

RESEARCH

Open Access



Visual transformer and deep CNN prediction of high-risk COVID-19 infected patients using fusion of CT images and clinical data

Sara Saberi Moghadam Tehrani^{1†}, Maral Zarvani^{1†}, Paria Amiri², Zahra Ghods¹, Masoomeh Raoufi³, Seyed Amir Ahmad Safavi-Naini⁴, Amirali Soheili⁵, Mohammad Gharib⁶ and Hamid Abbasi^{7*}

Abstract

Background Despite the globally reducing hospitalization rates and the much lower risks of Covid-19 mortality, accurate diagnosis of the infection stage and prediction of outcomes are clinically of interest. Advanced current technology can facilitate automating the process and help identifying those who are at higher risks of developing severe illness. This work explores and represents deep-learning-based schemes for predicting clinical outcomes in Covid-19 infected patients, using Visual Transformer and Convolutional Neural Networks (CNNs), fed with 3D data fusion of CT scan images and patients' clinical data.

Methods We report on the efficiency of Video Swin Transformers and several CNN models fed with fusion datasets and CT scans only vs. a set of conventional classifiers fed with patients' clinical data only. A relatively large clinical dataset from 380 Covid-19 diagnosed patients was used to train/test the models.

Results Results show that the 3D Video Swin Transformers fed with the fusion datasets of 64 sectional CT scans + 67 clinical labels outperformed all other approaches for predicting outcomes in Covid-19-infected patients amongst all techniques (i.e., TPR = 0.95, FPR = 0.40, F0.5 score = 0.82, AUC = 0.77, Kappa = 0.6).

Conclusions We demonstrate how the utility of our proposed novel 3D data fusion approach through concatenating CT scan images with patients' clinical data can remarkably improve the performance of the models in predicting Covid-19 infection outcomes.

Significance Findings indicate possibilities of predicting the severity of outcome using patients' CT images and clinical data collected at the time of admission to hospital.

Keywords Deep learning, Visual transformer, Predictive models, Convolutional neural network (CNN), Covid-19 detection, CT scan, Clinical data, Data fusion

[†]Sara Saberi Moghadam Tehrani and Maral Zarvani are joint first author.

*Correspondence:

Hamid Abbasi

h.abbasi@auckland.ac.nz

Full list of author information is available at the end of the article



Introduction

In the late 2019, Covid-19 pandemic was initially reported to rapidly infect residents of Wuhan city in China [1]. This previously unknown virus was then labelled as SARS-CoV2 by the International Committee on Taxonomy of Viruses (ICTV) and categorized under the family of corona viruses [2]. The infection caused by the Covid-19 was reported to be very similar to the disease due to the infection by SARS virus and could lead to severe respiratory syndromes and death [3, 4]. The fast and large increase in the number of infected individuals before vaccine roll-outs had resulted in a large increase in the number of referrals with critical conditions and admittance to the hospitals and clinics, imposing a burden on the healthcare sector, globally. This important factor could potentially result in an increase in critical human error that could lower the diagnosis accuracy, subsequently. Recent analytical enhancements could assist in finding practical solutions to the urgent need for developing automated diagnosis platforms that can provide prognostic information about the evolution of infection in patients. Clinical observations confirm a large variety of symptoms for the infected individuals, where the milder initial symptoms could rapidly develop to critical situations. This itself could limit the clinical assessments or in more severe cases can eliminate the chances of treatment [5]. Therefore, clinical monitoring of patients and accurate prediction of infection development during this period and/or even before their initial referrals can play an important role in saving lives [6]. Research suggest that the quality of patients' chest Computerized Tomography (CT) scans are interpretably linked to other observations from patients including their clinical examinations, laboratory tests, vital signals, patient history, and potential background illnesses [7]. Therefore, it is hypothesized that a proper combination of these data could be used for automatic prediction of both the severity and the developmental stage of the infection, more accurately [8].

Various applications of multi-modal data fusion techniques on Covid datasets have been addressed in the literature. Studies suggest that chest X-ray images and lung CT scans can be fed into deep-learning-based models for diagnosis and classification of Covid-19-related conditions [9–12]. Access to larger clinical datasets is currently a major challenge in the implementation of these techniques. Thus, various research have considered data augmentation techniques to cover these drawbacks [13–15]. Attempts show that predictive models fed with patients' clinical data, demographic/historical conditions and disorders, as well as laboratory tests can be used to predict outcomes [15–20]. Literature indicates possibilities of developing high-performance algorithms

to accurately predict the severity of infection and further diagnose healthy individuals from tested-positive cases. Successful algorithms have used combinational approaches through fusing clinical observations data, CT images, vital signals, and background/historical conditions [8, 17, 21–23]. These studies have initially combined features extracted from CT images with features from the patients' clinical data and fed the outputs into deep-net classifiers. For instance, studies show that the extracted features from the images can be combined with other available features/data (e.g., clinical observations/measures) to create a more robust and consistent dataset that can provide detailed information for the deep-net to predict the severity of infection in the high- and low-risk patients [8, 14, 24].

In this work, we use data fusion of lung CT scan images and clinical data from a total of 380 Iranian Covid-19-positive patients to develop deep-learning-based models to predict risk of mortality and outcomes in the high- vs. low-risk Covid-19 infected individuals. An overall schematic of the proposed approaches in this work is shown in Fig. 1. The article contributes to the field through:

1. Developing Visual Transformer and 3D Convolutional Neural Network (CNN) predictive models fed with a series of fusion datasets from patients' CT images and their clinical data. This includes introducing a novel heuristic concatenation approach, for integrating CT scan images with clinical data, which is inferred to have assisted with inter-network feature aggregations in the Transformer models.
2. Developing Visual Transformer and CNN-based predictive models fed with CT scan images only, and assessing the capabilities of genetic algorithm (GA) for hyper-parameter tuning of the 3D-CNN models fed with the fusion data and CT scan images.
3. Evaluating a series of conventional classifiers for predicting outcomes using patients' clinical data only, and investigating strategies to select a set of proper clinical labels from the pool of clinical data for the classification of imbalance data. The paper further discusses imputation techniques to deal with missing values in the dataset.

Related work

Clinical data-based detection

Here, only patients' clinical data, including patients' history and their lab test results, are used to develop predictive models. Yue et al. have demonstrated that the use of clinical data and patients' condition assessments at the time of admission can help to predict chances of mortality

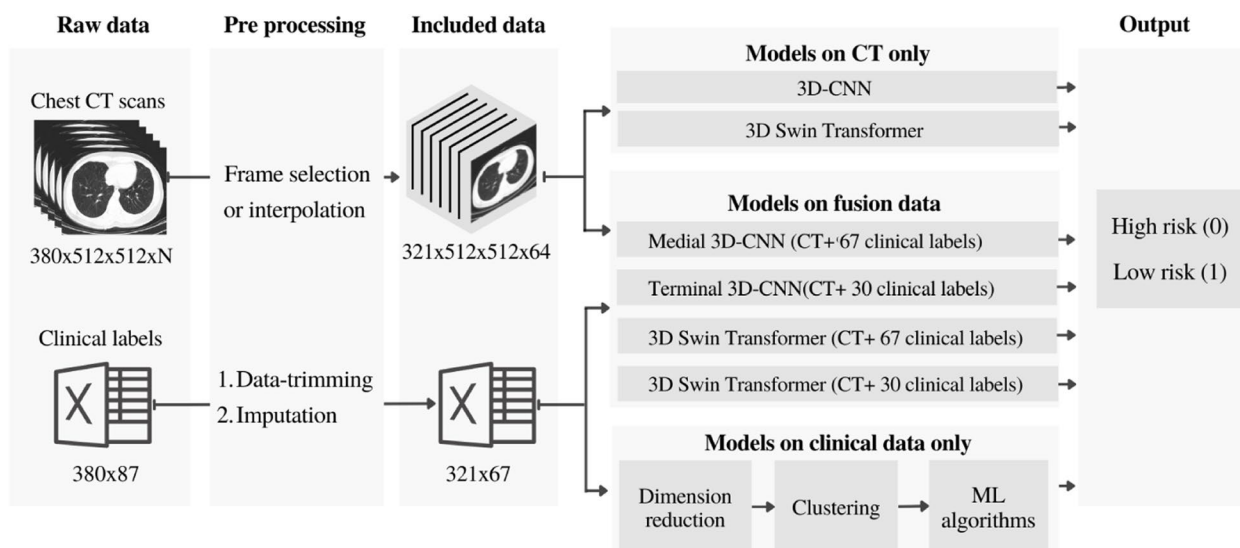


Fig. 1 The flow-chart schematic of the proposed predictive machine-learning approaches for the classification of high- and low-risk Covid-19 infected patients. “N” denotes the number of total CT scan slices from each patient

at around 20 days [14]. They have achieved promising results by integrating predictive models including logistic regression (LR), support vector machine (SVM), gradient boosted decision tree (GBDT), and neural networks (NN) to predict the mortality risk (AUC: 0.924–0.976) [14]. Dhruv et al. have also shown that patients’ clinical data, blood panel profiles, and socio-demographic data can be fed into conventional classification algorithms such as Extra Tree, gradient boosting, and random forest for predicting the severity of Covid-19 [15]. Similar works show that clinical parameters in the blood samples can be infused into a combined statistical analysis and deep-learning model to predict severity of Covid-19 symptoms and classify healthy individuals from tested-positive cases [17].

Image-based detection

In this approach, only chest X-ray or CT scan images are used for classification of Covid-19 infected patients. Purohit et al. have proposed an image-based Covid-19 classification algorithm and demonstrated that, among various image sharpening techniques, utilization of certain sharpening filters such as canny, sobel, texton gradient and their combinations can help to increase training accuracy in multi-image augmented CNN [13]. Research shows that deep neural networks are able to automatically diagnose Covid-19 infection in partial X-ray images of the lungs [25], or through fusing deep features of CT images [26–28]. Our team has also previously shown that chest X-ray images can be fed into CNNs for Covid detection [29].

Visual Transformer (ViT) networks, along with the CNN models, have recently shown remarkable capability in resulting higher performances in various applications, such as image classification, object detection, and semantic segmentation. Recent works show that ViT and in particular Video Swin Transformers can competitively achieve better accuracies, compared to the CNN-based methods, for the classification and identification of Covid-19 infected patients using chest CT scans [30] and X-ray images [31]. Research shows that the feature maps extracted from the CT scan images in the output of a ResNet model can be used as inputs to a transformer model for the identification of Covid patients (~1934 images, >1000 patients, recall accuracy 0.93) [32]. Transfer-learning in Visual Transformer models, fed with either CT images or their combinations with chest X-ray images, shows diagnostic possibilities of Covid-19 patients and localization of the infected regions in the lungs [31, 33]. A recent work has shown that a combination of parallelly extracted features from CT scans through simultaneous application of Visual Transformers and CNN can help to accurately classify Covid-19 patients [34]. Fan et al. have reported a high recall performance of 0.96 using 194,922 images from 3745 patients which suggests strong capabilities of combinational approaches [34].

Fusion-based detection

This approach mainly aims to fuse patients’ clinical data with any other possible information, such as chest X-ray and/or CT images, to use as the inputs for predictive

models. Using a relatively large CT image dataset from multiple institutions across three continents, Gong et al. have developed a deep-learning-based image processing approach for diagnosis of Covid-19 lung infection [21]. In their technique, a deep-learning model initially segments lung infected regions by extracting total opacity ratio and consolidation ratio parameters from CT images and then combines the outputs with clinical and laboratory data for prognosis purposes using a generalized linear model technique (reported AUC range: 0.85–0.93) [21]. Other studies have proposed robust 3D CNN predictive models fed with combined data from segmented CT images and patients' clinical data to predict whether a Covid-19-infected-individual belongs to the low- or high-risk group [8, 22, 23]. These studies have shown that their proposed approaches are independent of demographic information such as age and sex, and other conditions such as chronic diseases. Meng et al., have demonstrated that 3D-CNNs can perform much better when simultaneously fed with patients' segmented CT scans and clinical data compared to the singular use of clinical data or CT images in CNN-based or logistic regression models [22]. Ho et al., have compared performances of three 3D CNNs where each was trained on the (1) raw CT images, (2) segmented CT images, and (3) on the long lesion segmented data. They reported higher performance from the last approach amongst all [23].

A recent study has initially trained a speech identification model for Covid diagnosis using Long Short-Term Memory Networks (LSTM) that uses the acoustic aspects of patient's voice, their breathing data, coughing patterns, and talking [35]. The patients' chest X-ray images are also fed into general deep-net models, including a VGG16, a VGG19, a Densnet201, a ResNet50, a Inceptionv3, a InceptionResNetV2, and a Xception for Covid identification. Images and audio features were then combined and used as inputs to a hybrid model to identify non-Covid or Covid-positive patients. They have reported a lower accuracy for their hybrid model compared to their speech-based or X-ray image-based models [35].

Methods

Dataset

The dataset used in this research includes both lung CT scan images and their clinical data from a total of 380 Covid-19 infected patients. Patients were diagnosed by clinicians according to the Iran's National Health guidelines [36] through clinical assessments of their symptoms and lung CT images. The patients were hospitalized in the emergency unit at Imam Hussain Hospital, Tehran, Iran, between 22nd Feb 2020 to 22nd March 2020. All ethics of the current research have been approved by the Shahid Beheshti University's ethics committee (Ref:

IR.SBMU.RETECH.REC.1399.003). All patients have signed and submitted their informed consent to participate in the research and their data privacy has been fully considered [37]. Examples of the lung CT scans at different slice locations from a high- and a low-risk patient are shown in Fig. 2A, B and C, D, respectively.

From the total number of studied patients, 318 individuals have recovered from the illness while 62 have died. Since our top goal in this research was to correctly predict the severity of outcomes (mortality risk) using data collected at or around the time of initial referral (the first examination after the initial admission), we categorized died patients (including ICU-hospitalized deaths) in the high-risk group (class 0) and labeled the recovered individuals as the low-risk class (class 1). As data collection at the initial admission often results in 'data missing,' in the following sections, we will provide thorough explanations of our strategies for addressing missing data.

Our image datasets included a series of lung CT scans ranging between 50 and 70 images/slices depending on the length of the patient's lung. The clinical data used in this research were used as sets of numerical data collected for all patients, including demographic data, exposure history, background illness or comorbid diseases, symptoms, presenting vital signs, and laboratory tests data. A full list of these parameters as well as their mean and standard deviation (mean \pm std) are tabulated in Table A.1 of Appendix A.

Data pre-processing

An optimal data pre-processing is a critical initial step, prior to the initiation of training process, with possible boosting impacts on the overall performance of a model. A variety of pre-processing strategies can be chosen based on the type of data and/or algorithms used. In the following, we detail our pre-processing approaches for the numerical datasets and CT images.

Pre-processing of clinical data

Dealing with clinical data often presents a range of unique challenges, some of which include handling missing data which can require strategic imputations or conversion of qualitative measurements into numerical formats. In fact, one of the primary challenges in working with clinical data is addressing missing values. Missing data can occur for various reasons, including non-response from patients, incomplete records, or data entry errors. Addressing these challenges requires careful consideration and the implementation of specific strategies.

On the other hand, data often holds high dimensionality and strategies such as dimension reduction (data/

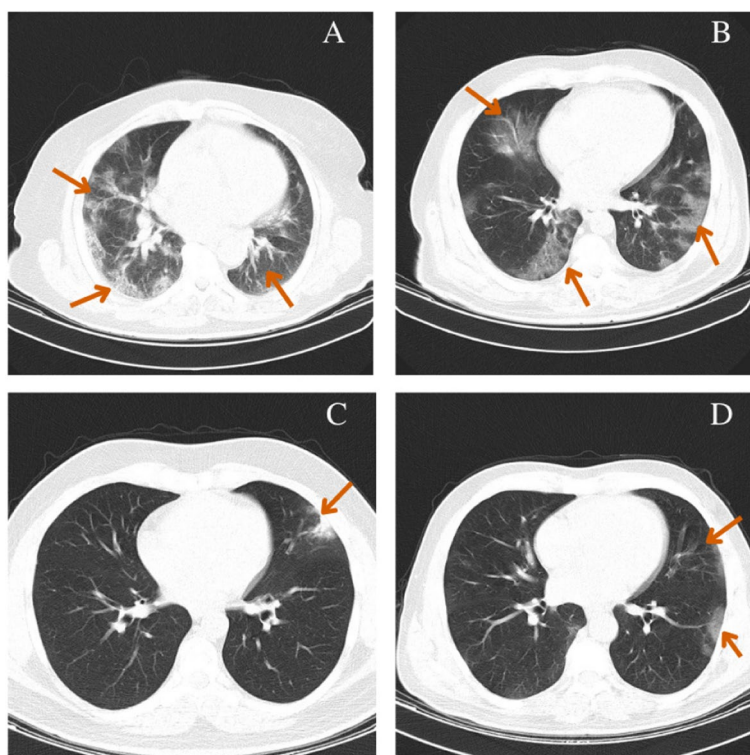


Fig. 2 Examples of the lung CT scans from the sequence of slices in the high- (A, B) and low-risk (C, D) Covid-19 infected patients. Arrows indicate infected regions

feature/label selection) and feature extraction can help to represent data in a simpler format.

In this work, we initially adopted the widely accepted one-hot encoding [38] approach to convert qualitative data such as gender, etc., into numerical representations. This transformation allows us to incorporate these essential variables into our analytical models effectively. In the following, we explain our strategies for clinical data trimming, data imputation, and preparation for analysis.

Clinical data trimming In the process of preparing our clinical dataset for analysis, we employed a series of meticulous data trimming strategies to enhance the quality and relevance of the data. These strategies aimed to strike a balance between maintaining a comprehensive dataset and ensuring data clarity and robustness.

Our initial data trimming step involved applying a thresholding criterion to identify and remove patients with a substantial amount of missing clinical data. Specifically, patients with at least 55% of their clinical data missing were excluded from the original clinical dataset. This step resulted in the removal of 59 patients, ensuring that our dataset primarily consisted of individuals with relatively complete clinical records. To further enhance

the dataset's robustness, we extended our data trimming efforts by identifying and removing clinical labels with at least 60% missing values across the entire dataset. This step eliminated 20 labels (out of a total of 87), streamlining our dataset and focusing our analysis on the most informative and complete clinical variables (i.e., 67 labels). These specific threshold values were chosen based on manual assessments and observations.

After implementing these data trimming strategies, our dataset was refined to contain 321 patients out of the initial 380 patients. To facilitate subsequent analyses, we categorized patients into two classes based on clinical outcomes. Patients who had unfortunately passed away were classified as "high-risk" ($n=57$, class 0), while the remaining participants were labeled as "low-risk" ($n=264$, class 1). These detailed data trimming strategies were essential in shaping our dataset to align with our research goals, ensuring that the resulting data analysis is meaningful.

Imputation of missing values The intensive work-load of clinical staff or unforeseen emergency situations/reasons may result in gaps or missing values within patients' data records. It's worth noting that working with the initial

format of imputed datasets can be challenging, making it crucial to employ appropriate algorithms to handle missing data [39]. In this work, to ensure the completeness of our clinical datasets, we employ imputation techniques to address these gaps effectively. Imputation involves estimating missing values based on the available data, ensuring that our datasets remain comprehensive and suitable for analysis. In the realm of machine learning, several methods can be applied for imputing missing data, and the choice often depends on the nature of the data and the research objectives. For instance, research shows that statistical approaches can be used to estimate missing data by utilizing key statistical parameters like mean and median derived from the entire dataset. Additionally, other machine-learning techniques such as linear regression or k-nearest neighbor (KNN) can be used to provide robust estimations for the missing values [40–44]. Here, for our specific implementation, we opted for the KNN algorithm with a parameter of $k=5$. This approach allowed us to impute missing values by considering the closest neighboring data points, enhancing the precision of our imputations. By employing these methods, we aimed to ensure that our clinical dataset remained comprehensive and robust, enabling us to conduct analyses and draw meaningful conclusions in the presence of missing data.

Dimension reduction Clinical datasets can exhibit high dimensionality due to the multitude of variables and features. Managing high-dimensional data can be challenging and may lead to issues such as overfitting. To mitigate these challenges and simplify the data representation, we employ dimension reduction techniques. Generally, dimension reduction is performed via feature/label selection or feature extraction operations. Feature/label selection approaches are mainly concerned with distinguishing the most dominant features/labels while feature extraction strategies are employed to transfer data values into a new domain and sometimes define novel features based on the original ones. In this research, we investigated the impact of both approaches on the feature-sets and assessed outcomes for each, both visually and by implementing a set of conventional classifiers, explained in the following, to identify the most informative and relevant variables, reducing the dimensionality of our dataset while retaining critical information. The final extracted features as well as the selected clinical labels from these attempts were later used in the training process. In this study, we often refer to the selected clinical data as “clinical labels”.

Feature extraction: here, we extracted features from the clinical data by utilizing a commonly used dimension

reduction technique, namely called principal component analysis (PCA) [45, 46]. PCA is an unsupervised and linear technique that uses eigen-vectors and eigen-values from a matrix of features to project lower dimensions from higher feature dimensions in the original matrix [47]. In the current study, an optimal number of required components in the PCA was found by using various numbers of extracted features. The output datasets from PCA were then fed into seven conventional classifiers including, SVM [38], MLP [46], KNN [47], random forest [48], gradient boosting [49], Gaussian naïve bayes [50], and XGBoost [51] to assess which number of feature-sets could lead to an optimal performance. This was accordingly found to be associated with a set of 25 components.

Feature/label selection: here we assessed the capabilities of two different approaches, namely “SelectKBest” [48, 49] and decision tree-based ensemble learning algorithms [50] to select a set of clinical labels from the pool of original clinical data. The SelectKBest algorithm uses statistical measures to score input features based on their relation to outputs and chooses the most effective features, accordingly. We used an ExtraTree classifier [51] for the decision tree-based ensemble learning approach where the algorithm randomly selects subsets of features to create the associated decision trees and evaluates minimal mathematical measures of each feature (typically the Gini Index [52]), while making the forest. Finally, all the extracted features are sorted in a descending order based on their measured Gini Index and user can choose to work with an arbitrary top k number of dominant features from the list. An optimal number of clinical labels was found by assessing the performance optimality of the aforementioned seven conventional classifiers across a set of various numbers for the SelectKBest algorithm and ExtraTree classifier. A set of 13 selected clinical labels from the SelectKBest and a set of 30 selected clinical labels from the ExtraTree classifier were found to result in better performances compared with other combination sets (see Appendix B).

We further visually assessed the selected features using an unsupervised non-linear technique based on manifold learning, called t-distributed stochastic neighbor embedding (t-SNE) [53]. The t-SNE is conventionally used for data visualization of large dimension datasets in 2 or 3 dimensions. t-SNE aims to find an optimized value for its cost function by measuring embedded similarities within the dataset at both higher- and lower- dimensions representations. In the t-SNE approach, a more visually separable data represents less complexity. Here, the t-SNE’s dimensionality parameter was set at 3 dimensions, and then applied to the (1) main dataset including

all 67 clinical labels (with no dimension reduction), (2) a dataset including 13 selected clinical labels from SelectKBest, 2) a dataset including 30 selected clinical labels from ExtraTree classifier, and 4) a dataset including 25 extracted features from the PCA. Outputs of the t-SNE were then scatter plotted to visualize the complexity within each dataset (see Results section for the plots). Finally, we chose to carry on with the set of 30 selected clinical labels from the ExtraTree classifier approach which were found to lead to better classification results and represented less visual complexity in the t-SNE approach.

Pre-processing of CT images

CT scan images of lung consist of a sequence of video frames at various sections (slices) along the patient's lung, where the number of frames varies in individuals according to their length of lung or device settings. These images can be used as the inputs for predictive/classification models where a certain number of input channels, that are compatible with the number of slices, must be used in the network's architecture. Since the number of CT video frames varies across patients, an appropriate slice selection approach should be used to shape a uniform volumetric 3D input size for consistency across all models [54]. Various slice selection strategies consider manual selection of frames from the beginning, middle and end of a video set. The major problem with such approaches is that they neglect information connectivity across slices which can lead to losing localized information and provide a false representation for the entire video set. On the other hand, there are strategies that initially select a fixed number of frames from the entire video, and then interpolate data to generate a desired set of frames that provides a more accurate representation of the whole video set [54] compared to the manual approach.

Here, we chose to work with Spline Interpolated Zoom (SIZ) frame selection technique. An arbitrary number of frames (N) is initially selected in this technique to construct a volumetrically uniform image-set that consists of a fixed number of CT slices for all individuals [54]. Then, depending on whether the patient's video set contained higher or lower number of slices compared to the N , the sequence of slices was evenly sampled using a spacing factor or interpolated to construct the missing slices, respectively. Here, the original size of the gray scale CT images was provided as $512 \times 512 \times 1$, and we chose to work with $N=64$ that represents the average number of frames in the videos from all patients. Therefore, the volumetrically uniform 3D inputs of the CNNs were reshaped to the size of $512 \times 512 \times 64$ for all patients.

Data fusion

Here, we combined the clinical data/measures with the 3D videos of the CT scan images from previous section, 'pre-processing of CT images', to shape more detailed fusion datasets for each patient.

To do so, we initially mapped the array of clinical labels ($1 \times N$) to a normalized vector with scaled numerical values ranging from 0 to 255. Next, we generated an empty 2D matrix with dimensions matching the size of 2D CT video-frames (i.e., 512×512). To organize the data correctly, we created a 3D matrix of size $512 \times 512 \times N$ by replicating this 2D matrix N times, with N representing the number of clinical labels. Subsequently, all data arrays in each 2D matrix within the 3D matrix were populated with the scaled values corresponding to the associated clinical labels/measures from the previous step. This approach helped to convert each of the scaled value from each clinical label into a 2D matrix (image). These 2D matrices were then concatenated to form a 3D matrix of clinical data, measuring $512 \times 512 \times N$ (i.e., N images of size 512×512). This 3D matrix was then added to the CT video (i.e., 64 image slices of size 512×512 forming a 3D matrix of $512 \times 512 \times 64$ frames) to form a final 3D fusion dataset measuring $512 \times 512 \times (64 + N)$ frames. This dataset was then used as inputs to the model described in '3D swin transformer models on fusion data' section, with N values varying - once with $N=30$ (suggested from 'Dimension reduction' section) and once with $N=67$.

Model training

The training process for each of the four previously outlined models, in the Introduction section and Fig. 1, are described in the following:

Approach #1: classification using clinical data only

Classification of datasets with imbalanced classes is associated with challenges and complexities which requires careful considerations. Data clustering is one of the useful approaches to handle such complexities and create more balanced datasets [55]. Here, we considered the following steps to create balanced datasets for training. Data were initially split into train and test sets (80% training, 20% test). The original ratio between class 1 and 0 in the clinical dataset is nearly 5 (imbalanced data), hence we used Gaussian mixture clustering algorithm [56] to divide the low-risk class in the training sets ($n=264$, class 1) into five different clusters. Each of these clusters were then combined with data from the high-risk class ($n=57$, class 0). This approach helped to create 5 separate balanced datasets which were then fed into seven different conventional classification algorithms, namely the SVM, MLP, KNN, random forest, gradient boosting, Gaussian naïve bayes, and XGBoost for classification. Each classification algorithm was accordingly

trained and tested on the five balanced train/test datasets resulting in five classification measures for classifier (e.g., 5x trained/tested random forests). A voting approach was then applied to the outputs of these five blocks to determine the winning class. The class (e.g., 0 or 1) with a larger number of votes (i.e., 3, 4, 5) from all blocks were chosen as the winning class.

Approach #2: training on CT images only

3D-CNN CT model Since the CT scan images are sequences of frames taken at different slices, therefore, they can be technically considered as 3D video data. Therefore, here we designed and trained a 3D-CNN with 3 convolutional layers on the training datasets. Here, inputs of the 3D-CNN classifiers are matrices of $512 \times 512 \times 64$ dimension from the pre-processing stage, where 64 is the number of CT scan frames (slices) for each patient. We further used Genetic Algorithm (GA) to automatically find and assign optimal values for the CNNs' hyperparameters [57]. The specified hyperparameters included the number of layers, number of neurons in each layer, learning-rate, optimization function, dropout size, and kernel size. The population size and the number of generation were set to 10 and 5, respectively. We also used Roulette wheel algorithm for parent selection followed by crossover mechanism. The GA were trained over 20 epochs and fitness values with lower FPRs were chosen, accordingly. The selected hyperparameters for the 3D-CNN-based models in this work are shown in Table C.1 in Appendix C.

3D swin transformer CT model Visual Transformers (ViT) are classes of deep neural networks that have been initially used for natural language processing (NLP) and sought as improved alternatives to other classes of deep-nets (i.e., CNNs) with competitive performances for multimodal inputs [58]. An input image to a ViT is initially shaped as a set of image-patches (equivalent to the set of words in NLP) which is then embedded with the localized information of the image to form inputs to an encoder network within the Transformer. The encoder unit consists of a multi-head self-attention layer [59], which highly improves features learning such as long-range dependencies and aggregation of global information [60]. The multi-head self-attention layer therefore aggregates spatial locations' information where global and local information are combined, accordingly. This operation is expected to help with inter-network feature extractions and lead to better outcomes compared to the CNN networks, where the receptive field sizes are fixed [61]. In CNNs, this can be equivalently achieved by increasing convolutional kernels which largely increases the computations. While ViT

generally require 4 times lower computational facilities, they can extraordinarily outperform the ordinary CNNs if trained on satisfactory data. Transformers generally require larger datasets for training, where transfer-learning and self-supervised techniques could greatly help to largely overcome such challenges. On the other hand, high-resolution input images can increase the computational burden and lower the computational speed, subsequently. To overcome this challenge, Swin Transformer models have been introduced to deal with higher resolution data in computer vision applications [62]. Video Swin Transformer (VST) models have been further introduced to work with 3D datasets such as videos [63], where the application of transfer learning and pre-trained models have been helpful. In this work, we normalized and augmented the CT scan images of each patient from 'pre-processing of CT images' section to form 3D inputs for a 3D Swin Transformer. With the aid of transfer-learning, we used a pre-trained model, Kinetics-400 [63], to set our model's initial weights. We trained the 3D Swin Transformer over 50 epochs using an Adam optimizer with a 0.02 learning rate and 0.02 weight decay.

Approach #3: training on fusion data

3D-CNN models on fusion data Here, we considered a terminal data fusion (on CTs + 30 clinical labels) and a medial data fusion approach (on CTs + 67 clinical labels) to combine the data. In the first approach, the terminal 3D-CNN, CT scans are initially fed into the 3D-CNN to extract their features-vector. The output features-vector were then terminally combined with the numerical data from dimension-reduction stage including 30 selected clinical labels for each patient to shape the final features-vector. The final features-vectors along with the labels were fed into the Naïve Bayes network [64] for training/classification. The schematic of this approach is shown in Fig. 3.

In the medial 3D-CNN approach, the extracted features-vectors of CT scans using the 3D-CNN in the previous structure were medially combined with extracted features from the original clinical data (67 clinical labels from each patient) using a 1D-CNN to create a more comprehensive features-set (Fig. 4). Two fully-connected layers were finally used at the end of this structure and the output was fed into a final classification layer.

3D swin transformer models on fusion data The complementary 3D fusion data from 'data fusion' section were used as inputs ($512 \times 512 \times (64 + N)$ frames) to the 3D Video Swin Transformer to assess effectivity of data fusion

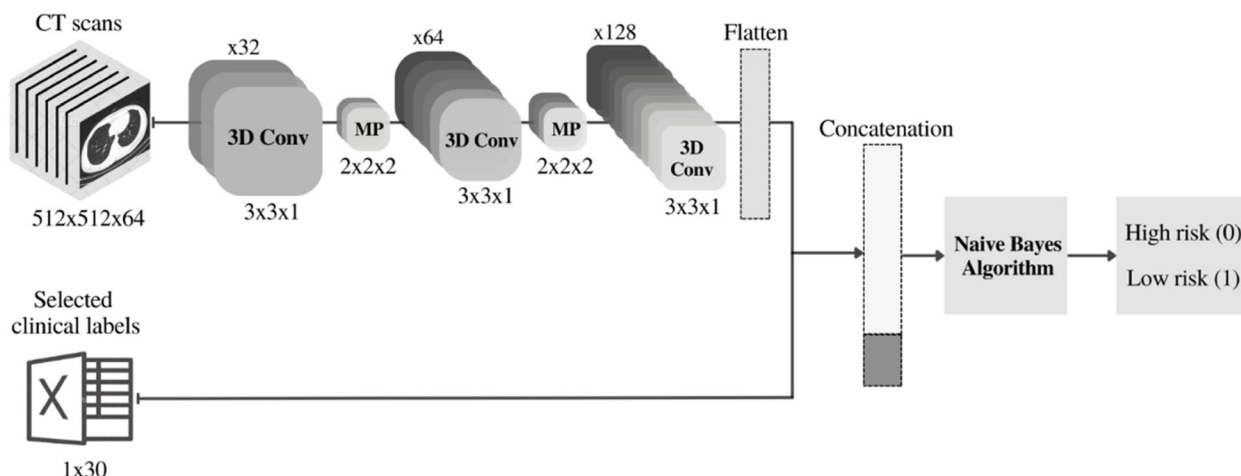


Fig. 3 Schematic of network architecture for the terminal 3D-CNN fusion model (on CTs + 30 clinical labels)

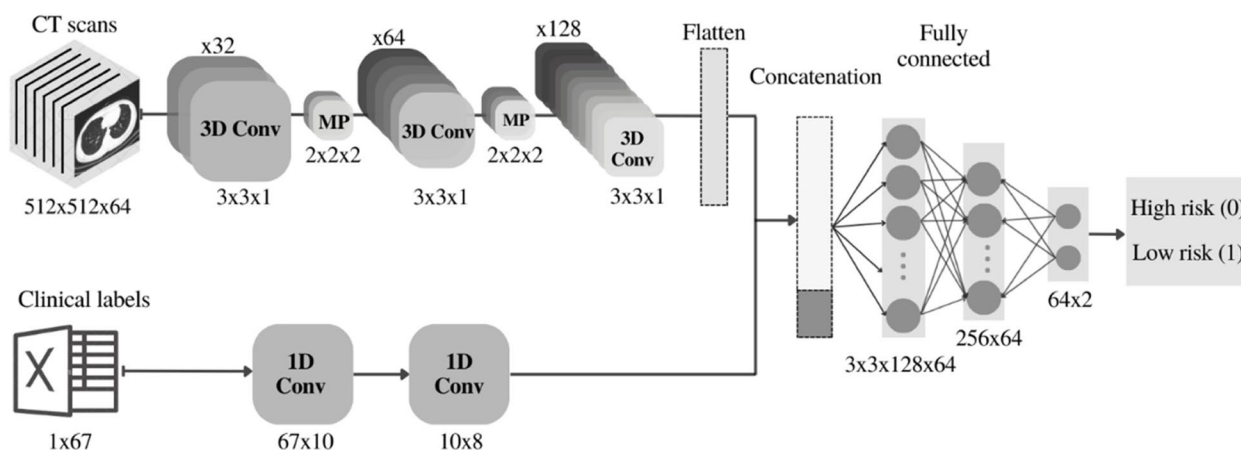


Fig. 4 Schematic of network architecture for the medial 3D-CNN fusion model (on CTs + 67 clinical labels)

approach. The schematic of our proposed data-fusion-based approach fed into the 3D Swin Transformer models is shown in Fig. 5. We tested the performance of the 3D Swin Transformer model under two different scenarios for $N=30$ (associated with the selected clinical labels in 'Dimension reduction' section) and $N=67$ (associated with all clinical labels).

Performance measure

A k-fold cross-validation approach ($k=5$) was used for overall performance assessments of all models. We also used the StratifiedKFold, a stratified cross-validator algorithm, to split the imbalanced dataset into train/test sets across 5 folds.

Performance measures such as Kappa and F0.5 score could provide better validation evaluations for the classification of imbalanced datasets compared to the

standard conventional measures such as “accuracy” and/or “precision/recall”. In fact, the later measures may not be reliable criteria when classification is performed on un-balanced data or when data is not normally distributed [65, 66]. AUC measure, however, includes the proportional impacts of the precision and recall metrics in validation assessments. Also, false positive rate (FPR – i.e., fall-out) is clinically a critical measure (e.g., compared to true positive rate (TPR) – i.e., sensitivity); this is mainly because this measure indicates how many of high-risk labels have been incorrectly classified in the low-risk class. Clinically, a high FPR rate is not acceptable as the misidentification of high-risk labels can be dangerous for patients who require treatments. Due to these reasons, our performance evaluation policy was focused on models that simultaneously achieved a minimal FPR, a higher TPR, a higher AUC, and higher F0.5 score and Kappa. This “trade-off” strategy was mainly targeted to

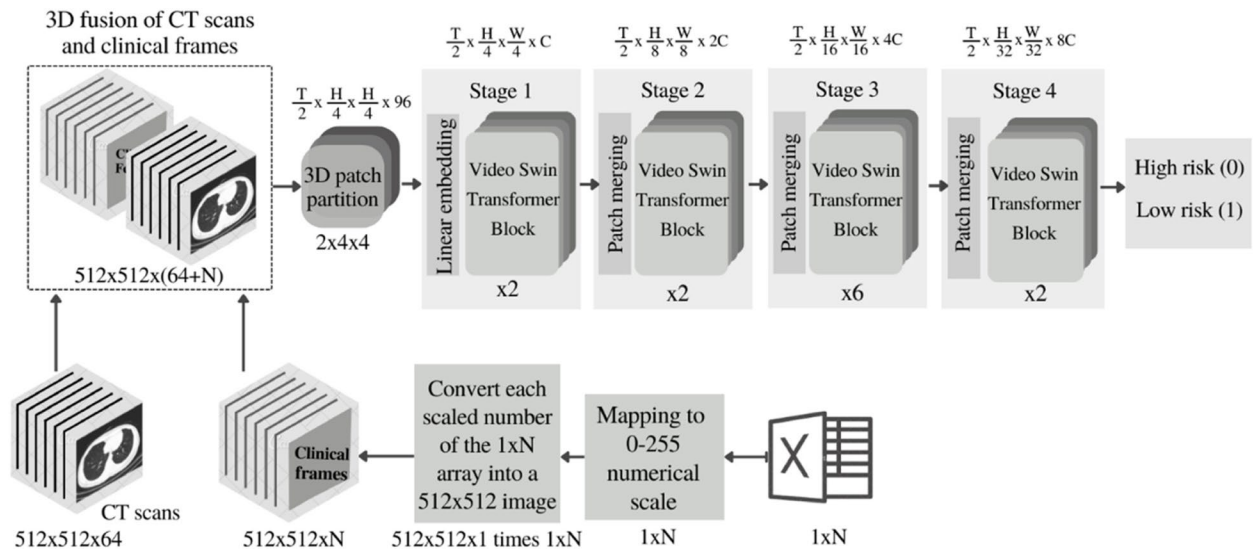


Fig. 5 Schematic of the 3D Swin Transformer model fed with the fusion of CT scan images and clinical data. “N” denotes the number of clinical labels

find a model with the lowest missed/wrong identifications for the high-risk class. These performance measures are described in the following sections.

Lastly, in machine learning, models are often evaluated using resampling methods such as k-fold cross-validation, which involves calculating and directly comparing mean performance scores of a model, across all folds. In this work, to assess whether the differences in mean scores among the top-performing models for all metrics across all folds are statistically significant, we further conducted a statistical analysis (i.e., p-values measures) between these models. This approach allows to further quantify the likelihood that the samples of scores in each fold were drawn from the same probability distribution across the 5-fold training sets, ensuring that the reporting performance evaluations are statistically significant. We employed a 95% confidence level (i.e., a p-value of 0.05) as the criterion to determine the statistical significance of the reported performance across these models.

Kappa

Kappa statistic is a performance measure that penalizes all positive or all negative predictions in its scoring regime. This approach is especially useful in multi-class imbalanced data classification and has been therefore commonly used in datasets with imbalanced classes [67, 68]. Moreover, Kappa has been shown to provide better insights than other metrics on detecting performance variations due to drifts in the distributions of the data classes. Kappa statistic ranges between -100 (total disagreement) through 0 (default probabilistic classification) to 100 (total agreement):

$$Kappa = \frac{n \sum_{i=1}^c x_{ii} - \sum_{i=1}^c x_{i.} x_{.i}}{n^2 - \sum_{i=1}^c x_{i.} x_{.i}} \times 100 \tag{1}$$

where x_{ii} is the count of cases in the main diagonal of the confusion matrix (successful predictions), n is the number of examples, c is the number of classes, and $x_{i.}$, $x_{.i}$ are the column and row total counts, respectively.

TPR, FPR and Precision

The TPR (also called sensitivity or recall), in this article, indicates how many of the data are correctly classified in the low-risk group (class 1) while the FPR indicates how many of the data in the high-risk group are incorrectly classified in the low-risk group (class 1). Also, precision (or positive predictive value (PPV)) evaluates the number of TPs out of the total number of positive predictions which indicates how good the model was able to make positive predictions.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

F0.5 score

F0.5 score is the weighted version of F1 score where more weight is considered to precision than to recall (Eq. 5). This is particularly important where more weight needs to be assigned to PPV for situations where FPs are considered worse than FNs.

$$1.25 \left(\frac{PPV \times TPR}{0.25PPV + TPR} \right) \tag{5}$$

To report realistic performance measures for the imbalanced classes in this work, we have employed the Macro averaging method [69] to evaluate the F0.5 score, recall, and precision metrics, in addition to reporting the actual TPR (sensitivity) and FPR (fall-out) values. Macro averaging is a commonly used technique for evaluating the overall performance metrics in imbalanced data against the most common class label(s). It is insensitive to the class imbalance within the dataset and treats all classes equally. With Macro averaging, these metrics are computed independently for each class and then averaged, ensuring equal treatment of all classes. Therefore, the reported metrics in this work include the TPR (sensitivity), FPR (fall-out), Macro-F0.5 score, Macro-recall, and Macro-precision measures.

Computing infrastructure

We used New Zealand eScience Infrastructure (NeSI) high-performance computing facilities’ Cray CS400 cluster for training and testing the models. The training process was executed using enhanced NVIDIA Tesla A100 PCIe GPUs with 40 GB HBM2 stacked memory bandwidth at 1555 GB/s. Intel Xeon Broadwell CPUs (E5-2695v4, 2.1 GHz) were used on the cluster for handling the GPU jobs. The algorithms were run under Python environments (Python 3.7) using Pytorch deep learning framework (Pytorch 1.11).

Results

This section provides the obtained results for the pre-processing, clinical-data-only trained models, CT scans-only trained models, as well as the fusion approaches for both CNNs and Transformer models.

Pre-processed clinical data

Full results from the feature selection and feature extraction in the ‘Dimension reduction’ section using the seven conventional classification algorithms for the (1) 67 original clinical labels, (2) 13 selected clinical labels from SelectK-Best algorithm, (3) 30 selected clinical labels from ExtraTree classifier, and (4) 25 extracted features from PCA algorithm are shown in Tables D.1 to D.4 in Appendix D, respectively. A trade-off performance criterion for a lower FPR and a higher TPR, F0.5 score, and Kappa in these tables showed that the Gaussian Naïve bays (NB) performed much better across the four approaches above. The abstracted results in Table 1 further confirm that the classification of the clinical data using Gaussian NB fed with 30 selected clinical labels from the ExtraTree classifier has led to better performances compared to other approaches.

In addition, features-space assessments using the t-SNE algorithm on the four above schemes are shown in Fig. 6A and D, respectively. The features-space plots in this figure hold high-complexity and a visual binary classification seems to be a challenging due to the negligible differences between the images. Nevertheless, the application of t-SNE on the 30 selected labels from ExtraTree classifier in Fig. 6C seems to provide a much better visually classifiable data.

Due to the above reasons, the dataset containing the 30 selected clinical labels from ExtraTree classifier were used as the clinical dataset for models in Approach #1 and Approach #3, where this data were further fused with the CT images to shape the fusion datasets.

Models on the clinical data only

Results from the seven classification algorithms in Approach #1 are shown in Table 2. Each classifier

Table 1 Comparison between the dimension reduction methods on the clinical data

Classifier	Dimension reduction method		Number of clinical labels /components	FPR	TPR	F0.5 score	Kappa
Gaussian Naive Bays	N/A	Raw data	67 labels	0.33	0.85	0.71	0.45
	Clinical labels selection	SelectKBest	13 labels	0.47	0.89	0.71	0.41
	Clinical labels selection	ExtraTree	30 labels	0.37	0.89	0.75	0.51
	Feature extraction	PCA	25 components	0.51	0.94	0.74	0.46

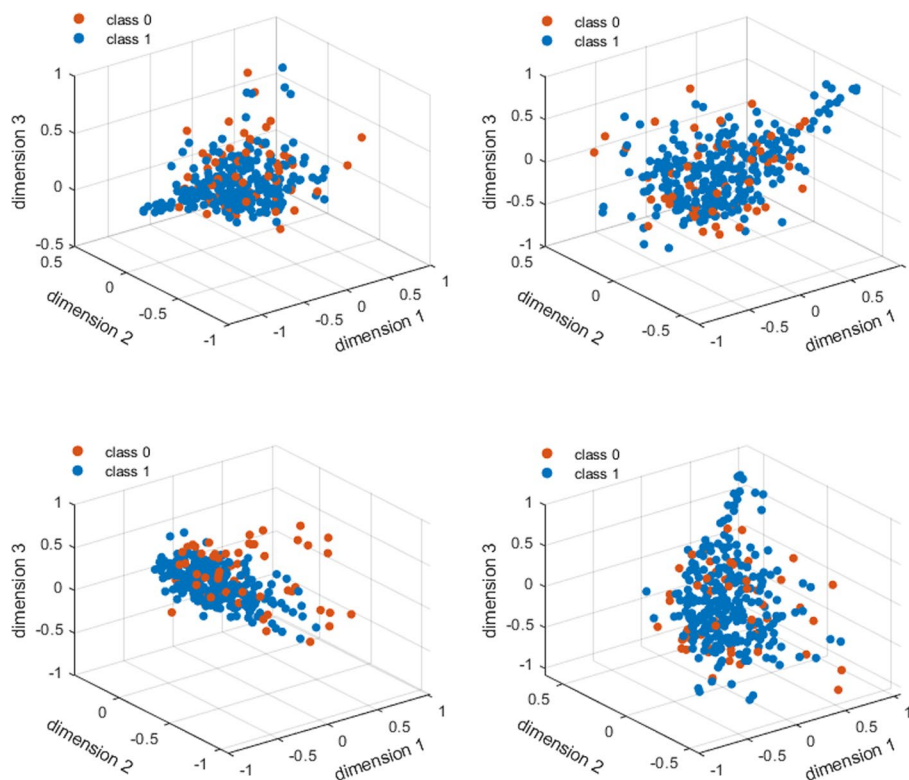


Fig. 6 Visual representations from the t-SNE approach using **A**: the main dataset including all 67 clinical labels (with no dimension reduction), **B**: 13 selected clinical labels from SelectKBest, **C**: 30 selected clinical labels from ExtraTree classifier, and **D**: 25 extracted features from PCA. blue: high-risk (class 0), orange: low-risk (class 1)

was assessed using the 30 selected clinical labels from ExtraTree classifier. As shown, here the gradient boosting algorithm has outperformed the other algorithms.

Models trained on CT images only

Results of the 5-fold cross-validation from the 3D-CNN and 3D Swin Transformer models in Approach #2 (trained on the CT-images only) are shown in Table 2. Results from the Transformer model on the CT images only shows improvement for all measures including FPR (0.45 lower), Kappa (0.27 higher), and F0.5 score (0.11 higher) compared to the 3D-CNN.

Models trained on fusion data

Results of the 5-fold cross-validation for each of the data-fusion approaches in Approach #3 are also shown in Table 2. As shown, the Transformer fusion models as well as the Terminal 3D-CNN have resulted in improved overall scores, across all measures, compared to the medial 3D-CNN fusion approach.

ROC curves of the top performing models from each section as well as the top performing 3D-CNN on the fusion data are shown in Fig. 7. Performance measures

of the top performing models in Fig. 7, including the 3D Swin Transformer on fusion data (CT+67 labels), 3D Swin Transformer on CTs only, Terminal 3D-CNN fusion (CT+30 labels), and Gradient Boosting, for all metrics across all 5 folds, showed statistical significance (p -value < 0.05).

Discussion

This paper, for the first time, demonstrated how a 3D data fusion approach of combining CT scan images and patients' clinical data can help to improve the performance of Visual Transformer and CNN models for predicting high-risk Covid-19 infection. Other studies have mainly focused on feeding such networks with either CT scan images or patients' clinical data. The paper explored a comprehensive set of strategies to evaluate optimal predictive model across a number of classifiers tested on a relatively large dataset of 380 patients. This research demonstrates the superiority of data-fusion approaches used in 3D Swin Transformers for better identification of high-risk Covid-19 infected patients.

Here we showed that the performance of a 3D Swin Transformer model tested on the fusion of CT scan images and the original set of 67 clinical

Table 2 Performance results of the classifiers

Model category	Model name	FPR	TPR	F0.5 score [†]	ROC/ AUC ^a	Recall ^b	Precision ^b	Kappa (-1, 1)
Approach #1: Clinical data only (on the set of 30 selected clinical labels from ExtraTree classifier)	Gaussian NB	0.49	0.70	0.57	0.60	0.61	0.59	0.19
	Random Forest	0.07	0.56	0.60	0.74	0.74	0.64	0.27
	Gradient Boosting	0.14	0.70	0.66	0.78	0.78	0.68	0.37
	XGBRF	0.16	0.65	0.62	0.72	0.73	0.64	0.28
	k-nearest neighbors	0.47	0.58	0.50	0.55	0.56	0.52	0.05
	SVM	0.18	0.18	0.32	0.50	0.50	0.42	0
	MLP	0.40	0.40	0.22	0.50	0.50	0.20	0
Approach #2: CTs only	3D-CNN	0.83	0.84	0.63	0.57	0.57	0.78	0.22
	3D Swin Transformer	0.38	0.89	0.75	0.75	0.75	0.75	0.49
Approach #3: Data fusion	Terminal 3D-CNN on CTs + 30 labels	0.36	0.90	0.75	0.76	0.76	0.76	0.51
	Medial 3D-CNN on CTs + 67 labels	0.65	0.98	0.70	0.66	0.66	0.67	0.37
	3D Swin Transformer on CTs + 30 labels	0.35	0.91	0.78	0.78	0.78	0.80	0.55
	3D Swin Transformer on CTs + 67 labels	0.40	0.95	0.82	0.77	0.77	0.83	0.60

^a ROC Receiver Operating Characteristics Curve, AUC Area under the ROC Curve

^b Denotes the macro-averaging evaluation method

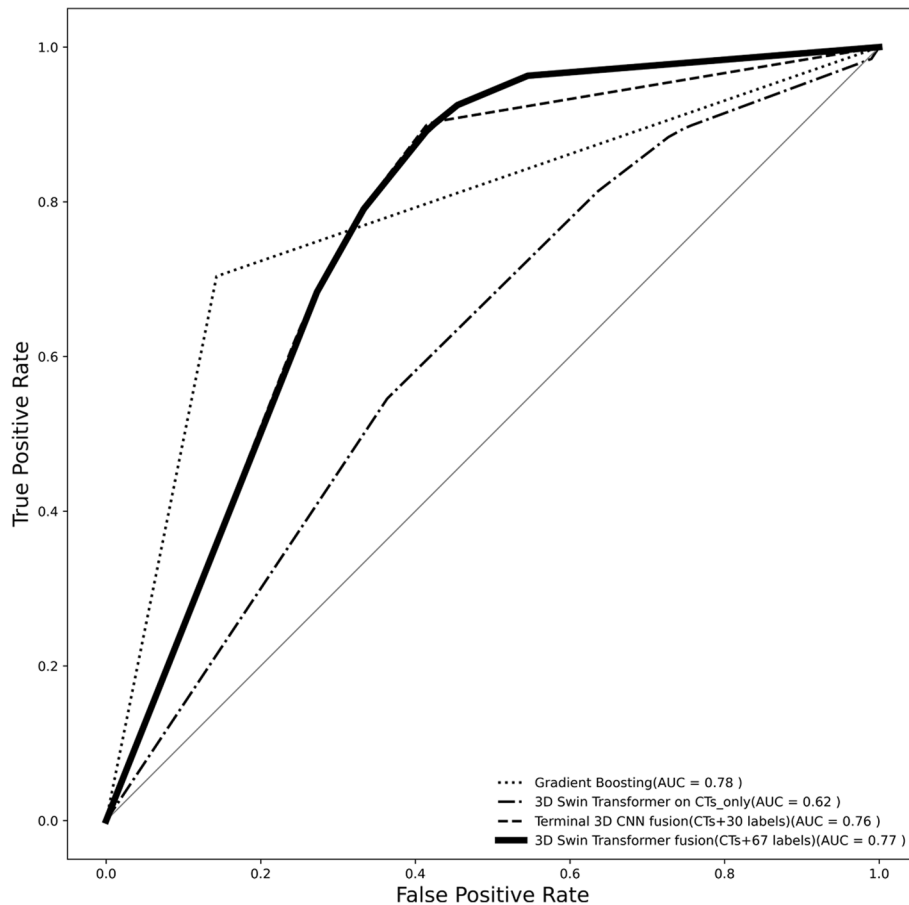


Fig. 7 Mean ROC curves from the top performing models. Bold line: 3D Swin Transformer on fusion data (CT + 67 labels), dashed-dotted line: 3D Swin Transformer on CTs only, dotted line: Gradient Boosting, dashed line: Terminal 3D-CNN fusion (CT + 30 labels)

labels outperformed all other strategies in this work (FPR=0.40, TPR=0.95, F0.5 score=0.82, AUC=0.77, Kappa=0.60) where the models were fed with fusion-type datasets, CT scan images only, and clinical data only. Here, we formed 3D fusion datasets by re-shaping the clinical data into $512 \times 512 \times \text{number of clinical labels}$ format and combined them with CT scan images of size $512 \times 512 \times 64$ to create our fusion dataset. It is inferred that our strategy for the dimension expansion of clinical data and fusing them with the CT scan images has successfully helped the self-attention layers within the Swin Transformer model to effectively rate interconnectivity between the clinical data and the CT images for better classifications.

We further tested and compared the performance of a terminal 3D-CNN model (on the CT+30 clinical labels), a medial 3D-CNN (on the CT+67 clinical labels), 3D Swin Transformers (on the CT+30 clinical labels and on the CT+67 clinical labels, respectively) to the original approach. Results from Table 2 indicates that our selected set of 30 clinical labels from the original pool of 67 clinical labels fused with the patients' CT scan images has been consistently and effectively helpful to achieve competitive performances compared to the 3D Swin Transformer on the CT+67 clinical labels. Here, the 3D Swin Transformer model on the fusion of CT+30 clinical labels achieved FPR=0.35, TPR=0.91, F0.5 score=0.78, AUC=0.78, Kappa=0.55, and the terminal 3D-CNN model on the fusion of CT+30 clinical labels achieved FPR=0.36, TPR=0.90, F0.5 score=0.75, AUC=0.76, Kappa=0.51. These closer performance measures from the selected set of 30 clinical labels suggest that these dominant labels may hold clinical values in the clinical settings for a better identification of the illness and could be looked at in details in future studies. These clinical labels have been listed in Table A.1 of Appendix A.

We also assessed classification capabilities of a 3D-CNN model and a Video Swin Transformer on sets of 3D CT scan images only. Here, the 3D Swin Transformer achieved much better results (FPR=0.38, TPR=0.89, F0.5 score=0.75, AUC=0.75, Kappa=0.49) compared to the 3D-CNN with higher FPR, lower AUC and Kappa (FPR=0.83, TPR=0.84, F0.5 score=0.63, AUC=0.57, Kappa=0.22).

Our assessments also showed that conventional classifiers, fed with patients' clinical data only, poorly classified the data compared to the other two approaches above, namely the fusion and CT-only strategies (see Table 2). Amongst the conventional classifiers, Gradient boosting was found to outperform the other ones when only fed with the clinical data (FPR=0.14, TPR=0.70, F0.5 score=0.66, AUC=0.78, Kappa=0.37).

An overall trade-off assessment shows that the 3D Swin Transformer fed with the fusion of CT scan images and the full set of 67 clinical labels identified high-risk patients from the low-risk class more accurately compared to the other approaches. This was closely followed by 3D Swin Transformer model fed with a fusion of CT images and the set of 30 selected clinical labels from ExtraTree classifier. The 3D Swin Transformer again demonstrated superiority compared to the 3D-CNN approach, even when both models were fed with CT images only; however, the overall performance was found to be lower than the data-fusion approach. Classification performances remarkably decreased across all the seven conventional models when only the clinical data were used. The mean ROC curves in Fig. 7 from the top performing models in each section demonstrate how the 3D Swin Transformer on fusion data (CT+67 labels) outperformed the other approaches. We have also provided the ROC curve of the top performing 3D-CNN model on the fusion data, namely the Terminal 3D-CNN fusion (CT+30 labels) to show how the choice of 30 selected clinical labels could also help our proposed CNN-based model to achieve competitive performances compared to the 3D Swin Transformer on the fusion data.

Overall, we expect potential clinical utility for the proposed 3D Video Swin Transformer fed with fusion datasets from patients' CT images and clinical data for reliable prediction of outcomes in Covid-19-infected patients. The improved performances of the Transformer models show robust capability for a future validation study on larger datasets. We encourage readers to apply the proposed fusion scheme in this work to larger clinical datasets for further validity assessments. Results from this research highlight the possibilities of predicting the severity of Covid-19 infection, at the time of admission to the clinical centers, when effectivity of early treatments is evident.

Limitations and future work

There are several limitations and avenues for future research emerging from the current study:

Data sample size: Our study's dataset comprises 380 Covid-19 diagnosed patients, which is relatively large. However, the study acknowledges the inherent variation in Covid-19 outcomes across demographic groups, regions, and healthcare systems. Consequently, the generalizability of our findings may be limited. To enhance the robustness of predictions, future research should focus on incorporating more diverse datasets, encompassing various patient profiles and geographical locations.

Data imbalance: The study employs imputation techniques to address missing data, a common practice in data analysis. Nevertheless, it is important to

acknowledge that imputation introduces potential biases that can affect the accuracy of predictive models. Although we provide transparency regarding our imputation methods, the chosen criteria and various factors within the strategy may impact results. Hence, one limitation of our study relates to potential biases introduced by these imputation procedures, which could have influenced algorithm performance.

Label selection: Our study involves the selection of clinical labels from a comprehensive pool of clinical data. The choice of labels significantly influences predictive model performance. Therefore, future research would benefit from a more comprehensive assessment of label selection criteria and an exploration of how this choice affects both model accuracy and clinical relevance.

Overfitting, external validation & model interpretability: To gauge the generalizability and real-world applicability of our models, external validation using an independent dataset is imperative. The absence of external validation in this study could be considered as a limitation that restricts our ability to demonstrate the models' practical utility. In the clinical context, validation on a larger, unseen dataset is essential to ensure interpretability. Future research efforts can be tailored to enhance model predictions, thereby providing trustworthy decision-making indicators.

Long-term outcome: While our study primarily focuses on predicting Covid-19 outcomes during or shortly after hospital admission, we recognize the potential for enduring effects and complications that extend beyond immediate hospitalization. Exploring the prediction and monitoring of post-recovery complications or long-term outcomes, especially in relation to the parameters examined in this study, represents a valuable avenue for future research. Such exploration holds significance for both patients and healthcare providers, enhancing the comprehensive management of Covid-19 cases.

Clinical utility: One limitation of this study was the inability to assess the robustness of the proposed models under conditions such as variations in data quality, imaging equipment, or specific clinical practices. Recognizing how these proposed models could potentially enhance patient outcomes is vital for their acceptance in healthcare settings. Consequently, future research should encompass clinical validation to evaluate the practical impact of employing these predictive models in clinical decision-making.

Conclusion

This paper demonstrated how the performance of Visual Transformers, namely a 3D Swin Transformer, could remarkably improve for predicting

Covid-19 outcomes when fed with a novel 3D data fusion approach of integrating CT scan images with patients' clinical data. The paper further explored and compared capabilities of a series of models including Transformers (on CT images only), 3D-CNNs (both on the fusion dataset and on CT images only) as well as conventional classifiers (on the clinical data only). Results showed that the use of fusion dataset provided opportunity for the 3D Swin Transformer model to better aggregate globally and locally interconnected features of the data and perform better compared to all other models. Results confirmed that this was valid for the larger fusion dataset of 64 CT scans + 67 clinical labels and the 64 CT scans + 30 selected clinical labels. The paper further discussed how genetic algorithm (GA) is a suitable choice for hyper-parameter tuning of the 3D-CNN models. We also investigated a series of strategies to find and select a proper set of clinical labels from the pool of clinical data for the classification of imbalance data. The paper further discusses imputation techniques to deal with missing values in the dataset. Overall, this paper demonstrates possibilities of predicting the severity of outcome in Covid-19 infected individuals at or around the time of admission to hospital using fusion datasets from patients' CT images and clinical data.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12911-023-02344-8>.

Additional file 1: Appendix A. Table A.1. Patients' clinical measurements (mean \pm std). X(Y%): X represents the label counts in the population, and the associated% of the label in the population is shown with Y. *Italic labels* represent the excluded labels previously explained in 'Clinical data trimming' section. **Appendix B.** Figure B.1. The suggested set of 30 selected clinical labels from ExtraTree classifier. Figure B.2. The suggested set of 13 clinical labels from SelectKbest algorithm. **Appendix C.** Table C.1. Selected hyper-parameters from the Genetic Algorithm for the 3D-CNN CT model ('3D-CNN CT model' section) and the medial 3D-CNN fusion model ('3D-CNN models on fusion data' section). **Appendix D.** Table D.2. Results of the seven conventional algorithms in 'pre-processing of clinical data' section on the 13 selected clinical labels from SelectKBest algorithm. Table D.3. Results of the seven conventional algorithms in 'pre-processing of clinical data' section on the 30 selected clinical labels from ExtraTree classifier. Table D.4. Results of the seven conventional algorithms in '3D-CNN models on fusion data' section on the 25 extracted features from PCA algorithm.

Acknowledgements

We would like to acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities to the results of this research. URL: <https://www.nesi.org.nz>.

We would like to thank Reza Azmi, Associate Professor at Alzahra University, Dr. Sara Abolghasemi from Infectious Diseases and Tropical Medicine Research Center at Shahid Beheshti University of Medical Sciences, and Narges Norouzi, Ph.D. Candidate at the Eindhoven University of Technology for their support and technical advice.

We would also like to thank Mohammad Arbabpour Bidgoli from the KTH Royal Institute of Technology in Sweden for assisting with data collection.

Authors' contributions

The algorithm development, data analysis and manuscript writing/preparation were undertaken by S.S, M.Z, and H.A. Data was collected and pre-processed by P.A, M.R, S.A.S, and A.S. Z.G contributed to algorithm development. M.G provided technical advice. Manuscript was reviewed, edited, and revised by H.A. H.A was the main supervisor of the work and assisted with study design, administration, and implementation. The final submitted article has been revised and approved by all authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All ethics of the current research have been approved by the Shahid Beheshti University's ethics committee (Ref: IR.SBMU.RETECH.REC.1399.003). All patients have signed and submitted their consent to participate in the research and their data privacy has been fully considered [37]. Informed consent was obtained from all subjects. All methods were carried out in accordance with relevant guidelines and regulations (Declaration of Helsinki).

Consent for publication

N/a.

Competing interests

The authors declare no competing interests.

Author details

¹Faculty of Engineering, Alzahra University, Tehran, Iran. ²University of Erlangen-Nuremberg, Bavaria, Germany. ³Department of Radiology, School of Medicine, Imam Hossein Hospital, Shahid Beheshti, University of Medical Sciences, Tehran, Iran. ⁴Research Institute for Gastroenterology and Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ⁵School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. ⁶Auckland City Hospital, Auckland 1010, New Zealand. ⁷Auckland Bioengineering Institute, University of Auckland, Auckland 1010, New Zealand.

Received: 8 January 2023 Accepted: 16 October 2023

Published online: 17 November 2023

References

- Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, Hu Y, Tao Z, Tian J, Pei Y. A new coronavirus associated with human Respiratory Disease in China. *Nature*. 2020;579(7798):265–9.
- of the International, Coronaviridae Study Group. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*. 2020;5(4):536.
- Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, Liu S, Zhao P, Liu H, Zhu L. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *The Lancet Respiratory Medicine*. 2020;8(4):420–2.
- Li Y, Bai W, Hashikawa T. The neuroinvasive potential of SARS-CoV2 may play a role in the Respiratory Failure of COVID-19 patients. *J Med Virol*. 2020;92(6):552–5.
- Li T, Lu H, Zhang W. Clinical observation and management of COVID-19 patients. *Emerg Microbes Infections*. 2020;9(1):687–90.
- Fang C, Bai S, Chen Q, Zhou Y, Xia L, Qin L, Gong S, Xie X, Zhou C, Tu D. Deep learning for predicting COVID-19 malignant progression. *Med Image Anal*. 2021;72:102096.
- Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, Li C. The Clinical and Chest CT Features Associated With Severe and Critical COVID-19 Pneumonia. *Invest Radiol*. 2020;55(6):327–31.
- Huang S, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in Pulmonary Embolism detection. *Sci Rep*. 2020;10(1):1–9.
- Bhattacharya S, Maddikunta PKR, Pham Q, Gadekallu TR, Chowdhary CL, Alazab M, Piran MJ. Deep learning and medical image processing for coronavirus (COVID-19) pandemic: a survey. *Sustainable Cities and Society*. 2021;65:102589.
- Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, Cai M, et al. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur Radiol*. 2021;31:6096–104.
- Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. *Comput Biol Med*. 2020;121:103795.
- Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. 2020;121:103792.
- Purohit K, Kesarwani A, Ranjan Kisku D, Dalui M. Covid-19 detection on chest x-ray and ct scan images using multi-image augmented deep learning model. In *Proceedings of the Seventh International Conference on Mathematics and Computing: ICMC 2021*. Singapore: Springer Singapore; 2022. p. 395–413.
- Gao Y, Cai G, Fang W, Li H, Wang S, Chen L, Yu Y, Liu D, Xu S, Cui P. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat Commun*. 2020;11(1):1–10.
- Patel D, Kher V, Desai B, Lei X, Cen S, Nanda N, Gholamrezanezhad A, Duddalwar V, Varghese B, Oberai AA. Machine learning based predictors for COVID-19 Disease severity. *Sci Rep*. 2021;11(1):1–7.
- Lassau N, Ammari S, Chouzenoux E, Gortais H, Herent P, Devilder M, Soliman S, Meyrignac O, Talabard MP, Lamarque JP, Dubois R, Loiseau N, Trichelair P, Bendjebbar E, Garcia G, Balleyguier C, Merad M, Stoclin A, Jégou S, Griscelli F, Tetelboun N, Li Y, Verma S, Terris M, Dardouri T, Gupta K, Neacsu A, Chemouni F, Sefta M, Jehanno P, Bousaid I, Boursin Y, Planchet E, Azoulay M, Dachary J, Brulport F, Gonzalez A, Dehaene O, Schiratti JB, Schutte K, Pesquet JC, Talbot H, Pronier E, Wainrib G, Clozel T, Barlesi F, Bellin MF, Blum MGB. "Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients," *Nat. Commun*, vol. 12, (1), pp. 634–4, 2021.
- Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad A, Uddin S, Kamal A, Alyami SA, Lin P, Islam SMS, Quinn JM. Machine Learning Approach to Predicting COVID-19 Disease Severity based on clinical blood Test Data: statistical analysis and Model Development. *JMIR Med Inf*. 2021;9(4):e25884.
- Yaşar Ş, Çolak C, Yoloğlu S. Artificial intelligence-based prediction of Covid-19 severity on the results of protein profiling. *Comput Methods Programs Biomed*. 2021;202:105996.
- Kivrak M, Guldogan E, Colak C. Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. *Comput Methods Programs Biomed*. 2021;201:105951.
- Khan YA, Abbas SZ, Truong B. Machine learning-based mortality rate prediction using optimized hyper-parameter. *Comput Methods Programs Biomed*. 2020;197:105704.
- Gong K, Wu D, Arru CD, Homayounieh F, Neumark N, Guan J, Buch V, Kim K, Bizzo BC, Ren H. A multi-center study of COVID-19 patient prognosis using deep learning-based CT image analysis and electronic health records. *Eur J Radiol*. 2021;139:109583.
- Meng L, Dong D, Li L, Niu M, Bai Y, Wang M, Qiu X, Zha Y, Tian J. A deep learning prognosis model help alert for COVID-19 patients at high-risk of death: a multi-center study. *IEEE J Biomedical Health Inf*. 2020;24(12):3576–84.
- Ho TT, Park J, Kim T, Park B, Lee J, Kim JY, Kim KB, Choi S, Kim YH, Lim J. "Deep learning models for predicting severe progression in COVID-19-infected patients: retrospective study," *JMIR Medical Informatics*, vol. 9, (1), pp. e24973, 2021.

24. Huang S, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med.* 2020;3(1):1–9.
25. Khan AI, Shah JL, Bhat MM. CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Comput Methods Programs Biomed.* 2020;196:105581.
26. Perumal V, Narayanan V, Rajasekar SJS. Prediction of COVID criticality score with laboratory, clinical and CT images using hybrid regression models. *Comput Methods Programs Biomed.* 2021;209:106336.
27. Qi S, Xu C, Li C, Tian B, Xia S, Ren J, Yang L, Wang H, Yu H. DR-MIL: deep represented multiple instance learning distinguishes COVID-19 from community-acquired Pneumonia in CT images. *Comput Methods Programs Biomed.* 2021;211:106406.
28. Shahid O, Nasajpour M, Pouriyyeh S, Parizi RM, Han M, Valero M, Li F, Aledhari M, Sheng QZ. Machine learning research towards combating COVID-19: Virus detection, spread prevention, and medical assistance. *J Biomed Inform.* 2021;117:103751.
29. Saberi S, Zarvani M, Amiri P, Azmi R, Abbasi H. "Deep learning classification schemes for the identification of COVID-19 infected patients using large chest X-ray image dataset," in *42nd Annual International Conference of the IEEE in Engineering in Medicine and Biology Society (EMBC), 2020.*
30. Ambita AAE, Boquio ENV, Naval PC. Covit-gan: vision transformer for covid-19 detection in ct scan images with self-attention gan for Data Augmentation. In *International Conference on Artificial Neural Networks.* Cham: Springer International Publishing; 2021. p. 587–598.
31. Krishnan KS, Krishnan KS. "Vision transformer based COVID-19 detection using chest X-rays." In *2021 6th International Conference on Signal Processing, Computing and Control (ISPC).* IEEE; 2021. p. 644–648.
32. Hsu C, Chen G, Wu M. "Visual transformer with statistical test for covid-19 classification," *arXiv Preprint arXiv:2107.05334*, 2021.
33. Mondal AK, Bhattacharjee A, Singla P, Prathosh AP. xViTCOS: explainable vision transformer based COVID-19 screening using radiography. *IEEE J Translational Eng Health Med.* 2021;10:1–10.
34. Fan X, Feng X, Dong Y, Hou H. "COVID-19 CT image recognition algorithm based on transformer and CNN." *Displays.* 2022;72:102150.
35. Nassif AB, Shahin I, Bader M, Hassan A, Werghi N. "COVID-19 Detection Systems Using Deep-Learning Algorithms Based on Speech and Image Data," *Mathematics*, vol. 10, (4), pp. 564, 2022.
36. Rahmzade R, Rahmzadeh R, Hashemian SM, Tabarsi P. Iran's Approach to COVID-19: Evolving Treatment Protocols and Ongoing Clinical Trials. *Front Public Health.* 2020;8:551889.
37. Raoufi M, Naini SAA, Safavi Z, Azizan FJ, Zade F, Shojaeian MG, Boroujeni F, Robatjazi M, Haghghi AA, Dolatabadi, Soleimantabar H. "Correlation between chest computed tomography scan findings and mortality of COVID-19 cases; a cross sectional study," *Archives of Academic Emergency Medicine*, vol. 8, (1), 2020.
38. Zheng A, Casari A. *Feature Engineering for Machine Learning: principles and techniques for data scientists.* O'Reilly Media, Inc.; 2018.
39. Little RJ, Rubin DB. *Statistical Analysis with Missing Data* John Wiley & Sons, 2019793.
40. Lee JY, Styczynski MP. "NS-kNN: a modified k-nearest neighbors approach for imputing metabolomics data," *Metabolomics*, vol. 14, (12), pp. 1–12, 2018.
41. Rafsunjani S, Safa RS, Al Imran A, Rahim MS, Nandi D. An empirical comparison of missing value imputation techniques on APS failure prediction. *IJ Inf Technol Comput Sci.* 2019;2:21–9.
42. Zeng D, Xie D, Liu R, Li X. "Missing value imputation methods for TCM medical data and its effect in the classifier accuracy," in *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 2017, pp. 1–4.
43. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiology.* 2017;70(4):407.
44. Batista GE, Monard MC. "A study of K-nearest neighbour as an imputation method." *His*, vol. 87, (251–260), pp. 48, 2002.
45. Fodor IK. A survey of dimension reduction techniques. No. UCRL-ID-148494. CA (US): Lawrence Livermore National Lab.; 2002.
46. Khalid S, Khalil T, Nasreen S. "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*, 2014, pp. 372–378.
47. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol.* 1933;24(6):417.
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–10.
49. Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS. An in-depth analysis of machine learning approaches to predict depression. *Curr Res Behav Sci.* 2021;2:100044.
50. Powell A, Bates D, Van Wyk C, de Abreu D. "A cross-comparison of feature selection algorithms on multiple cyber security data-sets." in *Fair*, 2019, pp. 196–207.
51. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn.* 2006;63(1):3–42.
52. Sandri M, Zuccolotto P. A bias correction algorithm for the Gini variable importance measure in classification trees. *J Comput Graphical Stat.* 2008;17(3):611–28.
53. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;9(11):2579–605. <https://www.jmlr.org/papers/volume9/vandemaaten08a/vandemaaten08a.pdf?fbclid>.
54. Zunair H, Rahman A, Mohammed N, Cohen JP. "Uniformizing techniques to process CT scans with 3D CNNs for tuberculosis prediction." In *Predictive Intelligence in Medicine: Third International Workshop, PRIME 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3.* Springer International Publishing; 2020. p. 156–168.
55. Lin W, Tsai C, Hu Y, Jhang J. Clustering-based undersampling in class-imbalanced data. *Inf Sci.* 2017;409:17–26.
56. McLachlan GJ, Lee SX, Rathnayake SI. Finite mixture models. *Annual Rev Stat its Application.* 2019;6:355–78.
57. Sun Y, Xue B, Zhang M, Yen GG, Lv J. Automatically designing CNN architectures using the genetic algorithm for image classification. *IEEE Trans Cybernetics.* 2020;50(9):3840–54.
58. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell.* 2022;45(1):87–110.
59. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Preprint arXiv:2010.11929*, 2020.
60. Li J, Yan Y, Liao S, Yang X, Shao L. "Local-to-global self-attention in vision transformers," *arXiv Preprint arXiv:2107.04735*, 2021.
61. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? *Adv Neural Inf Process Syst.* 2021;34:12116–28.
62. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022.
63. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H. "Video swin transformer," *arXiv Preprint arXiv:2106.13230*, 2021.
64. John GH, Langley P. "Estimating continuous distributions in Bayesian classifiers," *arXiv Preprint arXiv:1302.4964*, 2013.
65. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Comput Surv (CSUR).* 2016;49(2):1–50.
66. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: a review. *Int J Pat Recognit Artif Intell.* 2009;23(04):687–719.
67. Jeni LA, Cohn JF, De La Torre F. "Facing imbalanced data—recommendations for the use of performance metrics," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 245–251.
68. Brzezinski D, Stefanowski J, Susmaga R, Szczech I. Visual-based analysis of classification measures and their properties for class imbalanced problems. *Inf Sci.* 2018;462:242–61.
69. Huang C, Huang X, Fang Y, Xu J, Qu Y, Zhai P, Fan L, Yin H, Xu Y, Li J. Sample imbalance Disease classification model based on association rule feature selection. *Pattern Recog Lett.* 2020;133:280–6.

Publisher's Note

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