COVID-19 Disease Progression Prediction via Audio Signals: A Longitudinal Study

Ting Dang1 Jing Han1‡ Tong Xia‡ Dimitris Spathis1 Erika Bondareva1†
Chloë Brown1† Jagmohan Chauhan1,4† Andreas Grammenos1†
Apinan Hasthanasombat1† Andres Floto2 Pietro Cicuta3 Cecilia Mascolo1

1Department of Computer Science and Technology, University of Cambridge, UK
2Department of Medicine, University of Cambridge, UK
3Department of Physics, University of Cambridge, UK
4ECS, University of Southampton, UK
‡JH, TX contributed equally to this study
†listed in alphabetical order, contributed equally

Abstract

Recent work has shown the potential of the use of audio data in screening for COVID-19. However, very little exploration has been done of monitoring disease progression, especially recovery in COVID-19 through audio. Tracking disease progression characteristics and patterns of recovery could lead to tremendous insights and more timely treatment or treatment adjustment, as well as better resources management in health care systems.

The primary objective of this study is to explore the potential of longitudinal audio dynamics for COVID-19 monitoring using sequential deep learning techniques, focusing on prediction of disease progression and, especially, recovery trend prediction. We analysed crowdsourced respiratory audio data from 212 individuals over 5 days to 385 days, alongside their self-reported COVID-19 test results. We first explore the benefits of capturing longitudinal dynamics of audio biomarkers for COVID-19 detection. The strong performance, yielding an AUC-ROC of 0.79, sensitivity of 0.75 and specificity of 0.70, supports the effectiveness of the approach compared to methods that do not leverage longitudinal dynamics. We further examine the predicted disease progression trajectory, which displays high consistency with the longitudinal test results with a correlation of 0.76 in the test cohort, and 0.86 in a subset of the test cohort with 12 participants who report disease recovery. Our findings suggest that monitoring COVID-19 progression via longitudinal audio data has enormous potential in the tracking of individuals’ disease progression and recovery.

Introduction

Since the beginning of the SARS-CoV-2 pandemic in January 2020, different methods have been developed and employed for diagnostic testing and screening. In addition to the most commonly adopted laboratory tests via reverse transcription polymerase chain reaction (RT-PCR) [1, 2] or chest computer tomography (CT) scans [3] for diagnosis, a variety of digital technologies, often employing artificial intelligence, have also been investigated for COVID-19 screening [4, 5, 6]. Among these, automatic audio-based COVID-19 detection has drawn increasing attention due to its numerous advantages, including its flexi-
Fig. 1 | Overview of study design: COVID-19 detection and disease progression are assessed from audio data. Voice, cough and breathing sound recordings are collected from each participant over a period of time, together with self-reported COVID-19 test results. During the model development, audio recordings are chunked into segments consisting of 5 samples covering a few days, and processed using sequential modelling techniques (Gated Recurrent Units) for COVID-19 monitoring. Two subtasks are evaluated: i) COVID-19 detection (positive v.s. negative); and ii) Disease progression monitoring.
Sound recordings have distinct features during the disease progression. This is evident here in spectrograms of one participant, repeating the same sentence on 6 different days. The participant reported positive test results from 2020-11-10 to 2020-11-18, and reported negative test results from 2020-11-22 to 2020-12-26, indicating a recovery trend. The fundamental frequency and its harmonics (black box) for positive recordings demonstrates the lack of control in vocal cords, indicated by its non-separability. An increasing separability can be seen from positive to negative recordings over time, suggesting the recovery of voice characteristics. Similarly, the harmonics in the frequency range [2kHz 4kHz] (blue box) manifest an increasing separability, also reflecting the recovery trend.

Recent work has recognized the use of a 3-escalating-phase description of COVID-19 disease progression [25], namely, early infection, pulmonary involvement, and systemic hyperinflammation, which demonstrates commonality in disease progression. We hypothesize that this could be captured longitudinally via audio signals in automatic monitoring systems. Though the participants in our study may not experience all three stages, it is assumed that audio characteristics are affected during clinical progression of disease. Fig. 2 shows spectrograms of one participant reading the same sentence over a 43-day period. They reported COVID-19 infection followed by recov-
ery. As indicated in the black box, the fundamental frequency and its harmonics are not clearly separable when the participant tested positive (top row) especially on 2020-11-14, indicating a lack of control of vibration of the vocal cords. The separability increases after recovery. This matches the observed clinical course of COVID-19 disease progression [26], with the least separability 5 days after the first positive test result (2020-11-14), and an increase in separability 9 days after infection (2020-11-18). Similar patterns are also observed for the harmonics in the high frequency range between 2kHz to 4kHz (blue box). There is no obvious difference in the spectrogram patterns between 2020-11-18 (positive) and 2020-11-22 (negative), reflecting the difficulty of the COVID-19 detection task in general. Overall, this evolution of the spectrograms with disease progression shows that COVID-19 infection can manifest as changes in the acoustic representations, and modelling the audio sequences longitudinally might benefit COVID-19 detection.

Furthermore, the audio characteristics vary among individuals, e.g. one positive participant may produce a similar spectrogram as another negative participant. This is not considered in most conventional audio-based COVID-19 detection systems, which so far only used a single audio sample, rather than sequences. This makes the automatic detection a hard task and may lead to wrong predictions. While modelling the evolution of the spectrograms longitudinally, the individual's past audio signal can serve as a ‘baseline’, and the predictions can be corrected given this reference. Additionally, the spectrogram for each individual when healthy can be used as a reference for its own infected status, and modelling the relative changes in the audio sequences longitudinally is likely to be more accurate for COVID-19 detection. In a much generalised sense, the mean and the standard deviation of the audio recordings in individual’s healthy state, are both personalised. This provide a good threshold for the non-healthy states and benefits the COVID-19 detection. Motivated by these advantages, we explored the potential of modelling the sequential audio recordings longitudinally, as biomarkers of disease progression, focusing on how best to capture dynamic changes in the audio sequences over time, and aiming to demonstrate predictive power.

In this study, we developed an audio-based COVID-19 disease progression prediction model using longitudinal audio data. We adopted sequential deep learning models to capture longitudinal audio dynamics and to make predictions of disease progression over time. First, we examined whether modelling audio dynamics could aid COVID-19 detection, this shows strong performance compared to conventional models using a single audio sample. We then evaluated our model’s performance in predicting disease progression trajectories: our predictions successfully track test results and also match the statistical analysis in the timeline and duration of COVID-19 progression. In particular, we explored the use of audio signals for recovery prediction, as this may be useful in relation to setting home quarantine requirements.

From a public health perspective, an approach such as the one we propose has potential implications for how infected individuals are monitored, namely, it could allow more fine-grained remote tracking, and hence more efficient management of health system resources, by keeping individuals out of hospital as much as possible.

Results

Dataset preparation and statistics. An app\(^1\) was developed and released in April 2020 for data gathering purposes, aiming to crowd-source participants’ audio recordings, COVID-19 test results, along with the participants’ demographics, medical history, and COVID-19 related symptoms. Three different audio sounds were collected, including three cough recordings, three to five breathing sounds, and speech recordings, where each participant was asked to read a short phrase displayed on the screen, three times. The COVID-19 test results were self-reported and included a positive report, or a negative one or not having been tested. Participants were encouraged to provide data regularly. More details can be found in Methods.

\(^1\)https://www.covid-19-sounds.org/en/
Fig. 3 | Large datasets are required to identify, and avoid, biases. Data flow diagram and demographic statistics. 

**a,** Data selection process. **b,** Demographic statistics of the eligible participants including language, gender, age and symptoms. English is the dominant language, comprising 54.2% of the cohort. Age and gender are relatively balanced between positive and negative groups. 94.3% positive participants and 77.4% negative participants reported COVID-19 symptoms. **c,** Duration and reporting intervals in terms of days and number of samples. The median number of samples is 9 (left), corresponding to a time span of 35 days (middle left). Negative participants reported for a longer period compared to positive participants. The median reporting interval for the cohort is 3 days (middle right), validating the effective temporal dependencies of the audio data. The median duration after augmentation is 17 and 18 days for positive and negative participants respectively (right), showing that the augmentation has eliminated the confounding effects of the different duration for two subgroups.
From April 2020 to April 2021 (Fig. 3a), 3845 healthy participants (negative) and 1456 positive participants contributed audio samples with positive or negative clinical self-reported test results for at least one day. We used these participants’ data if five or more samples were provided, resulting in 447 and 168 negatively and positively tested participants, respectively. Label quality was manually checked to remove unreliable users and audio recording quality was examined using Yannet [27] to filter out corrupted and noisy samples, leaving 106 positive participants. To generate a balanced dataset, a cohort of 212 longitudinal users in total (106 positive and 106 negative) were selected across different countries.

The mobile app is a multi-language platform, and the cohort consists of eight different languages, with 54% English users as the dominant subgroup (Fig. 3b). Age and gender are relatively balanced between positive and negative groups, with 52.4% female participants, and 61.3% aged between 30-59. 100 out of 106 positive participants (94.3%) reported COVID-19 symptoms such as loss of taste or smell, fever, etc., while 82 out of 106 negative participants (77.4%) also reported these symptoms. The median number of samples for each user was 9, corresponding to a time period of 35 days (Fig. 3c middle left). The median duration for positive participants was smaller than for negative participants, namely 28 days and 45 days respectively. This duration is expected to contain adequate audio dynamics associated with disease progression, and to cover the complete course of disease progression for most participants [22]. In addition, the reporting interval for each participant was also computed as the average of the time intervals between two consecutive samples, and the median value was 3 days for both positive and negative groups (Fig. 3c middle right), validating the temporal dependencies of the data. To develop the machine learning models, data augmentation was carried out (see Methods), and the durations after augmentation for positive and negative participants are balanced, with a median value of 17 and 18 days respectively (Fig. 3c right). This duration aligns with the disease progression duration that is generally between 11 days to 21 days [20, 21, 22, 23, 24]. The similar duration for positive and negative participants after augmentation also helped to eliminate the confounding effect of the original different duration for two subgroups in model development (Fig. 3c upper right). The data was split into training, validation, and test partitions with 70%, 10% and 20% balanced positive and negative participants respectively, as well as the relatively balanced languages and genders (see Appendix Fig. 1).

**Study design and overview.** We investigated whether modelling audio biomarkers (cough, breath and voice) longitudinally can benefit COVID-19 detection and if it could be employed to monitor the disease progression accurately and in a timely manner (Fig. 1). The audio sequences are modelled by recurrent neural networks with Gated Recurrent Units (GRUs) to take into account the audio dynamics, which reflect disease progression (see Methods). The investigation is divided into two subtasks: one concerning the COVID-19 detection by predicting audio biomarkers as positive and negative, and the other concerning disease progression trajectory monitoring, examining the predicted probability of being positive over time. For instance, a decrease in the probability of being positive over time indicates a recovery trend, and an increase indicates a worsening trend. The first subtask aims to assess whether modelling past audio biomarkers in the input space benefits COVID-19 detection in general, while the second subtask focuses on longitudinal analysis of disease progression in the output space.

**COVID-19 detection.** To determine whether taking into account audio dynamics via the sequential modelling techniques of GRUs is effective in detecting COVID-19, the performance is compared against two benchmarks that do not capture audio dynamics (Fig. 4): one uses only audio biomarkers of the same day for prediction, while the other uses the average feature representation of audio sequences in the prior days for the last day prediction (Fig. 4a). The proposed system (Fig. 4b) outperforms the two benchmarks with the highest ROC-AUC of 0.79(0.74-0.84), sensitivity of 0.75(0.67-0.82), and specificity of 0.70(0.66-0.74), yielding 19.7% and 31.7% relative improvement over the two benchmarks, and demonstrating the effectiveness of modelling longitudinal
Fig. 4 | The proposed sequential model shows superior performance in COVID-19 detection compared to benchmarks leveraging only one isolated audio data point per user. a, ‘Average’ means using the average of feature representations within sequence for prediction, and ‘Single’ means using only the feature representation at the same day for prediction. None of these systems capture the longitudinal voice dynamics. b, The proposed sequential modelling outperforms two benchmarks, suggesting the advantages of capturing the disease progression via voice dynamics. c, Individual-level accuracy for 42 participants in the test cohort.

audio biomarkers.

For further assessment of whether sequential modelling using past audio biomarkers could provide an adjustable baseline for each individual, the prediction accuracy for each participant is evaluated and compared to the ‘single’ benchmark, defined as the ratio of correctly predicted samples over the total number of samples for each participant (Fig. 4c). We observe that the proposed sequential modelling outperforms the ‘single’ benchmark. The performance range of the sequential model for negative participants is larger than the benchmark, due to worse performance on two individuals.

**Disease progression prediction.** We analysed predicted disease progression trajectory by comparing it with the test results. The predicted progression trajectory is represented by the probability of positive prediction within the range 0 to 1 over time, with a higher value indicating a high possibility of positive test results (Fig. 5). Three different disease progression trajectories are shown in Fig. 5a, 5b and 5c. For the recovering participant P1 (Fig. 5a), a high probability is observed when P1 tested positive and low probabilities when P1 tested negative. The model performance is evaluated using point-biserial correlation coefficients γpb that measure the correlation between the predicted trajectory and the test results. Our model achieves 0.86 of γpb for P1, demonstrating a strong capability to predict disease progression. We can further categorize the probability over 0.5 as
Our approach enables to predict disease progression. Shaded orange and cyan indicate positive and negative test results, and '+' and 'o' represent positive and negative predictions. The green star indicates the presence of symptoms.

**a** disease progression of recovering participant P1; **b**, disease progression of positive participant P2; **c**, disease progression of negative participant P3; **d**, overall performance for the test cohort in terms of point-biserial correlation coefficient $\gamma_{pb}$ / accuracy $\gamma$.

![Graphs](image_url)

A positive prediction, and below 0.5 as a negative prediction. The predictions also match the test results.

For a positive participant P2 who consistently reports positive test results (Fig. 5b), our model outputs positive predictions matching these test results. It should be noted that $\gamma_{pb}$ is not applicable to participants who report consistently positive or negative test results, therefore, the accuracy $\gamma$ is instead used, which computes the ratio of correct predictions in terms of positive or negative over the total number of samples. P2 yields $\gamma = 1$ since all the predictions are positive. Further, the probability of positive increases from symptom onset day 0 to day 8, and decreases slightly to day 16, which matches the clinical course in general. For a negative participant P3 (Fig. 5c), the predicted probability is consistently below 0.5, corresponding to negative predictions that align with the test results. One type of disease progression that transitions from negative to positive is not included, due to the limited number of participants in the cohort. Even though we have adopted time inverse augmentation (cf. Methods) that reverses the audio biomarkers and their corresponding labels in time to enrich the different disease progressions, especially the negative to positive transitions, the time reversed audio biomarkers and disease progression may still not match the actual progression, and cannot be well captured in the model.

Fig. 5d demonstrates the model performance in disease progression monitoring for the entire test cohort, where $\gamma_{pb}$ and $\gamma$ are adopted for different participants. Though $\gamma_{pb}$ ranges between -1 and 1 and $\gamma$ is in the range 0 to 1, the higher the value is, the better the predictions for both metrics. It is also
Recovery trends can be predicted. The shaded orange and cyan area indicate the positive and negative test results. The predictions above 0.5 are categorized as positive predictions (plus mark) and below as negative predictions (dot). a, b Recovery predictions for two different participants P4 and P5. c, d Overall performance for recovery participants in the test cohort w/ and w/o dynamic time warping (DWT), which calculates the optimal match between predicted recovery trajectory and test results. d, Projection of latent vectors learnt by the model for three different participants. The y axis from top to bottom indicates the test results over time. A clear change in each latent vector dimension transitioning from positive to negative can be observed (recovering user), and consistent and different patterns can be observed for positive and negative participants. e and f, Scatter plot of symptoms vs. probability of positive for positive (e) and negative (f) participants, with a high correlation observed for positive participants and no correlation for negative participants.

expected that $\gamma_{pb}$ achieves positive correlation between 0 to 1 for a good model. Therefore, we report performance by combing these two measures. Our model achieves 0.75 for all the 42 test participants, and 0.86 and 0.71 for positive and negative participants respectively. Some more examples are given in Appendix Fig. 3 and Fig. 4.

Recovery trajectory prediction. We further examined the model performance for the recovery subgroup in the test cohort, where 12 out of 21 positive participants reported a recovery trend in their test
results. The predicted recovery trajectory of two randomly selected participants P4 and P5 are presented in Fig. 6a and 6b. For participant P4, a slight increase in probability is observed from day 0 to day 2, suggesting an increase in severity of illness. The predicted probability decreases from day 2 showing the recovery trend. The categorized positive and negative predictions also match the test results except for day 27, with our model predicting negative and the test result still showing positive. It should be noted that there may be a delay in participants taking clinical tests or reporting results, and that therefore, earlier negative predictions are acceptable. This also suggest the advantages of our audio-based data, which is precisely timed and can be instantly analysed.

For participant P5 in Fig. 6b, the predicted recovery trend with decreasing probability is clearly observed. However, the probabilities are all categorized as positive predictions even after the user tests negative from day 16 to 23. This is possibly due to i) individual differences in audio characteristics; or ii) an asymptomatic participant exhibiting minor changes in audio characteristics, thus leading to a slowly recovering trend prediction.

Overall performance for all 12 recovering participants is reported in Fig. 6c, with \( \gamma_{pb} = 0.76 \). As negative predictions with a time shift from negative test results are acceptable (as discussed in Fig. 6a), we align predictions and test results temporally using Dynamic Time Warping (DTW). The model further achieves \( \gamma_{pb} = 0.86 \), demonstrating a strong capability to monitor recovery. Some more examples can be found in Appendix Fig. 3 and Fig. 4.

**Visualisation of latent space.** To gain an in-depth understanding of the model, we analyse the intermediate audio representations. The latent vectors from the GRU outputs are projected from 64 dimensions to a four-dimension latent space using Principal Component Analysis (PCA) for visualisation (Fig. 6d). For the recovery users, a clear transition was observed for the first three dimensions when the participant recovers, while a consistent but different pattern was observed for the positive and negative participants respectively.

**Disease progression with symptoms.** We further studied the correlations between the predicted probability and symptoms for positive and negative participants. Participants reported 13 different symptoms, including fever, shortness of breath, wet and dry cough, loss of taste and smell, myalgia, headache, sore throat, chest tightness, rigors, runny or blocked nose, dizziness, confusion or vertigo. We assume that the number of symptoms is correlated with the severity of illness for positive participants, thus a high probability of positive predictions is expected for the audio recordings reported alongside with more symptoms. With an increase in number of symptoms, we can observe a general increase in the predicted probability (Fig. 6e). We further fit a line (red) for these participants, excluding those with more than 5 symptoms due to the limited number of samples. We can observe a clear positive correlation. Conversely, for negative participants (Fig. 6f), we observe no correlation between the probability and number of symptoms, suggesting that our model is not capturing symptoms, but information related to COVID-19.

**Disease progression in first 7 days.** Evidence on X-ray or chest tomography shows that 22.6% patients presenting with disease experience resolution 7 days after symptom onset, 12.1% show a stable condition, and the rest 65.3% patients worsen within 7 days from symptom onset [26]. We analysed the predicted trajectories over similar time period. Though many participants report symptoms on the first day they start recording, which may not be the first day they experienced symptoms, the increasing trend in the first few days could still suggest the initial worsening of patients’ conditions. We define the 7 day window to be from either the first day of reported symptoms, or the day of the first positive test if no symptoms are reported before that. This leads to 12 eligible participants in the test cohort, where we found 8 out of 12 (66.7%) show an increase in predicted probability, similar to the statistical analysis. We hypothesize that the predicted progression in the first 7 days could provide a more prompt indication of individual’s recovery rate.
Discussion

From a crowdsourced audio dataset, we studied 212 longitudinal participants and developed a deep learning model for COVID-19 disease progression monitoring via audio signals. We showed that modelling audio dynamics longitudinally shows benefit for COVID-19 detection. Individual-level performance also displays a significant improvement over baseline. The model capability to predict disease progression has been validated. Successful tracking of reported test labels shows strong performance in disease progression with $\frac{\gamma_{pb}}{\gamma} = 0.75$. We specifically focused on recovery prediction, with a correlation $\gamma_{pb}$ of 0.86 between the predicted progression trajectory and test results.

Individuals experience different disease progression trajectories, and our model can capture this variability among individuals. For the positive participant P6 (Appendix Fig. 3c), our model demonstrates an decrease in the predicted probability from day 0 to day 3, and this is followed by a slight increase from day 3 to day 6. For participant P7 (Appendix Fig. 3d), our model predicts a continuous decrease from day 0 to 11. This suggests effectiveness in predicting individual-specific disease progression trajectories. Though there is no reported severity of illness to validate the predicted probabilities, symptoms can be used as a reference. For participant P6, the number of symptoms increases from 3 to 6 at day 3 and decreases to 3 after. Therefore, it is reasonable to assume a worsening condition and an increase in the predicted probability, albeit with one day delay in prediction.

Similarly, the model can also predict individual-specific recovery trajectories. A sharper recovery trend is observed for participant P1 with a 49.2% relative decrease in 21 days (Fig. 5a) than for participants P4 (Fig. 6a) and P5 (Fig. 6b), with a 36.6% and 37.1% relative decrease in 21 and 22 days periods respectively. This is consistent with the evidence that recovery tends to be faster in younger people [21], as P1 is aged between 20-29 while P4 and P5 are aged between 30-39 and 40-49. Though it is difficult to draw statistical conclusions due to the limited number of participants, these results still suggest differences in the predicted recovery rate for different individuals. In terms of practical applications, the individual-specific recovery monitoring may be beneficial in providing prompt feedback to self-isolating patients, and more importantly, can provide treatment guidance for doctors according to each individual’s recovery status. Specifically, when a sharp decrease in the predicted probability is observed, it indicates that the individual is recovering well. Conversely, no decrease in the predicted probability over a long time period may require further or more effective treatment. Additionally, the predicted recovery trend could also be used to some extent for risk assessment of COVID-19 patients.

As our model has shown strong performance in using the longitudinal audio biomarkers, another important factor in the model deployment is the impact of sequence length, which is also analysed to provide insights into how many samples are needed for reliable predictions (Appendix Fig. 2). The cumulative histogram suggests that the longer the sequence, the better the performance. For sequence lengths with more than 2 samples (Fig. 2a) or around and more than 4 days (Fig. 2b), the model can produce reasonably good predictions. For telemonitoring purposes, the use of audio recordings from the last 4 days would offer a more reliable prediction.

Our study also has several limitations. First, the testing cohort is relative small with only 21 participants for the positive and negative groups respectively. This may not comprehensively represent the target population. In addition, the self-reported test results may inevitably be noisy to some extent, where a mismatch between audio recordings and test results may exist. This is due to possible delays in participants reporting the test results. This mismatch introduces confounding variability into the model development that is not fully taken into consideration.

The other limitation in our study is limited control over confounding factors. The age and gender groups are relatively balanced within and across the training, validation and test partitions, whilst language is only balanced between three partitions but still very unbalanced within each data partition. Our model using a multi-task framework mitigates the language
impact, but some language bias might remain due to the limited numbers of samples of some language subgroups.

We also observed that the predicted disease progression trend matches the test results, but for some users probabilities may be overall high or low over the course of COVID-19 disease progression. This suggests individual differences in audio characteristics. Though our model resolves this better than simple sample based models by capturing past audio signals, it is a universal model and therefore still imprecise. The development of participant-specific models is on our future agenda, but more data needs to be collected for this purpose.

In conclusion, having modelled longitudinal audio biomarkers with sequential machine learning techniques, we have proposed audio-based diagnostics with longitudinal data as a robust technique for COVID-19 disease progression tracking. We have showed that our system is able to monitor disease progression especially the recovery trajectory of individuals. This work not only provides a flexible, affordable and timely tool for COVID-19 disease tracking but it also provides a proof of concept of how telemonitoring could be applicable to respiratory diseases monitoring, in general.

**Methods**

**Data processing.** Data processing consists of data selection, audio pre-processing, sequence generation and data augmentation, aiming to increase the data size and generate effective sequence data for modelling.

**Data selection**

Positive participants are defined as those who have provided at least one audio recording alongside a positive test result, and negative participants are defined as those who have only ever tested negative. Since there are significantly more negative than positive participants, we first select the eligible positive participants to guarantee a relatively balanced dataset. We identified 168 positive participants who provide more than 4 samples, where the test results may also include negative (after recovery) and non-tested reports. Given the data was crowdsourced, we manually checked the label quality by removing the participants who reported contradictory labels, such as positive and negative test results reported on the same day, positive and negative test results alternating in a short period, etc. This resulted in 118 longitudinal positive participants. In addition, we also checked the audio quality to remove the poor samples using Yannet [27]: samples with background noise, clipped cough recordings due to larger amplitude than the maximum limit of mobile phones, etc. This resulted in 106 eligible positive participants.

For negative participants, similar selection criteria were used to identify eligible users. To generate a balanced dataset, we randomly selected 106 longitudinal negative participants from 447 participants.

**Audio pre-processing**

Audio recordings were first resampled to 16 kHz and converted to mono channel. These audio recordings were then pre-processed by removing the silence periods at the beginning and the end of the recording, and normalising them to have maximum amplitude of 1.

**Sequence generation**

For each participant, a sliding window of length 5 (samples) and shift 1 (sample) was used to segment the long sequence into short sequences, aiming to generate a large number of sequences for training as well as maintaining temporal dynamics within an effective period. A constraint was further applied to the short sequences, limiting the maximum time gap between two adjacent samples to 14 days. Any sequences that violated this constraint were removed. This guarantees that the sequence length ranges from 5 days to 56 days, with the median value of 18 days (refer to Fig. 3c).

**Data augmentation**

Three augmentation techniques were used to increase the data size and to balance the positive and nega-
tive samples. Though the dataset was selected to balance the number of participants for positive and negative subgroups, the number of samples for each participant is still different (refer Fig. 3c). The negative participants provided more samples than positive participants, and the positive participants also provided negative samples after recovery, leading to a significant larger number of negative samples than positive samples in the cohort. Thus, data augmentation was deemed necessary.

- The first augmentation technique is the commonly used Gaussian noise augmentation, where randomly sampled Gaussian noise is added to the original waveform. To balance the positive and negative samples to some extent, the noise augmentation was applied three times to the original positive group and only 1 time to the original positive group and only 1 time to the negative group.

- The second augmentation method is our proposed sequence augmentation, where 5 random samples are collected from each participant’s data pool and sorted in time order to generate new sequences. This increases the data size as well as enabling richer temporal dynamics. The same criterion of a maximum of 14 days between two adjacent recordings was applied for augmented sequences.

- The last augmentation method is the time inverse augmentation. Most of the reported sequences are consistently positive or negative, with only a limited number of samples reporting the transition from positive (negative) to negative (positive). The model might not effectively capture these transitions due to the limited data. Therefore, the segmented sequences were further time inverted during the training stage to enrich the dynamic changes.

All three augmentation techniques aided the model development by providing enough data, guaranteeing effective temporal dynamics, and increasing the various disease progression dynamics.

Model architecture. The proposed model consisted a pre-trained network of VGGish for feature extraction, and a recurrent neural network of Gated Recurrent Units (GRUs) for disease prediction (Fig. 7). Three different modalities including breathing, cough, and voice recordings are adopted as the input. For each modality, audio recordings are first converted to spectrograms, and fed into a VGGish pre-trained network for higher-level feature representations, which could help leverage and transfer the knowledge learnt from an external massive general-audio dataset [27]. The embeddings converted by VGGish from the three modalities are concatenated to form a multi-modal input vector for the subsequent GRUs based prediction network. The reason to choose GRUs over a Long Short-Term Memory neural network is due to fewer parameters in limited data size regimes. GRUs also use less memory and execute faster, which benefits potential model deployment. The outputs from the GRUs are evaluated in terms of two different tasks, one estimating the model capability in binary diagnosis by taking the binary output of the model, and the other predicting the disease progression trajectory by utilising the probabilities of positive predictions.

Feature extraction

The spectrograms were computed for 0.96 seconds for one recording, thus a series of spectrograms of each recording is converted to the same number of 128-dimensional feature vectors using VGGish. An average pooling layer is proposed to aggregate all vectors within one audio recording into one global latent feature vector. The global latent feature vectors obtained from each modality are concatenated to form a multi-modal feature vector. Due to the limited size of the data, the pre-trained network VGGish originally optimized for acoustic event detection [27] is frozen and not trained.

Irregular time intervals

The audio recordings for each participant are reported at irregular time intervals, where the days between consecutive audio recordings are not consistent. This is not compatible with GRUs modelling. Therefore, we assume any missing recording is the
same as the last recording and use forward imputation. This is carried out for each day, and employed for the feature embeddings after the VGGish and pooling layers as:

$$x_{t'} = x_t; t, t' \in [1, T]$$ \hspace{1cm} (1)

where $t < t'$. $t'$ can be any missing day and $t$ represents the days with recordings. In addition, we assume that the influence of the feature embeddings only have impact in a certain temporal context, and will fade away over time if the features have been missing for a long time. Therefore, a decay mechanism is designed for the feature embeddings, by incorporating the decaying factor in the additional feature dimension as:

$$\delta_t = 0.75^{\Delta t}$$ \hspace{1cm} (2)

$$\Delta t = t' - t$$ \hspace{1cm} (3)

$$\hat{x}_t = [x_t, \delta_t]; t \in [1, T]$$ \hspace{1cm} (4)

where $\delta_t$ is the decaying factor, which is 1 for the present audio features and exponentially decaying for missing features. The final feature representation $\hat{x}_t$ concatenates the original embeddings and the decaying factor.

**Sequential modelling**

One GRUs layer and one dense layer are cascaded and served as the COVID-19 prediction layers. 64 neurons are adopted for GRUs layer and 2 neurons for dense layers. Softmax is used as the activation function, generating the probability scores for positive and negative detection. The probability scores are i) categorized into binary outputs of positive and negative; and ii) adopted directly examining the probabilities of positive over time. These correspond to the two different tasks respectively (cf. Fig. 1).

**Multi-task learning** Since the data consists of eight different languages and the disease prevalence for

---

**Fig. 7 | Model structure.** A pre-trained CNN based model VGGish is used as the feature extractor, and Gated Recurrent Units (GRUs) are used as a classifier followed by dense layers, to account for longitudinal audio dynamics. It is a multi-task learning framework, with COVID-19 detection as the main task and language detection as an auxiliary task to avoid language bias. $h_i, i \in [1, 2, \ldots, N]$ represents the hidden vectors in the GRUs for time step $t_i$. The reverse layer is used for the language task, as shown in Eq. 5.
each language is different (cf. Fig. 3b), it might introduce language bias in the machine learning models, leading to the model recognizing the language instead of COVID-19 related information, e.g. classifying Italian speakers as positive and English as negative owing to higher prevalence in Italian-speaking users and lower prevalence in English-speakers. To reduce the language impact, a multi-task learning framework is proposed to include the auxiliary task of language recognition simultaneously with the COVID-19 prediction task. The subnetwork for language recognition consists of one dense layer (8 neurons) with softmax activation function, which takes the GRU outputs as the input and generate the language outputs. During the training phase, a reverse gradient layer is used to eliminate the language difference, with the loss as:

\[ L = \alpha L_{\lambda_1} - (1 - \alpha) L_{\lambda_2}, \]  

where \( L \) is the final loss, and \( L_{\lambda_1} \) and \( L_{\lambda_2} \) represent the weighted cross-entropy loss for COVID-19 detection and focal loss for language recognition respectively. \( \alpha \) is the scaling parameter that controls the balance between two subtasks.

**Model Training.** During the training process, all the model parameters were optimised on the development set and validated in the test set. During training, an Adam optimizer was used. The learning rate and momentum was set to be 1e-4 and 0.90, respectively. Early stopping was also used as no increase of ROC-AUC was observed in 5 successive epochs. Weighted cross-entropy was used as the loss for COVID-19 detection, with weight optimized in [1,5] with a step size of 1. Focal loss was used for language detection with \( \gamma = 2 \). For the multi-task learning, the weight \( \lambda \) was optimized within [0.5,1.0] with an interval of 0.1. Tensorflow 2.0 was used for the model development.

**Inference** As the sequence length in the training data varies within the range of [5,56] days, the model is able to capture longitudinal audio dynamics with varying length. Therefore, during the inference stage, the prediction is made given all the past audio recordings with no fixed number of samples. This is slightly different from the training phase but is more practical in real applications. In order to maintain the effective temporal dependencies of the audio recordings and match the training, a time constraint is applied to account only for the past audio recordings within 56 days before the current day, the maximum duration in the training phase. Further, we evaluate the model performance from the 2nd sample of each participant, to make sure the predictions capture the longitudinal audio dynamics.

**Evaluation metrics** The evaluation metrics are different for the two tasks. For COVID-19 detection, performance is measured using ROC-AUC, sensitivity and specificity. ROC-AUC illustrates the diagnostic ability of the binary classifier. Sensitivity shows the model’s ability to identify correctly positive patients, while specificity shows the model’s ability to identify correctly people without the disease.

For the disease trajectory monitoring, model performance is evaluated for each participant. There are two different metrics used based on the individuals’ test labels. For participants who have reported any transitions between positive and negative during the clinical course, we adopted Point-Biserial Correlation Coefficient \( \gamma_{pb} \) to measure the association of the predicted probability of positive and the test labels:

\[ \gamma_{pb} = \frac{\mu_1 - \mu_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \]  

Here \( \mu_1 \) and \( \mu_0 \) are the mean values of the predicted probabilities for the positive and negative participants, \( s_n \) is the standard deviation of the predicted probabilities for all the participants, \( n_1 \) and \( n_0 \) are the numbers of participants in the positive and negative groups respectively, while \( n \) is the total number \( n = n_1 + n_2 \). A higher \( \gamma_{pb} \) indicates a stronger correlation, thus a better disease progression trajectory prediction.

For the participants who reported positive and negative test results consistently over the clinical course, it is not possible to compute the correlation between the continuous predictions and the test labels. Therefore, we adopted the accuracy computed as the ratio of the correctly predicted samples \( N_i \) over the total
number of samples $N$:
\[ \gamma = \frac{N_i}{N} \] (7)

Data Availability
The data is sensitive, as voice sounds can be deanonymised. Anonymised data will be made available for academic research upon requests directed to the corresponding author. Institutions will need to sign a data transfer agreement with the University of Cambridge to obtain the data. A copy of the data will be transferred to the institution requesting the data. We already have the data transfer agreement in place.

Code Availability
Python code and parameters used for training of neural networks will be available on GitHub for reproducibility purposes.

References

[1] Cevik M, Kuppalli K, Kindrachuk J, Peiris M. Virolology, transmission, and pathogenesis of SARS-CoV-2. British Medical Journal. 2020;371:1–6.

[2] Vogels CB, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT-qPCR primer–probe sets. Nature Microbiology. 2020;5(10):1299–1305.

[3] Fan L, Liu S. CT and COVID-19: Chinese experience and recommendations concerning detection, staging and follow-up. European radiology. 2020;30(9):5214–5216.

[4] Deshpande G, Batliner A, Schuller BW. AI-based Human Audio Processing for COVID-19: A Comprehensive Overview. Pattern recognition:108289.

[5] Ates HC, Yetisen AK, Güder F, Dincer C. Wearable devices for the detection of COVID-19. Nature Electronics. 2021;4(1):13–14.

[6] Channa A, Popescu N, Skibinska J, Burget R. The rise of wearable devices during the COVID-19 pandemic: A systematic review. Sensors. 2021;21(17):5787.

[7] Miranda ID, Diacon AH, Niesler TR. A comparative study of features for acoustic cough detection using deep architectures. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2019. p. 2601–2605.

[8] Stasak B, Huang Z, Razavi S, Joachim D, Epps J. Automatic Detection of COVID-19 Based on Short-Duration Acoustic Smartphone Speech Analysis. Journal of Healthcare Informatics Research. 2021;5(2):201–217.

[9] Deshpande G, Schuller BW. Audio, Speech, Language, & Signal Processing for COVID-19: A Comprehensive Overview. arXiv preprint arXiv:201114445. 2020.

[10] Al Ismail M, Deshmukh S, Singh R. Detection of COVID-19 through the analysis of vocal fold oscillations. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2021. p. 1035–1039.

[11] Deshmukh S, Al Ismail M, Singh R. Interpreting glottal flow dynamics for detecting COVID-19 from voice. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2021. p. 1055–1059.

[12] Imran A, Posokhova I, Qureshi HN, Masood U, Riaz S, Ali K, et al. AHCOVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Informatics in Medicine Unlocked. 2020;20:1–20.

[13] Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD); San Diego, CA; Aug 22-27, 2020. 3474–3484.

[14] Laguarta J, Hueto F, Subirana B. COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. IEEE Open Journal of Engineering in Medicine and Biology. 2020;1:275–281.

[15] Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 Detection From Voice. IEEE Open Journal of Engineering in Medicine and Biology. 2020;1:268–274.
[16] Han J, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, et al. Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Toronto, Canada; June 6-11, 2021. 2021:8328–8332.

[17] Andreu-Perez J, Perez-Espinosa H, Timonet E, Kiani M, Giron-Perez MI, Benitez-Trinidad AB, et al. A Generic Deep Learning Based Cough Analysis System from Clinically Validated Samples for Point-of-Need Covid-19 Test and Severity Levels. IEEE Transactions on Services Computing. 2021:1–1. Https://10.1109/TSC.2021.3061402 (accessed Mar 1, 2021).

[18] Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller B. End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study. BMJ Innovations. 2021;7(2):356–362.

[19] Al Ismail M, Deshmukh S, Singh R. Detection of Covid-19 Through the Analysis of Vocal Fold Oscillations. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Toronto, Canada; June 6-11, 2021. 2021:1035–1039.

[20] Lechien JR, Chiesa-Estomba CM, Place S, Van Laethem Y, Cabaraux P, Mat Q, et al. Clinical and epidemiological characteristics of 1420 European patients with mild-to-moderate coronavirus disease 2019. Journal of internal medicine. 2020;288(3):335–344.

[21] Voinsky I, Baristaite G, Gurwitz D. Effects of age and sex on recovery from COVID-19: Analysis of 5769 Israeli patients. Journal of Infection. 2020;81(2):e102–e103.

[22] Wu J, Li W, Shi X, Chen Z, Jiang B, Liu J, et al. Early antiviral treatment contributes to alleviate the severity and improve the prognosis of patients with novel coronavirus disease (COVID-19). Journal of internal medicine. 2020;288(1):128–138.

[23] Yu F, Yan L, Wang N, Yang S, Wang L, Tang Y, et al. Quantitative detection and viral load analysis of SARS-CoV-2 in infected patients. Clinical Infectious Diseases. 2020;71(15):793–798.

[24] Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. The Lancet Infectious Diseases. 2020;20(8):911–919.

[25] Siddiqi HK, Mehra MR. COVID-19 illness in native and immunosuppressed states: A clinical-therapeutic staging proposal. The journal of heart and lung transplantation. 2020;39(5):405.

[26] Chen J, Qi T, Liu L, Ling Y, Qian Z, Li T, et al. Clinical progression of patients with COVID-19 in Shanghai, China. Journal of infection. 2020;80(5):e1–e6.

[27] Hershey S, Chaudhuri S, Ellis DP, Gemmeke JF, Jansen A, Moore RC, et al. CNN architectures for large-scale audio classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); New Orleans, LA; Mar 5–9, 2017. 131–135.

Acknowledgements
This work was supported by ERC Project 833296 (EAR) We thank everyone who volunteered their data.

Ethics
The study was approved by the ethics committee of the Department of Computer Science at the University of Cambridge, with ID #722. Our app displays a consent screen, where we ask the user’s permission to participate in the study by using the app. Also note that the legal basis for processing any personal data collected for this work is to perform a task in the public interest, namely academic research. More information is available at https://covid-19-sounds.org/en/privacy.html.

Author Contributions
AF, CM, PC designed the study. AH, AG, CB, DS, JC designed and implemented the app to collect the sample data. AG designed and implemented the server infrastructure. JH, TD, TX selected the data for analysis. DS, TD, TX developed the neural network models. TD conducted the experiments and performed the statistical analysis. TD wrote the main draft of the manuscript and generated tables and figures. JH, TX co-wrote the manuscript. All authors critically reviewed, contributed to the preparation of the manuscript, and approved the final version.

Competing Interests
All authors declare no competing interests.
Appendix

Fig. 1 | Data statistics in the training, validation and test partitions in terms of gender, age, and language. a, Gender; b, Age; c, Language. Gender and age are relatively balanced for three partitions as well as for positive and negative groups. Language is balanced for three data partitions, but still unbalanced within each partition between positive and negative participants.
Fig. 2 | The performance of COVID-19 detection with different sequence length in (a) samples and (b) days. X axis indicates the maximum sequence length of the subgroup, and y axis shows the accuracy for the sequences shorter than the maximum length.
Fig. 3 | Examples of strong disease progression predictions. **a and b**, Two recovery participants; **c and d**, Two positive participants P6 and P7 reporting consistent positive test results; **e and f**, Two negative participants. Wrong prediction at day 0 for participant in (f) is possibly due to the limited audio dynamics. It should be noted that some of the audio recordings are not provided with the test results, labeled with the white shaded area.
Fig. 4 | Examples of weak disease progression predictions. **a and b**, Two recovery participants. The predicted recovery trend is observed, while the categorized positive and negative predictions are not exactly matching the test results, which implies the potential personalisation. **c and d**, Two negative participants. Despite the accuracy of $\gamma = 0.79$ for participant c, the predictions fluctuate between positive and negative predictions, which suggests the possible solution of personalisation. Participant d was predicted positive for the 1st point with the probability of 0.55, which is possibly due to the inadequate audio dynamics. The predictions were predicted correctly after capturing richer audio dynamics by using a longer sequence.