an ensemble of machine-learning models. Shiri et al.²² achieved the best results in COVID-19 survival prediction by combining lesion-specific radiomics and clinical data. Gong et al.²³ improved the results for predicting severe COVID-19 outcomes by adding blood values to other clinical features and extracted radiomics.

Late fusion

Applying late fusion with penalized logistic regression, Ning et al.²⁴ reported an improvement in both COVID-19 severity and mortality outcome prediction compared to the stand-alone lung CT CNN and non-imaging Multilayer Perceptron (MLP) models. Tariq et al.²⁵ explored different fusion methods for predicting the need for hospitalization of COVID-19 patients and found the early fusion of different electronic medical record features to work best for this task.

Joint fusion

To the best of our knowledge, we are the first to propose a joint fusion method combining imaging and non-imaging data to predict ICU admission, ventilation, and mortality, or severity, depending on the dataset used.

Graph convolutional networks for medical applications

Previous studies have showcased the potential of Graph Convolutional Networks (GCNs) in medical applications, particularly in optimizing the processing of medical image information. Parisot et al.²⁶ pioneered using GCNs on population graphs to improve Alzheimer's and Autism Spectrum Disorder prediction. They also demonstrated that varying the patient information included in the graph setup significantly affects network performance⁵. Later works sought to diminish performance dependencies on graph generation, with Anirudh et al.²⁷ suggesting a bootstrapping strategy and ensemble learning for GCNs. Cosmo et al.²⁸ introduced a self-learning method for graph construction, integrating both imaging and non-imaging data for optimized GCN learning behavior. Further, GCNs have also been employed in medical image segmentation^{29–32} and Graph Attention Networks (GATs)⁷ have been utilized for patient diagnosis^{33,34}.

The aforementioned works leveraged already extracted image features. However, Burwinkel et al.³⁵ proposed a methodology that used GCNs on image data directly. They showed that end-to-end processing of imaging and clinical data within a GCN can improve performance due to optimized feature learning. At the same time, the proposed approach allowed for more effective usage of inter-class connections within the graph. We will expand upon this concept within our developed methodology and explain the implications in detail in section “[Method](#)”.

GCNs for COVID-19

In the context of COVID-19 diagnosis, GCNs have mainly been adapted for disease detection. Wang et al.³⁶ and Yu et al.³⁷ built graphs based on the similarity of extracted CT image features and classified the nodes for the presence of infiltrates. In addition to image features, Song et al.³⁸ and Liang et al.³⁹ used the acquisition site along with other features to improve COVID-19 detection. Instead of modeling a patient population, Saha et al.⁴⁰ converted edges detected in chest CT and X-ray images to graphs and leveraged these for detecting COVID-19. Huang et al.⁴¹ used GCNs to refine the segmentation of COVID-19 infections. Finally, Di et al.⁴² learned an uncertainty-vertex hypergraph to distinguish between community-acquired pneumonia and COVID-19. To the best of our knowledge, we propose the first graph-based end-to-end patient outcome prediction method by leveraging a population graph combining chest CTs and tabular patient data.

Multitask learning for COVID-19

Recent works^{4,22,43,44} on the radiological assessment of COVID-19 patients have shown a high correlation between disease burden and patient outcome, e.g., the probability of ICU admission. Several deep learning methods have been proposed to exploit this correlation with multitasking approaches^{45–47}. The majority of the proposed multitask methods focus on the joint detection of COVID-19 infection and the binary segmentation of related pathologies in lung CT images^{48–52}. Concerning COVID-19 patient outcome prediction, another set of works applied to multitask learning on the joint estimation of the severity of COVID-19 and various classification and segmentation tasks^{53,54}. Similar to our approach, Nappi et al.⁵⁵ used bottleneck features of a pretrained U-Net to predict COVID-19 progression and mortality. However, they did not optimize end-to-end, incorporate clinical patient data, or utilize a graph-based approach for the classification.

Method

Our proposed method provides an effective way to process multimodal patient information such as CT images X_I combined with clinical data X_C for disease outcome prediction of patients, as shown in Fig. 1. For a COVID-19 patient admitted to the hospital, the three outcomes we predict are the need for ICU admission, the need for mechanical ventilation, and the survival of the patient (for our in-house dataset), while we predict severity for the iCTCF dataset. Additionally, we use the segmentation of COVID-19 pathologies as an auxiliary target to

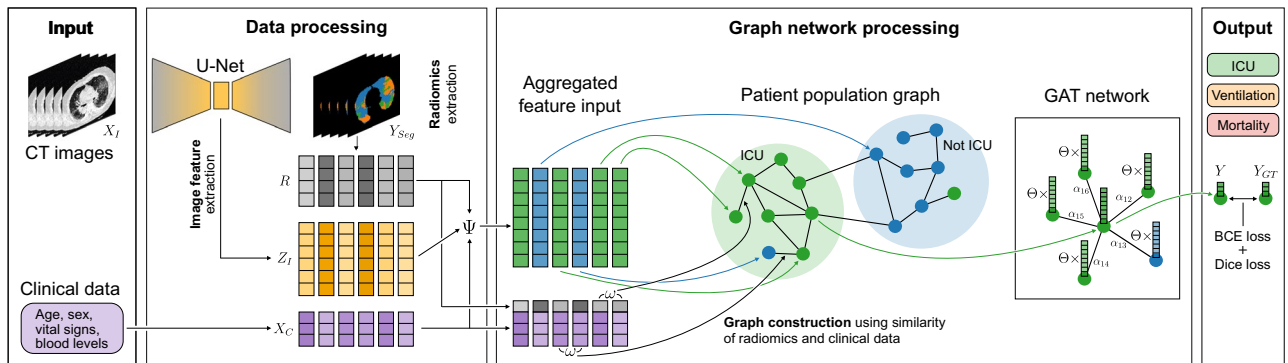


Figure 1. U-GAT is an end-to-end model, integrating learned image and radiomic features (Z_I and R) with clinical metadata X_C —such as age, sex, vital signs, and blood levels—for disease outcome prediction. Disease-affected area segmentation Y_{Seg} in CT images X_I aids in extracting radiomic features R and regularizes image feature Z_I extraction. These features coalesce into a multimodal vector via function Ψ . Test patients cluster with training patients in a graph based on radiomic and clinical data feature distance ω . A Graph Attention Network (GAT) then refines the features to predict the most probable outcome Y , utilizing learned linear transformation Θ and patient attention coefficients α_{ij} . Comparison to outcome ground truth Y_{GT} is facilitated by binary cross-entropy (BCE), while the Dice loss aids in the auxiliary segmentation task with manual ground truth. In the COVID-19 context, we segment lung CT image pathologies and predict patient ICU admission, ventilation need, and survival for the KRI dataset, and severity for the iCTCF dataset (not shown here).

improve the training. From the segmentation output, we calculate radiomic features R that represent the relative burden of the lung for each pathology class. To effectively incorporate the different modalities, we introduce a new framework that combines the segmentation capabilities of U-Net with the analytic strengths of GCNs. This network uses a population graph constructed with the similarity of clinical patient data X_C and radiomic features R to refine the image features of each patient. The proposed method operates end-to-end to perform an ideal combination of image feature representation learning, U-Net image segmentation, and graph data processing. The graph is pre-computed before training, and at test time, patients are dynamically connected to the graph of patients in the training set to ensure no data leaking during training and allow for usage flexibility in a clinical setting.

Graph-based image processing

To allow for inference on unseen data samples, we employ spatial graph convolutions. Compared to spectral methods, this approach allows an extension to unseen samples, not requiring retraining for every new patient. As explained in section “[Fusing imaging and tabular data](#)”, combining image data X_I with other modalities is essential for a holistic patient outcome prediction. For GCNs, image-based information is usually first extracted either manually or with a pretrained CNN. These extracted image features are then, in a second step, processed within the graph network. While this strategy lessens the memory demands of imaging data, it precludes the possibility of end-to-end optimization. Burwinkel et al.³⁵ showed that the image feature extraction process can potentially benefit from an underlying graph structure through an end-to-end feature extraction with a graph neural network since relevant graph information can backpropagate into the learned extraction process. We leverage this concept for the processing of the provided CT image information. Every CT image $x_{I,i}$ is processed by a U-Net to perform segmentation on the individual image slices. The calculated bottleneck feature maps of the U-Net are extracted (description in section “[Segmentation and image feature extraction](#)”) and processed to receive a corresponding representation $z_{I,i}$, usable within the graph neural network.

Equidistant subsampling

Utilizing GCNs for end-to-end feature extraction from high-resolution 3D images presents a major challenge due to high memory demands, which restricts the number of patient instances per batch. However, GCNs necessitate diversity in a single batch for effective feature aggregation. To accommodate larger batches, we suggest equidistant subsampling of S slices per volume along the axial view during training. If the main axis length is Z , each volume is divided into $\lfloor Z/S \rfloor$ stacks of S slices, omitting $(Z \bmod S)/2$ slices on both sides. This strategy not only enhances the likelihood of detecting disease-impacted areas but also mitigates overfitting by distributing scarce 3D volume data into multiple patient samples. At test time, the complete stack of slices is used, encompassing the entire 3D volume.

Graph construction method

We define a binary, directed graph $G(V, E)$ with vertices V and connecting edges E . Every vertex $v_i \in V$ corresponds to a stack of CT images $x_{I,i} \in X_I$ (sampling process described in section “[Graph-based image processing](#)”), a vector of radiomics features $r_i \in R$ (extraction process described in detail in section “[Segmentation and image feature extraction](#)”) and clinical data $x_{C,i} \in X_C$. For building the graph we concatenate the clinical data X_C and radiomics features R into one tabular feature and calculate the distance ω between two vertices based on these features. Each vertex v_i is connected with its k nearest neighbors. As an alternative to feature selection, we

propose to weight each feature based on a statistical analysis of the training data. Statistically important features should therefore have a bigger influence on the distance and similarity calculation. Possible weightings include correlation coefficients, e.g. the Pearson correlation for continuous features, or estimated mutual information⁵⁶ between the input features and the target labels like Y_{ICU} calculated on the training set. The motivation to use mutual information is to discover non-linear associations between the features and predicted labels, in addition to linear relationships. All distances are calculated on the z-scores normalized features. In Fig. 2, the k-nearest neighbors (KNN) graphs for one training set are visualized with and without weighting of the distance with mutual information.

Segmentation and image feature extraction

The proposed method is built on a joint image feature extraction and segmentation backbone. For this, any encoder-decoder-based architecture with a compressed bottleneck representation and segmentation output can be used. As described in more detail in section “Experiments”, we choose the original 2D U-Net architecture⁶ with small adaptations for our experiments. The S equidistant slices forming an input image $x_{I,i}$ (see section “Graph-based image processing”) are processed as a batch in parallel. Hence, for each slice, a 2D segmentation of the healthy lung and pathologies is generated. The image representation used for the classification task is extracted with a global average pooling of the two-dimensional bottleneck features of each slice, reducing the bottleneck size $c \times d_1 \times d_2$ with the number of channels c and the spatial dimensions d_1 and d_2 to a vector with the length of c per slice. The resulting S slice-wise image representations are then transformed into a single patient-wise representation. To achieve this, the slice features are aggregated by taking the element-wise maximum along the stacking dimension resulting in a single vector with size c . This vector is then passed through a final fully connected layer followed by a leaky ReLU activation to obtain the latent image representation $z_{I,i} \in Z_I$. Based on the improved performance reported by Goncharov et al.⁵⁴ using the final feature map of the U-Net instead of the bottleneck, we evaluated this approach, but initial results showed a substantial drop in performance which is why we did not investigate this concept any further.

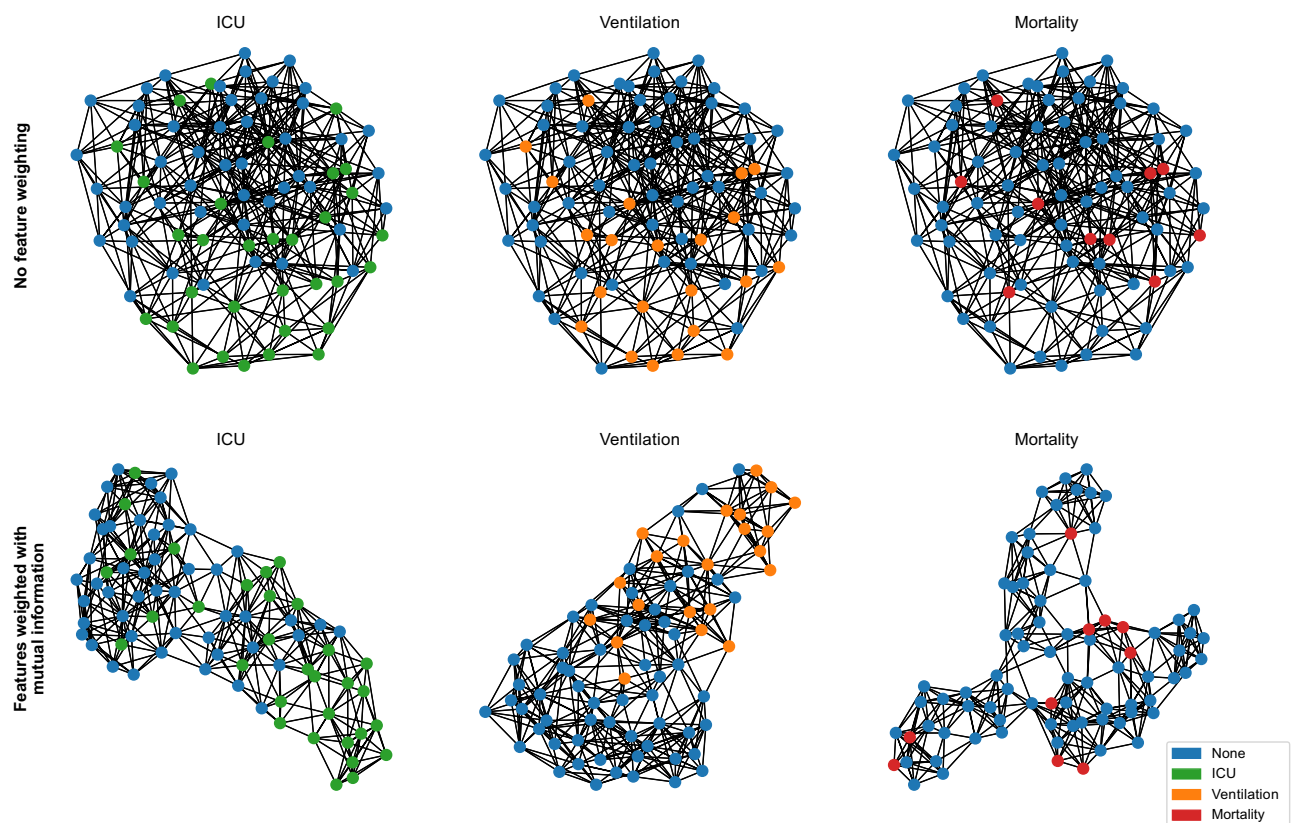


Figure 2. The initial patient clustering, visualized for the KRI dataset, is based on clinical and radiomic feature similarity. The top row displays graphs created by linking each node to its seven nearest neighbors based on Euclidean distance. To optimize this graph construction for the task at hand, we propose feature weighting in the distance calculation, informed by its task-specific mutual information⁵⁶ of features (bottom row). This prioritizes essential features in clustering and tailors the graph for specific tasks without needing feature selection or prior knowledge.

Extraction of radiomic features

Inspired by Burian et al.⁴, the clinical data is complemented with radiomics features R that are automatically extracted from the segmentation output Y_{Seg} . In addition to being more robust to overfitting than extracted image features, this improves the interpretability of the network by providing intermediate results that can easily be verified by visualizing the segmentation output. For instance, in the case of COVID-19-related tasks, one can use quantifications of COVID-19 pathologies in the segmented lung.

Multimodal feature fusion

Our methodology harnesses the multimodal data in a two-fold manner. On the one hand, radiomics extracted from the segmentation output and clinical patient parameters are employed to form the patient population graph. On the other hand, we synergistically fuse latent image features, extracted radiomics, and clinical data into the node features of this graph. This integrated representation encapsulates all salient attributes of a patient, providing a comprehensive patient characterization for subsequent processing. The three input sources provided by the image data $x_{I,i} \in X_I$ and resulting extracted features $z_{I,i}$, extracted radiomics features $r_i \in R$ and clinical data $x_{C,i} \in X_C$ constitute three separate modalities used within the graph network to perform the classification task for an individual patient node v_i within the graph. Especially the clinical data X_C can provide valuable orthogonal information to the imaging-based other two contributions. We have incorporated the latent bottleneck features $z_{I,i}$ of the U-Net to allow for end-to-end feature optimization, facilitating an image feature extraction beyond hand-crafted radiomics. To assure that the influence of every modality is equally considered during processing, we are using a linear transformation on every modality to receive a feature representation of equal size. These representations are then processed within an aggregation function Ψ to receive the corresponding fused representation $z_{f,i}$ used within the graph network:

$$z_{f,i} = \Psi \left(\sigma \left(\Theta_I z_{I,i} \right), \sigma \left(\Theta_R r_i \right), \sigma \left(\Theta_C x_{C,i} \right) \right), \quad (1)$$

where σ is a non-linear activation function and $\Theta_I \in \mathbb{R}^{F_I \times F_f}$, $\Theta_R \in \mathbb{R}^{F_R \times F_f}$, $\Theta_C \in \mathbb{R}^{F_C \times F_f}$ are learnable linear transformations, which map the incoming feature dimension onto dimension F_f . Possible approaches for Ψ are concatenation, averaging, pooling, or attention mechanisms. Concatenation was experimentally chosen for our proposed method as discussed later in section “Experiments”.

Classification of patient outcome

The graph processing of our proposed method is based on graph attention layers (GAT)⁷. They combine effective processing of the provided neighborhood with the possibility for direct inference on new unseen data samples while maintaining filter localization and low computational complexity. The attention-based graph processing allows us to incorporate the clinical patient data X_C effectively into the learning process by basing the graph construction on the similarity of tabular features and creating $N(i)$ for every $z_{f,i}$. Further, the attention mechanism allows for an intelligent learned weighting of the neighbors. Now, a transformation of representation $z_{f,i}$ does not only rely on the representation itself but receives weighted contributions from all $z_{f,j} \in N(i)$. This process has the potential to stabilize the prediction for patients with an uncharacteristic initial representation of its corresponding class, but which is localized within the correct data cluster.

Experiments

Datasets

KRI dataset

The KRI dataset (“in-house” dataset) consists of 132 COVID-19 patients, expanding on the dataset with 65 patients described in⁴. To assess the patient outcome, different parameters were collected: admission to the ICU, the necessity of mechanical ventilation, and the patient’s survival. These outcomes presented themselves immediately or sometime after general admission to the hospital. The complete dataset is available on request for research purposes in the frame of the BFS project AZ-1429-20C. For each CT volume, the total lung, healthy lung tissue, ground-glass opacifications (GGO), consolidations, and pleural effusions area were annotated by expert radiologists (4–8 years of experience). We combined pleural effusion and consolidation into a single class named “Other pathologies” since distinguishing between the two classes is a highly challenging task, even for senior radiologists⁵⁸ as both have almost the same Hounsfield unit range. Moreover, pleural effusion is only present in the most severe cases in only 1.2% of all available patients. See radiomics statistics for this dataset in the supplementary material.

iCTCF dataset

To substantiate the versatility of our method, we have extended our evaluation to a larger and publicly available dataset: the iCTCF dataset²⁴ (“external” dataset). It comprises 1,521 patients and includes high-resolution CT images, clinical data, and patient outcomes. The main difference to the KRI annotations is the lack of image annotations of different pathologies in the lung. Since our work focuses on triaging patients infected with COVID-19, we exclude the control group and only predict the outcome severity of PCR-positive COVID-19 patients. This results in 620 patients with mild (Type I) and 274 patients with severe outcomes (Type II)²⁴ leading to a total of 894 patients. Since the iCTCF dataset does not contain any annotations of the CT images, we employ a U-Net, pretrained on a diverse dataset^{59–61} of lung CT-slices of Hofmanninger⁶², to generate lung masks and a nnU-Net by Isensee et al.⁶³, pretrained on the COVID-19 Lung CT Lesion Segmentation Challenge⁶⁴, to infer the pathology annotation. The radiomic *COVID-19 burden* was extracted using this annotation, resembling the percentage of the lung affected by COVID-19 pathologies.

Experimental setup

We first evaluate the proposed method on the KRI dataset using a nested 5-fold cross-validation⁶⁵ stratified by the ICU labels. For this, the dataset is split into five equally sized folds, each containing a similar amount of ICU patients. In nested cross-validation, there are outer and inner evaluation loops for testing and validation. In each of the five outer loops, one fold is selected as a test set, and the remaining four folds are used for training and validation. In the four inner loops, three folds are selected for training and one for validation. This is repeated until every combination has been used for testing and validation, resulting in a total of 20 repetitions.

For the experiments presented here, following Burian et al.⁴, the static lung CT images taken at admission were used in combination with the following clinical features and blood test results: age, sex, body temperature, percutaneous oxygen saturation, leukocytes, lymphocytes, C-reactive protein (CRP), creatine, D-Dimer, lactate dehydrogenase (LDH), creatine kinase, troponin T, interleukin 6 (IL-6), thrombocytes. The outcomes included: the need for mechanical ventilation, admission to the ICU, and patient survival (mortality). All three tasks are binary classification tasks. We focus on evaluating the main task of ICU prediction and extend some experiments on ventilation and mortality outcome tasks to explore multitasking and the translation to other tasks. The experiments were conducted with ten equidistant samples ($Z = 10$) of the chest CT images, producing nine subvolumes per patient. During training, a random subvolume is chosen for each patient. At validation and test time, the whole patient volume is sampled. Since there is only a single test patient per batch, the pre-computed image features and radiomics of the other patients can be used. During the test phase, a batch graph consists of one test node and 18 neighboring nodes from the training set that serves as a context for this new patient. For all our experiments, we set the modality aggregation function ψ to perform concatenation.

For the iCTCF dataset, following the evaluation of Ning et al.²⁴, we split the data in a 10-fold cross-validation regime. In every run, eight folds are used for training and 1 for validation and testing, respectively. Given only a single radiomic of the COVID-19 burden of the lung is available, we concatenate the extracted radiomic with the clinical data and encode this tabular data into a joint embedding vector of size 64 for each patient. Since the dataset contains many features, of which most have only low mutual information with the target outcome, only features with estimated mutual information higher than 0.05 were used for graph construction. All available clinical features were used as patient node features. We stopped training when there was no improvement in the validation classification loss for five epochs.

Network parameters and training

We conducted all experiments in PyTorch 1.7.0⁶⁶ and PyTorch Geometric 1.7.0⁶⁷ using the Adam optimizer with a base learning rate of 5×10^{-4} and a weight decay of 3×10^{-5} . As the segmentation and image feature extraction backbone, we choose the classical 2D U-Net architecture proposed by Ronneberger et al.⁶ with the following modifications in the double convolution blocks: an added batch normalization layer after each activation for faster convergence and a padding of one pixel in each convolution layer to align input and output image size of the network. The final layer consisted of a one-dimensional convolution to the number of output classes followed by a softmax layer. We used a Dice loss as introduced by Milletari⁶⁸ for segmentation and a binary cross-entropy (BCE) loss for classification. Further training details can be found in the supplementary material. For graph processing, we used a two-layer GAT⁷.

Graph construction

We employed the KNN graph construction method introduced in section “Graph-based image processing” using a mutual information weighted distance metric for the following experiments after comparing it to other methods on the validation set. For ω we chose the weighted Euclidean distance (Minkowski distance of second order, $p = 2$). Here, every feature dimension was weighted by its approximate mutual information with the respective outcome label. The mutual information was estimated using the method proposed by Ross et al.⁵⁶ with 3 neighbors averaging the results of 30 repetitions. We compared weighting the KNN with mutual information against weighting with Pearson correlation. To understand the impact of weighting features, we also compared these weighted methods against an unweighted KNN. For the unweighted setup, we evaluated different subsets of manually selected features as can be seen in Table 3. The number of neighbors k used for graph construction was set in a hyperparameter search on the validation set.

Ablative testing and comparison to baselines

To investigate the effect of the different components of our method, we show ablative results on the test set. We mainly evaluate two components: the image and radiomics feature extraction of the U-Net and the GAT classification. The end-to-end U-GAT feature extraction is compared with features extracted from a simple frozen U-Net trained on the same annotations but without any multi-tasking, and the end-to-end image features from a ResNet18 as proposed by He et al.⁶⁹. It is important to note that radiomics were not used in the ResNet18-GAT architecture because ResNet18 does not produce segmentations. To evaluate the contribution of GAT, we compare it with the following classification method alternatives:

- Weighted K-nearest neighbors (KNN): The default scikit-learn weighted k-nearest neighbor classifier using the inverse Euclidean distance of all features as the similarity metric for neighbor selection and for weighting of neighbor labels⁷⁰.
- Multilayer Perceptron (MLP): This classifier is a simple neural network with a hidden layer size of 64 followed by a leaky ReLU activation and a 10% dropout.
- GraphSAGE: replacing the GAT operator with GraphSAGE⁵⁷.

In addition to ablative testing, we compare unimodal vs. multimodal approaches by evaluating the performance of using an MLP classifier using only clinical data or only image features extracted by a ResNet18. An overview of the type of data used in each method is given in Table 1.

U-GAT ensemble and comparison with Random Forest

Random Forest is an ensemble method that is an effective classifier for small datasets since they are less prone to overfitting due to the Law of Large Numbers⁷¹ and provide the additional benefit of interpretability. As discussed in section “Fusing imaging and tabular data”, Burian et al.⁴ and Chao et al.¹⁵ have successfully deployed Random Forests to use tabular radiomics and clinical data for ICU prediction. In this experiment, we focus on the task of ICU prediction and explore if an ensemble of our proposed model can improve its performance due to increased robustness against overfitting and how it compares to the well-established Random Forest classifier. To form an ensemble we average the predicted probabilities of the 4 models trained on the inner loops of the nested cross-validation and evaluate them on the 5 test sets of the outer loop of the nested cross-validation.

Metrics for segmentation and classification

As our proposed method follows a multitask approach including the CT segmentation and each of the tasks of ICU, ventilation and mortality prediction individually, the evaluation criteria can be divided into segmentation and classification metrics. To measure the overlap between segmented regions and ground truth, we use the Dice score (DS). The main metrics for evaluating the binary classification performance are average precision (AP) and the area under the receiver operating characteristic curve (AUC), as they are independent of selected classification thresholds. Given that all tasks have a severe class imbalance, the F1 score (F1) has been chosen as the main threshold-dependent metric. In the ensemble experiments, the balanced accuracy score (bACC), sensitivity, and specificity are additionally reported. For all threshold-dependent metrics, the optimal threshold is set using the validation results and maximizing the Youden’s J statistic⁷²: $J = \text{sensitivity} + \text{specificity} - 1$. The classification metrics are all binary and were calculated using scikit-learn 0.24.1⁷⁰.

Results and discussion

Population graph construction

In the first phase of experiments on our KRI dataset, we optimized the population graph construction method. This involved evaluating various feature selections and distance weights to improve the KNN-based graph construction. We found that connecting each node with its seven nearest neighbors provided optimal results, based on a hyperparameter search using a simple, unweighted KNN classifier. Two measures - mutual information and Pearson correlation - were used to weight features in the distance calculation of the similarity metric used for KNN neighbor selection. Table 2 shows the top 10 of the average of both measures for the ICU task. While a Pearson correlation > 0.3 and mutual information > 0.1 can be observed in the ICU and ventilation tasks for some features, the mortality showed significantly lower values indicating the difficulty of the task at hand (see supplementary material). The percentage of the healthy lung has the highest mutual information for all tasks. The results shown in Table 3 confirmed that our proposed weighting with the mutual information method yielded the best outcomes, particularly for the ICU task, as indicated by an AP of 0.722 ± 0.096 and an AUC of 0.757 ± 0.142 . The comparison with manual feature selection, e.g., only using clinical data, showed that using all available

Architecture	Multitasking	Multimodal	Patient modalities			Patient similarity	
			Images	Radiomics	Clinical	Radiomics	Clinical
MLP-Clinical	–	–	–	–	✓	–	–
RF-Clinical	–	–	–	–	✓	–	–
ResNet18	–	–	✓	–	–	–	–
ResNet18-GAT	–	✓	✓	–	✓	–	✓
U-Net*+RF	–	✓	–	✓	✓	–	–
U-Net*+KNN	–	✓	–	✓	✓	✓	✓
U-Net*+MLP	–	✓	✓	✓	✓	–	–
U-Net*+GraphSAGE	–	✓	✓	✓	✓	✓	✓
U-GAT*	–	✓	✓	✓	✓	✓	✓
U-GAT	✓	✓	✓	✓	✓	✓	✓

Table 1. Backbones and classifiers used for evaluation with the respective features for patients and the distance metric (similarity). *Images* describes the latent image features extracted with an image encoder. *Radiomics* stands for the radiomics extracted from the segmentation networks. *Clinical* data includes vital signs, blood values, and demographic information. We compare U-GAT to other end-to-end trained methods only using clinical data (MLP-Clinical), only using image data (ResNet18), and a GAT with a CNN backbone without an auxiliary segmentation task (ResNet18-GAT). In addition, we compare the performance of different classifiers on the image features extracted from a frozen U-Net, marked with a *, i.e., U-Net*. KNN is a k-nearest neighbors classifier. GraphSAGE is a graph convolutional method without an attention mechanism⁵⁷. Multitasking refers to the joint training of classification and segmentation.

Task	Feature	Category	Mutual information	Pearson correlation
ICU	Healthy lung (%)	Radiomics	0.244 ± 0.052	-0.596 ± 0.033
ICU	Ground-glass opacity (%)	Radiomics	0.184 ± 0.043	+0.577 ± 0.026
ICU	Other pathologies (%)	Radiomics	0.144 ± 0.055	+0.471 ± 0.048
ICU	C-reactive protein	Clinical	0.104 ± 0.038	+0.372 ± 0.071
ICU	Interleukin 6	Clinical	0.091 ± 0.023	+0.091 ± 0.137
ICU	Age	Clinical	0.087 ± 0.031	+0.018 ± 0.062
ICU	Lymphocytes	Clinical	0.047 ± 0.027	-0.062 ± 0.112
ICU	Temperature	Clinical	0.043 ± 0.040	-0.016 ± 0.116
ICU	Serum creatinine	Clinical	0.041 ± 0.045	+0.009 ± 0.125
ICU	Thrombocytes	Clinical	0.039 ± 0.037	-0.007 ± 0.060
ICU	Creatine kinase (total)	Clinical	0.037 ± 0.040	+0.113 ± 0.110

Table 2. Top 10 features sorted by the mutual information for each task and its Pearson correlation in the KRI dataset. The average is calculated on the training sets of all repetitions.

Task	Architecture	Distance features	Distance feature weights	AP	AUC
ICU	U-GAT*	Age, sex	-	0.512 ± 0.109	0.573 ± 0.109
ICU	U-GAT*	Clinical	-	0.671 ± 0.152	0.720 ± 0.135
ICU	U-GAT*	Radiomics	-	0.670 ± 0.145	0.720 ± 0.116
ICU	U-GAT*	All	-	0.704 ± 0.080	0.733 ± 0.073
ICU	U-GAT*	All	Pearson correlation	0.697 ± 0.122	0.751 ± 0.088
ICU	U-GAT*	All	Mutual information	0.722 ± 0.096	0.757 ± 0.142

Table 3. Evaluation of edge features and their weighting used for distance calculation on the validation set of the KRI dataset. Highest values are in bold.

features is most effective, but mutual information estimation can further help identify the most relevant features and give them a higher weight in the similarity metric. The external dataset confirmed the feature importance of radiomic data. Here, the COVID-19 burden has the highest mutual information with the severity labels (see supplementary material). A key benefit of using a weighted distance for KNN graph construction is that the graph can adapt to each task without prior knowledge. Fig. 2 shows the graph for each task on the KRI dataset with and without weighting the distance measure with mutual information. Besides improving classification, an effective similarity measure can be used to identify relevant patients that have been treated in the past and support the decision-making process of physicians by enabling them to analyze the disease progression of similar patients.

U-GAT evaluation

In the next set of experiments shown in Table 4, we evaluate the different components of the proposed method and compare the results to baseline methods. Our multimodal method outperforms the unimodal MLP, limited to only clinical data as input. The same picture presents when limiting the model to solely use imaging data, as is the case for the ResNet18 method. Here, again our proposed methods outperform ResNet18 on all tasks. These experiments showcase the benefit of a multimodal approach. U-GAT achieves a higher AP than the other methods in all ablations of replacing the U-Net with a ResNet18 and replacing the GAT with an MLP or a GraphSAGE. This shows that leveraging similar patients from the training set is useful for refining the features of test patients. We see similar results on the external dataset where U-GAT has a higher AP of 0.593 ± 0.106 than the single modality models MLP and ResNet18 with 0.556 ± 0.099 and 0.525 ± 0.140 , respectively, highlighting the advantage of multimodal learning.

The results of joint end-to-end training of the segmentation and classification task seem to improve the AP slightly for all tasks on both datasets. While the average Dice score is lower in all multitask setups than in the segmentation single-task setup (see supplementary material), this makes the segmentation task a suitable auxiliary task to improve classification results. On the KRI dataset, both the ICU and ventilation predictions reached the highest AP of 0.699 ± 0.149 and 0.644 ± 0.142 , respectively, when multitasking with segmentation. The mortality task generally achieves worse results. One main explanation for this effect is the immense data imbalance that is present for the mortality task, with only 19 out of 132 positive samples. Additionally, we observe low mutual information of the radiomics and clinical features with the mortality outcome (supplementary material, Table S5). This indicates that the features at hand might not be sufficiently predictive for this specific task. Several relevant clinical aspects closely connected to multiorgan failure, such as heart, kidney and liver parameters, were not available in the datasets. The evaluation on the external dataset shows the same picture where joint end-to-end training of severity classification and pathology segmentation with U-GAT increases

Dataset	Task	Architecture	AP	AUC	F1
KRI	ICU	MLP-Clinical	0.577 ± 0.109 [†]	0.654 ± 0.104 [†]	0.560 ± 0.107 [†]
KRI	ICU	ResNet18	0.670 ± 0.097	0.716 ± 0.077	0.560 ± 0.084 [†]
KRI	ICU	U-Net*+KNN	0.632 ± 0.113	0.677 ± 0.112	0.519 ± 0.131 [†]
KRI	ICU	U-Net*+MLP	0.615 ± 0.127 [†]	0.687 ± 0.128	0.612 ± 0.085
KRI	ICU	U-Net*+GraphSAGE	0.628 ± 0.114 [†]	0.690 ± 0.107 [†]	0.574 ± 0.085 [†]
KRI	ICU	ResNet18-GAT	0.637 ± 0.165	0.678 ± 0.160	0.595 ± 0.084 [†]
KRI	ICU	U-GAT*	0.672 ± 0.129	0.725 ± 0.107	0.651 ± 0.104
KRI	ICU + Seg.	U-GAT	0.699 ± 0.149	0.743 ± 0.103	0.661 ± 0.084
KRI	Ventilation	MLP-Clinical	0.527 ± 0.167	0.692 ± 0.109 [†]	0.475 ± 0.188
KRI	Ventilation	ResNet18	0.573 ± 0.127	0.715 ± 0.086 [†]	0.390 ± 0.160 [†]
KRI	Ventilation	U-Net*+KNN	0.527 ± 0.180 [†]	0.674 ± 0.112 [†]	0.368 ± 0.192 [†]
KRI	Ventilation	U-Net*+MLP	0.587 ± 0.183	0.741 ± 0.119	0.488 ± 0.134
KRI	Ventilation	U-Net*+GraphSAGE	0.603 ± 0.151	0.758 ± 0.109	0.481 ± 0.205
KRI	Ventilation	ResNet18-GAT	0.570 ± 0.152	0.689 ± 0.152 [†]	0.423 ± 0.178 [†]
KRI	Ventilation	U-GAT*	0.618 ± 0.137	0.788 ± 0.106	0.592 ± 0.130
KRI	Vent. + Seg.	U-GAT	0.644 ± 0.142	0.788 ± 0.112	0.539 ± 0.179
KRI	Mortality	MLP-Clinical	0.261 ± 0.135	0.544 ± 0.134	0.224 ± 0.152
KRI	Mortality	ResNet18	0.210 ± 0.116 [†]	0.461 ± 0.155 [†]	0.155 ± 0.138
KRI	Mortality	U-Net*+KNN	0.257 ± 0.137	0.512 ± 0.166	0.184 ± 0.147
KRI	Mortality	U-Net*+MLP	0.252 ± 0.157	0.502 ± 0.191	0.190 ± 0.157
KRI	Mortality	U-Net*+GraphSAGE	0.270 ± 0.143	0.568 ± 0.180	0.236 ± 0.163
KRI	Mortality	ResNet18-GAT	0.247 ± 0.151	0.520 ± 0.156	0.184 ± 0.157
KRI	Mortality	U-GAT*	0.271 ± 0.137	0.549 ± 0.188	0.230 ± 0.172
KRI	Mort. + Seg.	U-GAT	0.287 ± 0.186	0.586 ± 0.187	0.199 ± 0.173
iCTCF	Severity	MLP-Clinical	0.556 ± 0.099	0.735 ± 0.068	0.539 ± 0.064
iCTCF	Severity	ResNet18	0.525 ± 0.140	0.739 ± 0.083	0.513 ± 0.102
iCTCF	Severity	U-Net*+KNN	0.456 ± 0.070 [†]	0.705 ± 0.060	0.318 ± 0.129 [†]
iCTCF	Severity	U-GAT*	0.558 ± 0.102	0.740 ± 0.096	0.505 ± 0.114
iCTCF	Severity	U-GAT	0.593 ± 0.106	0.763 ± 0.085	0.521 ± 0.109

Table 4. Ablative testing and comparison with an MLP only using clinical data and a ResNet18 only using image data as input on all tasks. Highest values per task are in bold. U-GAT* refers to the proposed method using image and radiomic features extracted from frozen U-Net trained on the same annotations as the end-to-end U-GAT. Values marked with † indicate statistical significance with $p < 0.05$ based on the Wilcoxon's rank test comparing the proposed method with every other baseline.

the AP from 0.558 ± 0.102 to 0.593 ± 0.106 compared to U-GAT* that uses segmentations from a frozen U-Net trained on the same annotations.

Multitasking evaluation

We conducted additional experiments on the synergistic effects of multitasking segmentation with classification and the concurrent prediction of different patient outcomes since all of these tasks are interdependent. Results detailed in the supplementary material showed that classification can benefit from joint segmentation (Supplementary, Table S7) but mortality prediction was the only task that improved with the simultaneous prediction of all outcomes (supplementary material, Table S8).

U-GAT ensemble and comparison with Random Forest

As discussed in section “Ablative testing and comparison to baselines”, we also compare our method against Random Forests used in previous works to perform classification from fused tabular radiomics with clinical data. The comparison, shown in Table 5, illustrates the enhancement in U-GAT's average precision from 0.699 ± 0.149 to 0.745 ± 0.137 , elevating it to marginally outperform the Random Forest, which stands at 0.729 ± 0.089 . The results indicate that ensembling our method increases the robustness of our method to overfitting, showing comparable results as a Random Forest.

Interpretability and inter-patient graph attention

In addition to its performance boost over GraphSAGE, using GAT offers another important advantage. The attention mechanism of our model learns to identify the neighbors in the graph that are the most relevant for the prediction task, providing insight into the decision process of the model. The analysis of attention scores could suggest patients that the model deems relevant for the individual outcome prediction. These connections within

Architecture	AP	AUC	bACC	F1	Sens.	Spec.
RF-Clinical	0.635 ± 0.098	0.707 ± 0.086	0.624 ± 0.056	0.519 ± 0.070	0.475 ± 0.131	0.773 ± 0.175
U-Net*+RF	0.729 ± 0.089	0.774 ± 0.057	0.716 ± 0.075	0.649 ± 0.011	0.651 ± 0.177	0.781 ± 0.166
U-GAT ensemble	0.745 ± 0.137	0.770 ± 0.098	0.735 ± 0.111	0.700 ± 0.114	0.736 ± 0.067	0.734 ± 0.174

Table 5. Comparative analysis of ICU outcome prediction on the KRI dataset: U-GAT vs its cross-validation ensemble, a random forest model using only clinical data, and another random forest model incorporating all available tabular data, including radiomics extracted with a pretrained U-Net. Highest values are in bold.

the patient population graph can help uncover new information about a disease that is still poorly understood and provide valuable insights to physicians. Combined with the segmentation results, our attention mechanism allows the clinicians to thoroughly evaluate our model output and decision-making process, giving them potentially higher confidence in the prediction. For each of the two GAT layers, the model assigns attention scores to the neighbors of each node in the graph. These scores define how much the node representation after the layer will be based on the representation of its different one-hop neighbors. These attention scores can be thought of as a weighted directed adjacency matrix $A \in [0, 1]^{N \times N}$, where N is the number of nodes in the batch and all rows in A add up to 1. We can multiply the attention matrices of both layers to receive a matrix that shows how the representation of a node is based on its two-hop neighborhood, i.e., all nodes that are at most two edges away. These attention scores are visualized in Fig. 3. Our results on the test patient shown in Fig. 3 highlight that the attention mechanism succeeds in assigning high importance to neighbors of the same class and lower importance to those of the opposite class, thus implicitly refining the neighborhood constructed by the KNN algorithm. Furthermore, we can see that the attention mechanism does not necessarily assign high attention to neighbors that are particularly similar in their radiomic or clinical features. In contrast to a simple KNN classifier, which can only base its prediction on feature similarity, our method evidently can identify the most relevant neighbors that go beyond a simple correlation and are connected through more complex patterns and thus introduces orthogonal information to that embedded in the KNN graph.

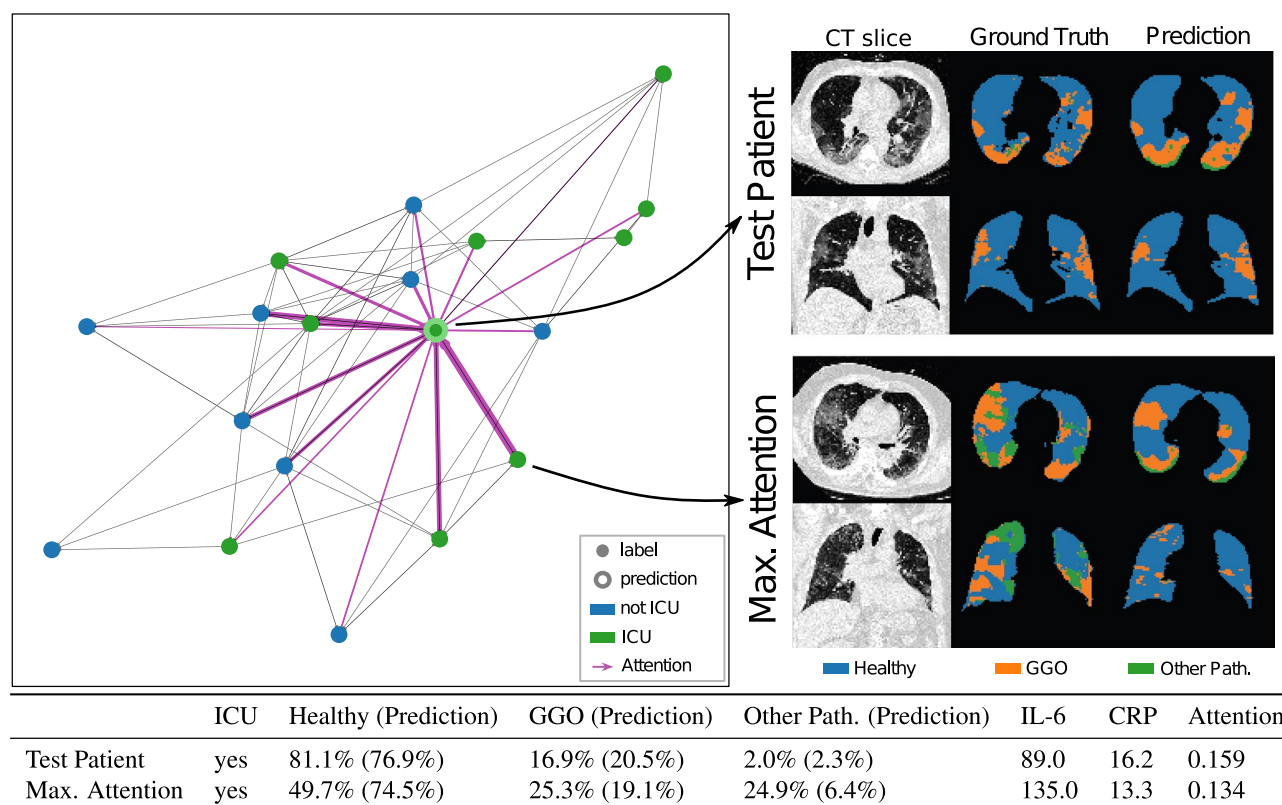


Figure 3. KRI dataset—Left: Batch graph showing the attention scores of a single test patient. The line's thickness corresponds to the respective neighbors' attention score after two hops. Right: CT images, segmentation ground truth, and predicted segmentation of a single axial and coronal slice from the test patient and the neighbor with maximum attention. Bottom: Most important features for the test patient and the neighbor with maximum attention. In brackets, the radiomics predicted by the pretrained U-Net are shown.

Challenges and future outlook

In a future iteration of our current model, the segmentation of infrequent lung pathologies, such as pleural effusion could be improved along with the prediction of imbalanced outcomes, notably mortality. Our approach to enhance the model involved constructing the population graph based on the mutual information of each feature. This has effectively improved the graph structure, and importantly, the features identified through this method are consistent with established radiological findings. It should be noted, however, that the mutual information displays a pronounced standard deviation and is notably lower for the mortality prediction task, indicating the inherent complexity of this specific prediction task and the potential sparsity of highly informative features given the available parameters in the dataset. In subsequent studies, these areas can be addressed by incorporating more annotated data and expanding the patient cohort, particularly the clinical data.

Conclusion

In this work, we developed and evaluated a method to effectively leverage multimodal information for the outcome prediction of COVID-19 patients. Here, the said information in the form of CT lung scans, clinical data, and radiomics was incorporated into a graph structure and processed within a GAT to stabilize and support the prediction based on data similarity. With U-GAT, we propose an end-to-end methodology that segments patient pathologies in medical images and uses a combination of imaging and non-imaging data to predict clinical outcomes. We explicitly incorporate automatically extracted lung radiomics in our architecture and demonstrate increased performance. We show that the auxiliary segmentation of COVID-19 pathologies indeed improves outcome prediction. To create the patient population graph, we propose a novel graph construction based on feature weighting utilizing mutual information, effectively clustering relevant patients. Our attention analysis imparts an additional layer of transparency, potentially increasing clinicians' confidence in our predictive approach. This added clarity can assist in identifying comparable patients from previous cases, thus informing and guiding the treatment trajectory for the current patient under consideration. This study underscores the potential of graph-based, data-driven strategies in improving patient care and decision-making in challenging clinical settings using multiple modalities.

Data availability

The iCTCF dataset²⁴ is publicly available and can be accessed at <https://ngdc.cncb.ac.cn/ictcf/>. The complete KRI dataset is available on request for research purposes in the frame of the BFS project AZ-1429-20C.

Received: 3 July 2023; Accepted: 3 November 2023

Published online: 09 November 2023

References

1. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *The Lancet* **395**, 470–473 (2020).
2. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: What next?. *Lancet (Lond., Engl.)* **395**, 1225–1228. [https://doi.org/10.1016/S0140-6736\(20\)30627-9](https://doi.org/10.1016/S0140-6736(20)30627-9) (2020).
3. Ryberg, J. Covid-19, triage decisions, and indirect ethics: A model for the re-evaluation of triage guidelines. *Ethics Med. Public Health* **17**, 100639 (2021).
4. Burian, E. *et al.* Intensive care risk estimation in covid-19 pneumonia based on clinical and imaging parameters: Experiences from the munich cohort. *J. Clin. Med.* **9**, 1514. <https://doi.org/10.3390/jcm9051514> (2020).
5. Parisot, S. *et al.* Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease. *Med. Image Anal.* **48**, 117–130 (2018).
6. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 234–241 (Springer, 2015).
7. Veličković, P. *et al.* Graph attention networks. *Int. Conf. Learn. Represent.* **2018**, 859 (2018).
8. Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18 (AAAI Press, 2018).
9. Wolf, T. N. *et al.* Daft: A universal module to interweave tabular data and 3d images in cnns. *NeuroImage* **260**, 119505 (2022).
10. Cai, G. *et al.* A multimodal transformer to fuse images and metadata for skin disease classification. *Vis. Comput.* **2022**, 1–13 (2022).
11. Taleb, A., Kirchlner, M., Monti, R. & Lippert, C. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 20908–20921 (2022).
12. Sudlow, C. L. M. *et al.* Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, 85 (2015).
13. Duanmu, H. *et al.* Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II* 23 242–252 (Springer, 2020).
14. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Dig. Med.* **3**, 136. <https://doi.org/10.1038/s41746-020-00341-z> (2020).
15. Chao, H. *et al.* Integrative analysis for covid-19 patient outcome prediction. *Med. Image Anal.* **67**, 101844. <https://doi.org/10.1016/j.media.2020.101844> (2021).
16. Tang, Z. *et al.* Severity assessment of covid-19 using ct image features and laboratory indices. *Phys. Med. Biol.* **66**, 74. <https://doi.org/10.1088/1361-6560/abbf9e> (2021).
17. Cai, W. *et al.* Ct quantification and machine-learning models for assessment of disease severity and prognosis of covid-19 patients. *Acad. Radiol.* **27**, 1665–1678. <https://doi.org/10.1016/j.acra.2020.09.004> (2020).
18. Xu, Q. *et al.* Ct-based rapid triage of covid-19 patients: Risk prediction and progression estimation of icu admission, mechanical ventilation, and death of hospitalized patients. In *medRxiv: Preprint Server for Health Sciences* 2020.11.04.20225797. <https://doi.org/10.1101/2020.11.04.20225797> (2020).

19. Jimenez-Solem, E. *et al.* Developing and validating covid-19 adverse outcome risk prediction models from a bi-national european cohort of 5594 patients. *Sci. Rep.* **11**, 14 (2021).
20. Homayounieh, F. *et al.* Computed tomography radiomics can predict disease severity and outcome in coronavirus disease 2019 pneumonia. *J. Comput. Assist. Tomogr.* **44**, 640–646. <https://doi.org/10.1097/RCT.0000000000001094> (2020).
21. Chassagnon, G. *et al.* Ai-driven quantification, staging and outcome prediction of covid-19 pneumonia. *Med. Image Anal.* **67**, 101860. <https://doi.org/10.1016/j.media.2020.101860> (2021).
22. Shiri, I. *et al.* Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest ct images in covid-19 patients. *Comput. Biol. Med.* **132**, 104304 (2021).
23. Gong, K. *et al.* A multi-center study of covid-19 patient prognosis using deep learning-based ct image analysis and electronic health records. *Eur. J. Radiol.* **139**, 109583 (2021).
24. Ning, W. *et al.* Open resource of clinical data from patients with pneumonia for the prediction of covid-19 outcomes via deep learning. *Nat. Biomed. Eng.* **4**, 1197–1207. <https://doi.org/10.1038/s41551-020-00633-5> (2020).
25. Tariq, A. *et al.* Patient-specific covid-19 resource utilization prediction using fusion ai model. *NPJ Dig. Med.* **4**, 1–9 (2021).
26. Parisot, S. *et al.* Spectral graph convolutions for population-based disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 177–185 (Springer, 2017).
27. Anirudh, R. & Thiagarajan, J. J. Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 3197–3201 (IEEE, 2019).
28. Cosmo, L., Kazi, A., Ahmadi, S.-A., Navab, N. & Bronstein, M. Latent-graph learning for disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 643–653 (Springer, 2020).
29. Soberanis-Mukul, R. D., Navab, N. & Albarqouni, S. Uncertainty-based graph convolutional networks for organ segmentation refinement. In *Medical Imaging with Deep Learning* 755–769 (PMLR, 2020).
30. Tian, Z. *et al.* Graph-convolutional-network-based interactive prostate segmentation in mr images. *Med. Phys.* **47**, 4164–4176 (2020).
31. Meng, Y. *et al.* Cnn-gcn aggregation enabled boundary regression for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 352–362 (Springer, 2020).
32. Wolterink, J. M., Leiner, T. & Išgum, I. Graph convolutional networks for coronary artery segmentation in cardiac ct angiography. In *International Workshop on Graph Learning in Medical Imaging* 62–69 (Springer, 2019).
33. Burwinkel, H. *et al.* Decision support for intoxication prediction using graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II* 23 633–642 (Springer, 2020).
34. Du, H., Feng, J. & Feng, M. Zoom in to where it matters: A hierarchical graph based model for mammogram analysis. [arXiv:1912.07517](https://arxiv.org/abs/1912.07517) (2019).
35. Burwinkel, H. *et al.* Adaptive Image-Feature Learning for Disease Classification Using Inductive Graph Networks. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, vol. 11769 of *Lecture Notes in Computer Science* (eds. Shen, D.) 640–648. https://doi.org/10.1007/978-3-030-32226-7_71 (Springer International Publishing, 2019).
36. Wang, S.-H., Govindaraj, V. V., Górriz, J. M., Zhang, X. & Zhang, Y.-D. Covid-19 classification by fgcn with deep feature fusion from graph convolutional network and convolutional neural network. *Inf. Fus.* **67**, 208–229 (2021).
37. Yu, X., Lu, S., Guo, L., Wang, S.-H. & Zhang, Y.-D. Resgnet-c: A graph convolutional neural network for detection of covid-19. *Neurocomputing* **2020**, 859 (2020).
38. Song, X. *et al.* Augmented multi-center graph convolutional network for covid-19 diagnosis. *IEEE Trans. Ind. Inform.* **2021**, 859 (2021).
39. Liang, X. *et al.* Diagnosis of covid-19 pneumonia based on graph convolutional network. *Front. Med.* **7**, 1071 (2021).
40. Saha, P. *et al.* Graphcovidnet: A graph neural network based model for detecting covid-19 from ct scans and x-rays of chest. *Sci. Rep.* **11**, 1–16 (2021).
41. Huang, H. *et al.* Graph-based pyramid global context reasoning with a saliency-aware projection for covid-19 lung infections segmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1050–1054 (IEEE, 2021).
42. Di, D. *et al.* Hypergraph learning for identification of covid-19 with ct imaging. *Med. Image Anal.* **68**, 101910. <https://doi.org/10.1016/j.media.2020.101910> (2021).
43. Colombi, D. *et al.* Well-aerated lung on admitting chest ct to predict adverse outcome in covid-19 pneumonia. *Radiology* **296**, E86–E96. <https://doi.org/10.1148/radiol.2020201433> (2020).
44. Wang, D. *et al.* Study on the prognosis predictive model of covid-19 patients based on ct radiomics. *Sci. Rep.* **11**, 1–9 (2021).
45. Yang, X. *et al.* A novel multi-task deep learning model for skin lesion segmentation and classification. [arXiv:1703.01025](https://arxiv.org/abs/1703.01025) (2017).
46. Mehta, S. *et al.* Y-net: Joint segmentation and classification for diagnosis of breast biopsy images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 893–901 (Springer, 2018).
47. Le, T.-L.-T., Thome, N., Bernard, S., Bismuth, V. & Patoureaux, F. Multitask classification and segmentation for cancer diagnosis in mammography. [arXiv:1909.05397](https://arxiv.org/abs/1909.05397) (2019).
48. Alom, M. Z., Rahman, M. M. S., Nasrin, M. S., Taha, T. M. & Asari, V. K. Covid_mtnet: Covid-19 detection with multi-task deep learning approaches. [ArXiv 2020](https://arxiv.org/abs/2005.08599), 859 (2020).
49. Wu, Y.-H. *et al.* Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Trans. Image Process.* **30**, 3113–3126 (2021).
50. Amyar, A., Modzelewski, R., Li, H. & Ruan, S. Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation. *Comput. Biol. Med.* **126**, 104037. <https://doi.org/10.1016/j.compbiomed.2020.104037> (2020).
51. Gao, K. *et al.* Dual-branch combination network (dcn): Towards accurate diagnosis and lesion segmentation of covid-19 using ct images. *Med. Image Anal.* **67**, 101836. <https://doi.org/10.1016/j.media.2020.101836> (2021).
52. Bao, G. & Wang, X. Covid-mtl: Multitask learning with shift3d and random-weighted loss for diagnosis and severity assessment of covid-19. [ArXiv 2012](https://arxiv.org/abs/2012.08599), 85 (2020).
53. He, K. *et al.* Synergistic learning of lung lobe segmentation and hierarchical multi-instance classification for automated severity assessment of covid-19 in ct images. *Pattern Recogn.* **113**, 107828 (2021).
54. Goncharov, M. *et al.* Ct-based covid-19 triage: Deep multitask learning improves joint identification and severity quantification. *Med. Image Anal.* **71**, 102054 (2021).
55. Näppi, J. J. *et al.* U-survival for prognostic prediction of disease progression and mortality of patients with covid-19. *Sci. Rep.* **11**, 1–11 (2021).
56. Ross, B. C. Mutual information between discrete and continuous data sets. *PLoS one* **9**, e87357 (2014).
57. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 1025–1035 (2017).
58. Kim, S. T. *et al.* Longitudinal quantitative assessment of covid-19 infection progression from chest cts. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII* 24 273–282 (Springer, 2021).
59. Goksel, O. *et al.* Overview of the visceral challenge at isbi 2015. In *VISCERAL Challenge@ISBI* (2015).

60. Yang, J. *et al.* Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. *Med. Phys.* **45**, 4568–4581 (2018).
61. Rudyanto, R. D. *et al.* Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study. *Med. Image Anal.* **18**(7), 1217–32 (2014).
62. Hofmanninger, J. *et al.* Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *Eur. Radiol. Exp.* **4**, 50 (2020).
63. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
64. Roth, H. R. *et al.* Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Res. Square* **2021**, 74 (2021).
65. Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125–127 (1974).
66. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
67. Fey, M. & Lenssen, J. E. Fast graph representation learning with pytorch geometric. [arXiv:1903.02428](https://arxiv.org/abs/1903.02428) (2019).
68. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* 565–571 (IEEE, 2016).
69. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
70. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
71. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
72. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

Acknowledgements

The authors would like to thank the team at the Radiology Department at Klinikum rechts der Isar, particularly Matthias Zierhut, for his help annotating CTs and Friedericke Jungmann for her help collecting patient metadata. Further, the authors acknowledge the critical views and feedback provided by Anees Kazi, Roger Soberanis-Mukul from the Chair for Computer Aided Medical Procedures and Augmented Reality at Technical University Munich and Gerome Vivar and Ahmad Ahmadi from the German Center for Vertigo and Balance Disorders at Ludwig-Maximilians Universität München. The CT annotation and segmentation work was partially funded by EIT Health's rapid response project 20882 "FastRAi". The developments on Graph Convolutional Networks were covered by the Bavarian Research Foundation (BFS) grant AZ-1429-20C.

Author contributions

M.K., H.B., and D.B.-H. conceptualized, designed, and implemented the proposed methodology. M.K., H.B., D.B.-H., M.P., T.C., N.N., and T.W. designed and evaluated all technical experiments. M.P., T.C., N.N., and T.W. consulted M.K., H.B., and D.B.-H. on the design of the proposed model. E.B., M.R.M., R.B., and T.W. collected and curated the KRI dataset, defined the clinical goals, and performed the clinical evaluation. M.K., H.B., D.B.-H., M.P., T.C., and T.W. wrote different parts of the manuscript. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-46625-8>.

Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023