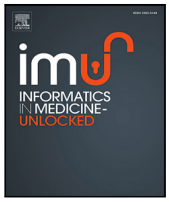
Informatics in Medicine Unlocked 46 (2024) 101478

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



iPREDICT: AI enabled proactive pandemic prediction using biosensing wearable devices

Muhammad Sajid Riaz [a,*], Maria Shaukat [a], Tabish Saeed [a], Aneeqa Ijaz [a], Haneya Naeem Qureshi [a], Iryna Posokhova [b], Ismail Sadiq [c], Ali Rizwan [c], Ali Imran [a,c]

[a] AI4Networks Research Center, School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, USA
[b] Hudson College of Public Health, University of Oklahoma, USA
[c] James Watt School of Engineering, University of Glasgow, Scotland, UK

ARTICLE INFO

Keywords:
Pandemic prediction
Wearable biosensors
Biosensing devices
AI for healthcare
Pandemic prognostics
Pathogen spread
iPREDICT

ABSTRACT

The emergence of pandemics poses a persistent threat to both global health and economic stability. While zoonotic spillovers and local outbreaks may not be fully preventable, early detection of infections in individuals before they spread to communities can make a major difference in containing an infectious disease and stopping it from becoming an epidemic and then a pandemic.

In this paper, we propose a novel Artificial Intelligence (AI)-based pandemic prediction framework called iPREDICT—a concept framework designed to leverage the power of AI and crowd-sensed data for accurate and timely pandemic prediction. The core idea of iPREDICT is to leverage the deluge of data that can be harnessed from connected and wearable biosensing devices. iPREDICT system then works by monitoring anomalies in the biomarkers at the individual level and correlating them with similar anomalies observed in other members of the community. Using AI-based anomaly detection in conjunction with analysis of the spatiotemporal growth of the correlated anomalies, iPREDICT thus can potentially detect and monitor the emergence of a local outbreak in near real-time to predict a potential pandemic.

However, not every outbreak has the potential to become a pandemic. We illustrate how tools like graph neural networks can be leveraged to determine optimal thresholds as a function of a large number of demographical, social, and geographical factors that determine the spatiotemporal spread of an outbreak, thus quantifying the risk of it becoming an epidemic or pandemic.

We also identify essential technological and social challenges that require attention to transform iPREDICT from an idea into a globally deployable solution for future pandemic prediction and management. To provide deeper insights into iPREDICT design challenges and trigger research towards possible solutions we present a COVID-19 based case study. The results signify the impact of variation in biosensing hardware, data sampling rate, and compression rate on the performance of AI models that underpin various components of the iPREDICT system.

1. Introduction

While COVID-19 is not the first pandemic in the 21st century, it is one of the most devastating ones taking millions of lives and annihilating trillions of dollars from the global economy [1–3]. A list of major pandemics in the last millennium is given in Table 1. In the US alone, the social and economic damage of the COVID-19 pandemic has surpassed that of all the natural disasters in the last century combined [4]. The absence of proactive and scalable outbreak detection or pandemic prediction mechanisms is one of the core reasons why pandemics spread and thus cause greater socioeconomic damage than natural disasters like hurricanes and tsunamis. Predictive or early detection systems for calamities have helped humanity in minimizing human fatalities and economic losses in the past. The prevention of the catastrophic repercussions of potential pandemics requires establishing a similar early detection system. Such a system needs to be scalable to allow global surveillance and detection of infectious disease outbreaks, and thus predict pandemics at the pre-emergence or local outbreak stage.

Previous epidemiological studies of the pathogenic diseases [14–17] and extensive insights from different coping strategies for the COVID-19 pandemic [18–23], provide evidence that the early stage detection of

* Corresponding author.
E-mail addresses: riazsajid@ou.edu (M.S. Riaz), maria.shaukat.1@ou.edu (M. Shaukat), t.saeed@ou.edu (T. Saeed), aneeqa@ou.edu (A. Ijaz), haneya@ou.edu (H.N. Qureshi), ali.imran@ou.edu (A. Imran).

https://doi.org/10.1016/j.imu.2024.101478
Received 10 January 2024; Received in revised form 8 March 2024; Accepted 15 March 2024
Available online 16 March 2024

Table 1
Significant human symptoms caused by global pandemics and related biological markers (biomarkers).

Pandemic	Russian Flu 1889–1893 [5]	Spanish Flu 1918–1920 [6]	Asian Flu 1957–1959 [7]	HIV/AIDS 1981-ongoing [8]	Swine Flu 2009–2010 [9]	Middle east respiratory syndrome (MERS) 2012-ongoing [10,11]	Ebola Virus 2014–2016 [12]	SARS-CoV-2 2019-ongoing [13]	
Pathogen		Influenza A/H1N8	Influenza A/H1N1	Influenza A/H2N2	Human Immunodeficiency Virus	Influenza A/H1N1	MERS-CoV	Ebola, Sudan, and Bundibugyo viruses	SARS-CoV-2
Symptoms	Fever and chills	✔	✔	✔	✔	✔	✔		✔
	Cough	✔	✔	✔		✔	✔		✔
	Sore throat	✔	✔	✔		✔	✔		✔
	Diarrhea	✔			✔		✔	✔	
	Breathing difficulty			✔		✔	✔		✔
	Delirium	✔	✔						✔
	Dizziness	✔	✔						✔
	Malaise	✔		✔		✔			
	Vomiting	✔					✔	✔	
	Headache	✔	✔						
	Skin rash				✔			✔	
	Muscle pain				✔			✔	
	Insomnia	✔							
	Heliotrope cyanosis		✔						
	Impaired color vision		✔						
	Blurred vision		✔						
	Nausea			✔					
	Eye redness			✔					
	Ocular pain			✔					
	Sore genitals				✔				
	Mouth ulcer				✔				
	Runny nose					✔			
	Stomach pain						✔		
	Internal bleeding							✔	
	Respiratory distress								✔
	Loss of taste/smell								✔
	Myocarditis								✔
Biomarkers	Heart rate variability	✔	✔	✔	✔	✔	✔	✔	✔
	Skin temperature	✔	✔	✔	✔	✔	✔		✔
	Cough recording	✔	✔	✔		✔	✔		✔
	Electrolyte imbalance	✔		✔	✔		✔	✔	✔
	Electroencephalogram (EEG)	✔	✔	✔		✔		✔	
	Blood pressure	✔	✔		✔			✔	✔
	Skin images	✔	✔		✔	✔		✔	✔
	Oxygen saturation (SpO2)			✔		✔	✔		✔
	Sweat rate	✔			✔		✔		
	Electrocardiogram (ECG)		✔		✔				✔
	Eye images/scans		✔	✔		✔			
	Photoplethysmography (PPG)		✔		✔				✔

the pandemic when it is just a local outbreak can be a game changer in containing the infection and preventing it from becoming a full-blown epidemic and then pandemic. The current laboratory-based diagnostic tests, that are often conducted after the infection has spread at the local level, do not offer the continual screening, agility, safety, scalability, and ubiquity to serve as a fast and proactive outbreak detection and thus a pandemic prediction and prevention system. Without such a system, COVID-19 cannot be expected to be the last pandemic of its scale and resultant catastrophic impact on the global health and economic system.

In Table 1 we provide an analysis of the major pandemics in the last millennia in terms of the common symptoms they have presented among humans. We note that most infectious diseases present symptoms that can be detected and monitored through commodity wearable or ambient sensors. The biomarkers that can be measured to screen for these symptoms and thus detect an infection are also identified in this table. With advances in biosensing, nanotechnology, and wireless communications most of these biomarkers can be measured nonintrusive at population level and analyzed centrally. This observation combined with promising results and the impact of our seminal work [23] on screening for COVID-19, anytime anywhere just from the cough sounds by using an app installable on any commodity phone or watch, motivates us to propose iPREDICT (see Fig. 1 for the schematic of the iPREDICT) an innovative framework that can enable in-situ and continuous screening at the population level to detect a new outbreak of an existing or a new disease at an early stage thus serving as potential pandemic prediction and prevention system that world direly needs.

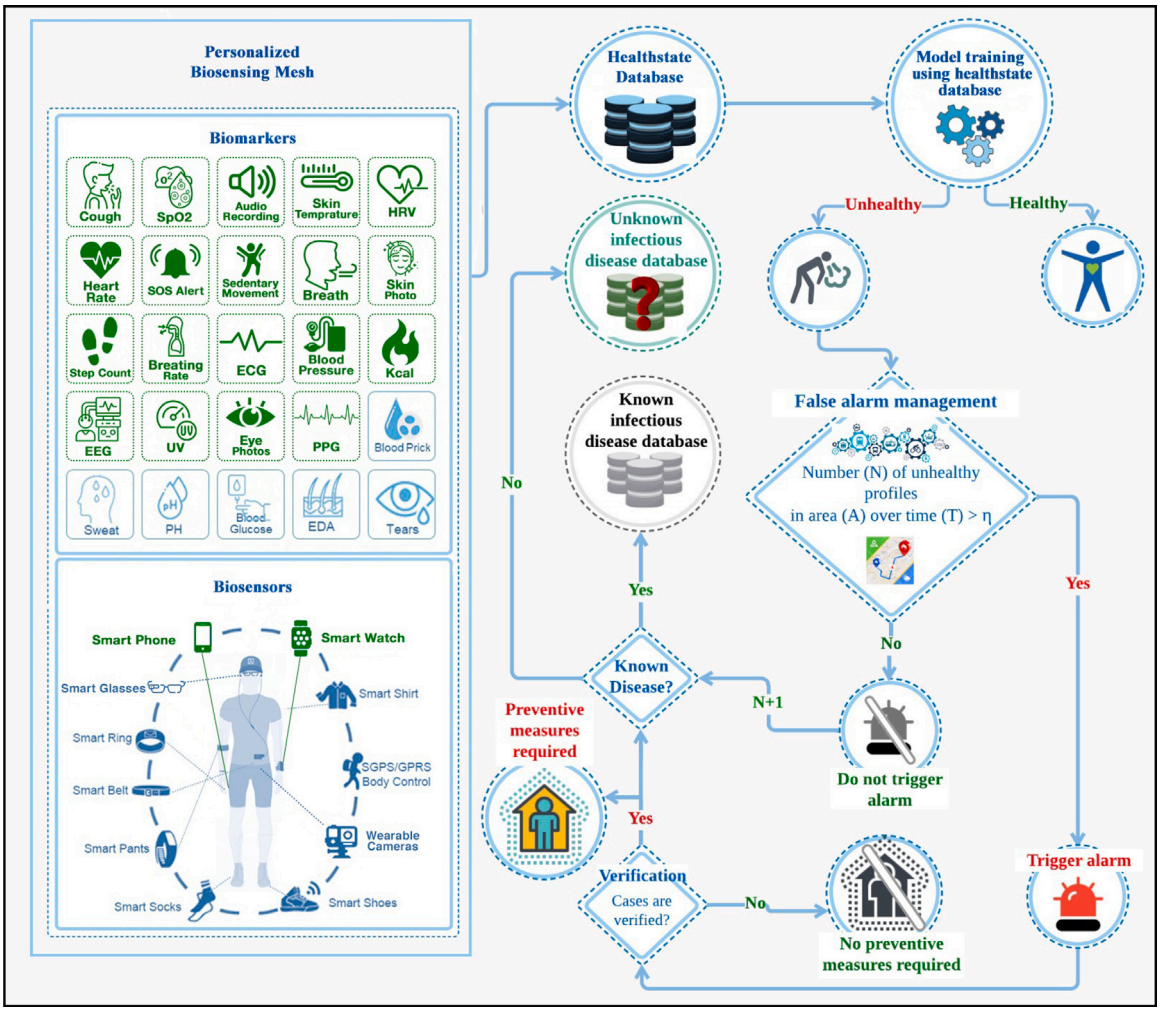


Fig. 1. iPREDICT: An AI-enabled proactive pandemic prediction framework using wearable biosensing devices. *Biomarkers: SpO2(oxygen saturation), HRV (heart rate variability), ECG (electrocardiogram), Kcal (kilocalories burnt), EEG (electroencephalogram), UV (ultraviolet exposure), PPG (photoplethysmography), pH (saliva pH), EDA (electrodermal activity). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1.1. Related work

The topic of Artificial Intelligence (AI)-based pandemic prediction and preparedness is still in its infancy. The complex and multidisciplinary challenges faced by researchers interested in pursuing this topic that includes but is not limited to: the novelty of each pathogenic spillover that usually leads to an outbreak; a large number of environmental, social, cultural, behavioral, demographic, political, and geographical factors that determine if a local outbreak will become pandemic; and the lack of required large multi-modal data to train and test AI-based solutions are among the most common challenges. Nevertheless, in the wake of COVID-19 some reviews have emerged that highlight the need to use AI for future pandemic prediction and technology-based interventions [24–29]. In [24] a survey of human and technology-based intervention methods is presented to control the spread of the virus among humans, specifically COVID-19. In [25] the authors explore the use of state-of-the-art bio-sensing technologies for disease diagnosis and management. In [26] the authors emphasize the importance of interdisciplinary collaboration and the need for developing rapid and accurate diagnostic tools by combining existing research efforts in engineering, medicine, chemistry, and biosensing technologies. In [30] a fairly latest research authors proposed the use of a wearable device (near field communication (NFC)-based wristband) to monitor individual biomarkers like oxygen saturation (SpO2), and body temperature to detect and manage the spread of COVID-19.

These studies [31,32] have identified the need and potential for developing an AI-based platform to control the spread of viruses among humans. However, primarily these studies are focused on COVID-19 and not future pandemics, therefore these solutions are more of a COVID-19 spread (hotspot) detection and not the pandemic prediction. Furthermore, current literature lacks an in-depth understanding of the challenges involved in designing a proactive pandemic prediction framework. Two critical factors that exacerbate the problem of pandemic prediction are: (1) Though similar in nature each pathogen is different based on the medium of spread, reproduction rate, incubation period, and virus life. (2) The symptoms of disease among different infected individuals will be different based on their age group, gender, and lifestyle. Due to the aforementioned challenges, even if we try to generalize the current solutions for future pandemics, data from the previous pandemics alone would not suffice for training an accurate AI-based solution. However, current research [30] proposes acquiring free-life individualized data for robust AI-based model training for infection detection, which has the potential to be a game changer in proactive pandemic prediction.

1.2. Contributions and organization

To address the limitations in the current literature highlighted in Section 1.1, we propose a novel surveillance system named iPREDICT that can detect broiling infection spread and alert the authorities to take

action to prevent future pandemic development. iPREDICT proactively monitors the pre-emergence state of a pandemic by leveraging recent advances in wearable biosensing, AI, and ubiquitous wireless networks.

The core idea is to collect the big data of biomarkers of the population by leveraging readily available and widely used wearable and ambient devices/sensors with ubiquitous wireless connectivity. The biosensing devices encompass a broad range including smartphones, smartwatches, wristbands, sensor patches, and a growing variety of biosensors that can be embedded in clothing, shoes, and other accessories [33–36]. Combining this data with an optimally designed and trained backend AI-engine, and intelligent analytics can help to detect an early infectious outbreak by monitoring the number of individuals infected by a specific virus, at a certain geolocation in a given time.

The key contributions of this work can be summarized as follows:

1. We present iPREDICT, an AI-powered proactive pandemic prediction framework that uses wearable biosensors. The framework integrates expertise from various fields such as AI, epidemiology, and distributed system software development, to provide a comprehensive solution for accurate prediction of future pandemics. iPREDICT acquires essential biomarkers from free-life biosensors, analyzes the transmission pattern of infectious diseases based on location, and employs AI algorithms to raise alerts and prevent the rapid spread of the disease and potential pandemic outbreaks.
2. We provide a comprehensive overview of global pandemics from the last millennium to serve as a foundation for iPREDICT. The survey focuses on the pathogens that caused the previous pandemics, their predominant symptoms in humans, and the relevant biomarkers that can be autonomously used to track the symptoms of epidemic-inducing diseases. Therefore iPREDICT can use historical data from past pandemics, and the correlation between disease symptoms and biomarkers to effectively predict and prevent future pandemics.
3. We propose an approach that employs graph neural networks (GNNs) to determine the pandemic prediction threshold, taking into account various environmental, geographical, and biological parameters. As the problem is complex, with no existing mathematical model that includes all these parameters, the proposed approach uses historical pandemic data to build a GNN-based framework capable of predicting epidemic thresholds at different scales and resolutions of the population. The expertise of the authors in applying AI in different domains informs the development of this method.
4. We present several crucial challenges both social and technological that must be addressed in the widespread deployment of iPREDICT. With a strong focus on the engineering challenges within our research domain, which include AI, signal processing, and cellular networks. The challenges we address include the collection of audio data (cough sounds) using various smartphone devices, at different audio sampling rates (for the efficient storage of audio data, which is critical for large-scale systems), and the transfer of audio data in different file sizes and formats, over cellular networks for analysis and diagnosis.
5. We demonstrate the feasibility of iPREDICT by leveraging our previous work AI4COVID-19 [23] as a case study and provide an analysis of the quantitative impact of four different engineering challenges on the performance of AI4COVID-19 by considering one biosensor (microphone) and one biomarker (cough sound) out of a massive list of available biosensors and biomarkers in a variety of biosensing devices, see Fig. 1.

The rest of the paper is organized as follows: iPREDICT and its respective components, data privacy challenges, and respective mitigation methods are briefly discussed in Section 2. A detailed discussion of a few of the salient engineering challenges for developing iPREDICT is provided in Section 3. The feasibility of iPREDICT with COVID-19 case study while addressing the associated engineering domain-specific challenges is provided in Section 4. Section 5 concludes this study and mentions future research directions.

2. iPREDICT: Proposed methods of (AI)-based pandemic prediction framework and study design

iPREDICT presents a comprehensive framework given in Fig. 1 for future pandemic prediction comprising four integral components. First, a personalized biosensing mesh forms the foundation, enabling real-time data collection. Second, a curated array of biomarkers, efficiently gathered through the biosensing mesh, facilitates intricate health assessments. Third, the synergy of AI models leverages individual biosensor data streams to facilitate personalized training, enhancing predictive accuracy. Fourth, by diligently analyzing these streams, the framework adeptly identifies burgeoning anomalies through the AI models' predictions, thereby enabling timely outbreak alarms, exemplifying iPREDICT's potential in proactive pandemic prediction. A detailed description of the individual components of iPREDICT is provided in the following subsections.

2.1. Personalized biosensing mesh

Within the iPREDICT framework, we propose the "personalized biosensing mesh" as a pivotal component, which capitalizes on the capabilities of diverse wearable devices such as smartwatches and smartphones. By ingeniously integrating these commodity wearables, a comprehensive biosensing ecosystem is forged, capable of capturing an array of vital biomarkers highlighted (in green color and dotted border) in Fig. 1 such as skin temperature, SpO2, audio recording, heart rate variability, and cough sounds (proposed an even detailed list of biomarkers in Fig. 1. Moreover, a description and how these biomarkers can be used as a symptom for the detection of various diseases is presented in Table 2). These biomarkers can be acquired using readily available wearable devices through their built-in biosensors [30,37].

In iPREDICT we propose the use of smartwatches and smartphones as wearable devices due to their usage convenience, acceptance, and availability to the masses. These wearable devices encompass an assortment of sensors, each equipped to measure specific biomarkers. For instance, heart rate sensors embedded in smartwatches meticulously track pulse rate variability and resting heart rates [38]. Accelerometers, commonly featured in smartphones can monitor movement patterns and quantify activity levels. Also, both the devices have microphones that can collect cough and audios that is used for respiratory disease diagnosis with impressive results [23]. Moreover, cutting-edge wearables incorporate photoplethysmography (PPG) sensors that ascertain blood oxygen saturation, while electrodermal activity sensors gauge stress levels. Temperature sensors integrated into devices like smartwatches serve as sentinels, detecting fluctuations indicative of fever or irregularities [39].

By harnessing this confluence of wearable devices and their inherent biosensors, a rich and diverse multimodal data stream is cultivated. The cough sound data is analyzed for the preliminary diagnosis of several respiratory diseases like Bronchitis, Pertusis, and COVID-19 [23]. Likewise, heart rate data, culminate in a comprehensive portrayal of an individual's activity levels and overall health. We highlighted a few of the biomarkers (cough, audio i.e counting, heart rate) that showed promising results in COVID-19 screening [23,40]. This cumulative biomarker dataset serves as the foundation for constructing a multidimensional individual health profile, emblematic of the iPREDICT framework's prowess.

2.2. Creation of biomarker profiling

The next component of iPREDICT is the creation of a comprehensive database of historical biomarkers of the population, we call it Health-state database in the iPREDICT framework presented in Fig. 1. Table 2

Table 2
Description of different biomarkers for disease diagnosis.

Biomarker	Description
Breath	Provides latent information via sound, smell, and intensity about a person's health. [41]
Audio Recording	Carries latent features that can be used in the acoustic analysis of respiratory diseases. [23]
Step Count	Provides insights about the lifestyle of a person that can relate to overall health. [42]
Burnt Kilocalories	Can be associated with cachexia which can be caused due to cancer or other chronic diseases. [43]
SOS Alert	Can be used in emergency cases when a person's vital signs fall outside a normal range. [44]
Heart Rate	Is a major biomarker for respiratory disease diagnosis. [41]
Sweat	Can be used for cystic fibrosis that causes damage to lungs and digestive system. [45]
Sedentary Movement	Can be used for cardiovascular disease diagnosis. [46]
Skin Temperature	Is a major symptom of the diseases that cause fever. [47]
Skin Photos	Provides information about allergic viruses.
Heart Rate Variability	Is a major biomarker for cardiovascular disease diagnosis. [41]
Oxygen Saturation (SpO_2)	Can be used as a biomarker for respiratory disease diagnosis such as COPD. [48]
Electrocardiogram	Is widely used biomarker for coronary heart disease diagnosis. [41]
Blood Pressure	Is a basic biomarker used by physicians for cardiovascular disease diagnosis. [41]
Saliva pH	Is used as a biomarker for stress examination and monitoring. [49]
Blood Glucose	Is a commonly used biomarker for the diagnosis of diabetes. [41]
Cough	Contains the signature of several respiratory diseases e.g. asthma, pertussis, bronchitis etc. [23]
Breathing Rate	Can be used as a biomarker for several diseases and conditions such as asthma, COPD, and pneumonia. [41]
Retinal Images	Is used as a biomarker for the diagnosis of diseases like chronic kidney problems and anemia. [41]
Electroencephalogram	Is a widely used biomarker for neurodegeneration problems like epilepsy, sleep disorders, and brain injuries. [41]
Photoplethysmograph	Is a biomarker used for cardiovascular disease detection. [50]
Ultraviolet Exposure	Can be used as a biomarker for systemic oxidative stress. [51]
Electrodermal Activity	Is used as a biomarker for the diagnosis of anxiety disorder and Parkinson's disease. [52]
Tears	Contains useful information in the fluid that can be used for the diagnosis of ocular and breast cancer. [53]

presents a list of biomarkers and the respective description of what latent information these biomarkers provide which can be exploited for the disease diagnosis. Healthstate database consists of biomarker measurements of individuals either labeled as 'healthy/normal' or as 'not normal' i.e., profiles of individuals that have been pre-identified to have some medical condition. These profiles will enable the detection of anomalous data points that lie within the health data stream of the individuals. However, biomarker profiling comes with several challenges brought by the variability and complexity of the biomarker data. These challenges can be broadly categorized into two categories explained below:

1. **Challenge 1: Inter-person Variability**: The creation of a personalized biosensing mesh comes with a complexity challenge. One way to address this is to create the models on edge devices, and only send triggers along with select when anomaly is noted for central examination by AI and or medical and public health professionals. To cope with the low computational power of the edge devices, instead of advanced deep learning (DL) models, simpler template matching methods can be used.
2. **Challenge 2: Intra-person variability**: Non-infectious diseases, seasons, lifestyle changes, and stress can cause variations. Addressing this challenge requires not only the fusion of multiple biomarkers that reflect a multi-system state of the human body but also deep medical expertise. As an example, HRV is a biomarker that drops usually with the onset of most type of sickness. However, these sicknesses may not be the cause of concern for the iPREDICT system as they may not be infectious. Therefore, a reliable method to detect the spread of an infection is to have a higher-level model, that looks for patterns of anomalies among people who have been in close proximity. For example, if HRV of multiple people who have been in close contact starts dropping within a time window, then it can be considered as a case for further analysis for iPREDICT system. This further analysis is carried out in iPREDICT components described in the next sections.

2.3. AI-based anomaly detection using biomarker profiles

In tandem with the challenges highlighted in the previous section, the high dimensionality (multiple biosensors capturing multiple biomarkers) and the large volume of data in an individual's biomarker profile add to the complexity of the anomaly detection component of iPREDICT. Due to such challenges, we propose a potential disease outbreak to be modeled as a time series anomaly detection problem. To achieve this, we propose a novel mechanism for identifying anomalous readings at an individual's biomarker level for detection of viral infections, at their onset. The biosensor time series data will be used to train an AI model for identifying individuals with biomarker levels deviating from their normal trend. Thus, the trained AI models will be patient-specific, mitigating the effects of intra as well as inter-personal variability and promoting precision medicine. Time series anomaly detection can be achieved using several machine learning (ML) models such as ARIMA, SARIMAX, etc. [54], and DL models (e.g., recurrent neural networks (RNNs), long short-term memory (LSTM) [55], and autoencoders [56]). The predictive results from these models will identify a potential disease outbreak and suspected individuals will be further tested to verify if a cluster of such anomalous data is present in close spatio-temporal proximity.

2.4. Adaptive thresholding for disease prevalence

This component of iPREDICT identifies the trends of similar irregularities in the biomarker values of multiple individuals residing in close proximity over a brief time duration. Once the cluster of infected people is identified, iPREDICT performs as explained in algorithm 1, triggering an alarm based on a disease-specific threshold to alert the authorities about a potential outbreak. We propose a pandemic threshold η based on several magnitude/number (N), area (A), and time (T) of infection parameters given in Fig. 2. The threshold η is modeled based on **NAT** in Eq. (1).

$$\eta(N_d, A, T) > \eta_d \tag{1}$$

Where N_d represents the number of infected individuals by the disease d, A is the area under consideration and T represents time, while η_d represents the alarm threshold of a specific infectious disease.

Finding the quantitative value of η_d is a challenging task for the epidemiology research community. An even bigger challenge lies in adaptively setting this threshold to minimize the intervention time for the authorities to take necessary measures. While we know from the epidemiology literature [57–61] that **NAT** depends on a wide range of factors as listed in Fig. 2, we do not have a quantitative representation of **NAT** that includes all the factors of Fig. 2, and developing a quantitative understanding will take decades of research by the

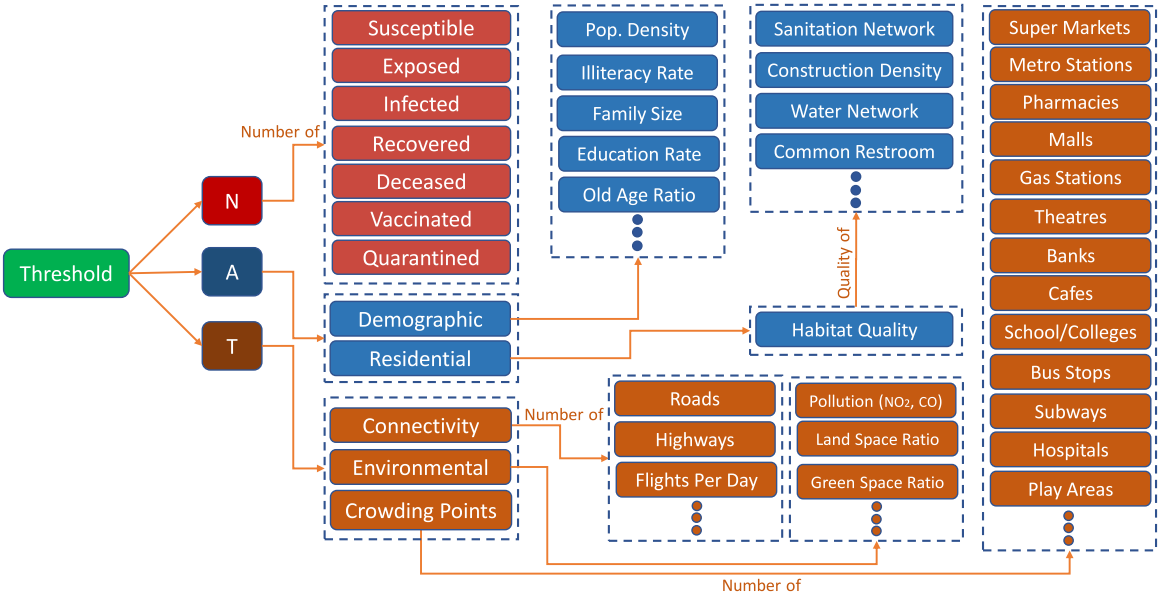


Fig. 2. A qualitative representation of the pandemic threshold based on NAT using associated parameters from epidemiology.

epidemiology research community or we may never know its closed-form mathematical equation. The challenge is due to the diversity of the nature of infectious diseases (reproduction rate, nature of spread (airborne or touch), association with other infectious diseases, etc.) and general human behaviors (mobility pattern, response to intervention policies, etc.). Therefore, leveraging the capability of AI to learn such complex relationships between a large variety of parameters and the availability of data on the recent pandemics, we propose a GNN-based framework for alarm management as the next component of iPREDICT.

Algorithm 1 False alarm management using a disease-specific threshold

Require: number of infected profiles N_d, in an area A over a time T and a infectious disease specific threshold η_d. Sig_d disease signature based on respective biomarker profiles. DB_k database containing known disease profiles, and DB_u containing unknown/novel disease profiles.

Ensure: Boolean value of $TriggerAlarm$

$N_d \leftarrow 0$
while $f(N_d, A, T) \leq \eta_d$ **do**
  $N_d \leftarrow N_d + 1$
  $TriggerAlarm \leftarrow False$
  if $f(N_d, A, T) \leq \eta_d$ **then**
    if Sig_d **Belongs to** DB_k **then**
      $DB_k \leftarrow Sig_d$
    else if Sig_d **Belongs to** DB_u **then**
      $DB_u \leftarrow Sig_d$
    end if
  else if $f(N_d, A, T) > \eta_d$ **then**
    $TriggerAlarm \leftarrow True$
    if Sig_d **Belongs to** DB_k **then**
      $DB_k \leftarrow Sig_d$
    else if Sig_d **Belongs to** DB_u **then**
      $DB_u \leftarrow Sig_d$
    end if
  end if
end while

2.5. Population resolution and scale-agnostic graph neural network system for alarm management

To overcome the challenges highlighted and discussed in the previous section, we propose an AI-enabled data-driven approach to learn pathogen and population dynamics-specific values for η_d by learning from historical data of epidemics. We aim to take benefit from the recent advances in DL on graphs, i.e., Graph Neural Networks [62] which learn to predict the η_d by performing convolutions on a graphical representation of the population and its features listed in Fig. 2.

Extensive literature in epidemiology exists where historical epidemic data is used for forecasting the future state of an epidemic. Firstly, compartmental models such as SIR [63], SIERD [64], and SIRV [65], etc., comprise of systems of ordinary differential equations which predict epidemic parameters and spread. Secondly, ML models like SARIMA [66] predict future infection rates through time series forecasting on past infection rates. Thirdly, time series forecasting is also performed via Deep Neural Networks (DNNs) such as LSTMs [67–69]. However, these approaches rely solely on the temporal aspect of the epidemic, i.e., historical infection rates, while not accounting for the spatial dynamics of population such as density, distribution, inter-mobility, and population characteristics like hygiene, humidity, etc., which can be vital in driving an epidemic. This is evident from [70], where integrating rainfall data with infection rates significantly improved the forecast of Dengue, since humidity and stagnant water caused by rain breed Dengue carrier mosquitoes. Similarly, [71] shows that meteorological factors like atmospheric pressure positively influenced the forecast of Influenza B and [72] examined the influence of mobility data on Influenza spread modeling. Despite such studies advocating for the efficacy of spatial features in epidemic prediction, an all-encompassing predictive model with spatial as well as temporal features is yet to be established. Recently, several studies [73–78] emerged where a population is modeled as a knowledge graph such that it captures the temporal characteristics of population as graph node features and spatial dynamics as well as mobility as graph structure i.e., adjacency matrix. Such graphical modeling aligns with the widespread use of graphs in epidemiology where spot maps, heat maps, and area (Patch or Choropleth) graphs are employed to illustrate the geographical spread of outbreaks on a 2D plane [79]. However, the representation of such maps as knowledge graphs which can train DL models, and the DL on graphs with Graph Neural Networks are emergent research directions in AI which have demonstrated improved prognostic capability over classical ML and DL for epidemic forecasting [73–78]. Therefore, based on (1) the limitation of compartmental, ML, and DNN models to capture population features and mobility, (2) the conventional capability of graphs in epidemiology to encapsulate these factors effectively, and (3) the recent advancement in DL on

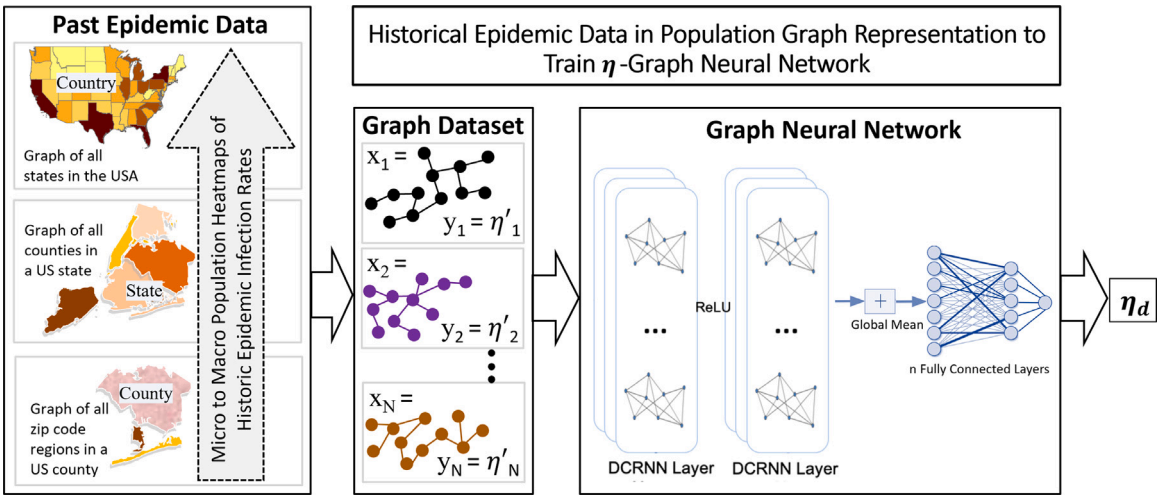


Fig. 3. A Graphical modeling of historical epidemic data and training of Graph Neural Network to predict η_d.

graphs to build predictive models from spatio-temporal graph datasets, we suggest a GNN model that learns from the dynamic graphical representation of populations during past epidemics to predict η_d future epidemics.

In [73,75,77,78] Graph neural networks embed graphs of US counties as low-dimensional latent embeddings which capture the population characteristics. As graph features (e.g., infection rates) vary against time, the embeddings of population graphs from past time $\{t-d, t-d+1, t-d+2, \ldots, t\}$ are treated as a time series with either Transformer or LSTM for forecasting of future $t+1$ infection rates. However, these approaches do not take into account the resolution and scale of the population graphs as variables. As the resolution increases from country to state to county and further, the properties of the population graphs change and hence a GNN model trained with data exclusively at low resolution (such as at the state level) cannot be employed at high resolution such as zip code and vice versa. In [78] and [77], the authors acknowledge resolution as a significant variable, but the multi-resolution nature of their model comes from the clustering of graph nodes and making condensed graphs from the pooled features of the clustered nodes. However, the clustering of graph nodes is data-driven therefore it disregards the natural clustering of regions due to standard geographic divisions e.g., all counties in one state can exhibit similar characteristics due to proximity, inter-mobility, cultural and environmental similarities, so they should be clustered together. The clustering of regions based on geography takes advantage of Tobler's first law of geography [80], which states that spatially closer regions have higher similarity than spatially distant regions. Such geography-aware clustering, therefore, eliminates the need for training data and hyper-parameter search required in data-driven clustering. In addition to resolution, the scale of the graphs is a bottleneck. For instance, [77] makes a graph with all the counties in the US as nodes. Given that there are 3142 counties and equivalent regions in the US, one graph will contain as many nodes and up to 4.9 million edges. Further increasing the resolution will result in a graph of 40,000 nodes = no. of five-digit zip codes in the US [81]. On the other hand, the lower the resolution, the smaller the graph but the crucial early-stage infection data is lost. For instance, if all the US states are modeled as a graph of 50 nodes (high scale, low resolution), then the pathogen breakout can be detected when it is already epidemic across states while the goal should be early detection during spreads over one county. Therefore, the amount of area covered in a graph, i.e., the scale of the graph should be inversely proportional to the resolution of the graph. Building on this understanding, we propose a multi-resolution multi-scale hierarchical approach for modeling populations as graphs and training an end–end resolution and scale-agnostic GNN which learns to predict pandemic alarm threshold η_d from population features listed in Fig. 2.

We divide the population in a nested manner based on the Standard Hierarchy of Census Geographic Entities [82], from micro to macro-region i.e., ZIP Code Tabulation Areas (ZCTA), county, and state. The set of all ZCTAs in one county forms a county-level graph as shown in Fig. 4. Similarly, all counties in a state form one state-level graph, and so on. Therefore, the total number of spatial graphs is, $S_{total} * C_{avg} * Z_{avg}$ where $S_{total} = 50$ for the states in the US, C_{avg} is the average number of counties in US states, and Z_{avg} is the average number of ZCTAs in US counties. For each geographic level, there can be multiple temporal graphs that have the same spatial structure but vary in node/edge features as each temporal graph consists of historical data from time t to $t+T$ where T is the time span of a week.

To summarize, the historical epidemic data is to be modeled into a set of graphs $\{G_h^t | h \in H, t \in T\}$ where H is the spatial granularity/hierarchy such as {ZCTA, County, State} and T is time granularity of data such as $\{week_0, week_1, \ldots, week_T\}$. Each graph G in G_h^t is represented as $G = (V, E)$ where V is the set of nodes representing regions at hierarchy h and E is the set of edges between nodes. Each node has a set of features $X = \{x_0, x_1, \ldots, x_k\}$ which represent **NAT** features listed in Fig. 2. Hence, in node features, we combine a plethora of data sources which together affect the risk of pathogen breakouts. Historical data of past epidemics and pandemics consisting of the number of infected, susceptible, recovered individuals, etc., over time in a region form the dynamic node features. Properties of the population, such as census data, population density, death/birth rate, poverty, literacy rate, age and gender demographics, etc. along with environmental factors which highlight the population's limitations and resources such as weather, pollution, and terrain form the static features. One node thus consists of all the relevant features to qualify the breakout within that node i.e., a population section (such as Zip code no. 11005 or the Queens county, depending on whether the population resolution is zipcode-level or county-level). To join nodes with edges, we determine the geographical connectivity between regions by using spatial distance as well as road network density and borders between the regions which are modeled as nodes. For each edge, the weight e is a function of human mobility pattern from Facebook Data for Good [83] which uses the location history from mobile devices to track air, road, or train travel between two regions and also specifies normal mobility ranges of communities, cohabitation and co-movement of groups. So, in summary, the graphical modeling of historical epidemic and pandemic data is such that the node features represent all the variables that can quantify pathogen spread within a region while edges and edge weights represent the variables that account for the spread of a pathogen from one region to another.

The most significant aspect of this geographically nested graph dataset is that η_d from the lower hierarchy serves as a node feature in

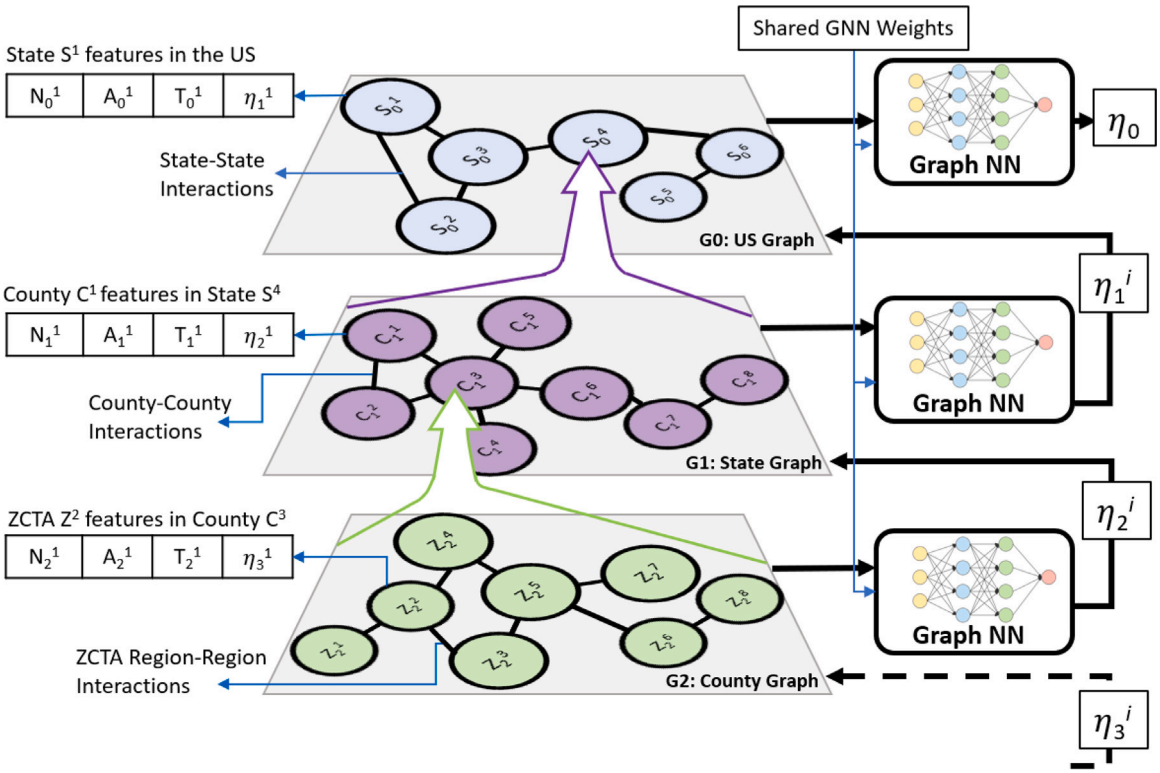


Fig. 4. Representation of multi-hierarchy and multi-scale modeling of populations into graphs where η_d is predicted from lower hierarchy graphs inform the prediction of η_d at higher hierarchy.

the higher hierarchy graph as depicted in Fig. 4. In the training dataset, the η_d at each hierarchy level comes from the averaging of η_d of nodes at the lower level, e.g.,

$$\eta_{NY\,State} = \frac{\sum_{i=1}^{C_{NY\,State}} \eta_i}{|C_{NY\,State}|} \tag{2}$$

Where $C_{NY\,State}$ is the number of counties in the New York state. At the inference time, however, the lower hierarchy η_d comes from the trained GNN which learns to predict η_d at varying resolutions and scales.

This dataset of spatio-temporal graphs described above is ingested by a Graph Neural Network (GNN) as shown in Fig. 3, specifically Diffusion Convolution Recurrent Neural Network (DCRNN). DCRNN, introduced by Y. Li et al. [84] is a type of GNN designed for addressing spatio-temporal forecasting tasks. It combines diffusion convolutional layers and recurrent layers to capture spatial and temporal dependencies, respectively, and integrates them into a unified framework. The diffusion step captures spatial dependencies by propagating information through the graph structure of the data. It allows each node to aggregate information from its neighboring nodes. The recurrent step captures temporal dependencies by incorporating historical information from previous time steps using a recurrent architecture, such as a Gated Recurrent Unit (GRU). The matrix multiplication in GRU is replaced with the diffusion convolution described above, thus integrating the diffusion, convolutional, and recurrent steps in DCRNN, i.e., effectively modeling both the spatial dependencies among different locations in the graph and the temporal dependencies over time. DCRNN transforms the input node features into lower dimensional embedding in latent space. The embeddings are optimized at every training step to best capture the information from node features and node neighbors. The latent embeddings from all the nodes are then combined by either concatenation, mean pooling or trainable pooling layers such as hierarchical pooling [85] and Self-attention graph pooling [86]. Then, the pooled embedding is passed through fully connected neural network layers to finally output η_d, a real number that embodies the threshold for **NAT** parameters such that when $\eta(N_d, A, T) > \eta_d$, the alarm is triggered in the system as shown by Fig. 1. As GNNs are independent of the graph structure, a GNN such as DCRNN trained on graphs of multiple resolutions and scales, can learn features that are resolution and scale-agnostic. Hence, the resultant trained GNN can be deployed at ZCTA, county, or state level with the shared weights as shown in Fig. 4.

2.6. Verification and preventive measures

Once the adaptive threshold is learned and the false alarm maintenance block accurately triggers the alarm when needed, it is essential for the respective authorities to verify the presence of the spreading disease. Among the methods that can be used for verification include laboratory testing of potentially infected individuals. Moreover, it is imperative to classify if the spread is from a disease that although infectious, has no risk of escalating as a future pandemic, or a known/unknown disease that can break out into a pandemic. The pathogens that have the risk of evolving into a pandemic can be either re-emerging or newly emerging. For the re-emerging pathogens, it is crucial to inform the health authorities at the earliest as the preventive measures required to curtail its spread are well-defined and priorly known. Among these preventive measures include social distancing, traveling constraints at this geolocation, promoting better personal hygiene measures, and could also include medication based on prior experience. On the other hand, if the alarm is triggered by a newly emerged pathogen, it becomes essential to alert experts in the fields of pathology, virology, and epidemiology. Because as their research and input expertise on the characteristics (spread pattern, reproduction rate, and mode of spread i.e. airborne or using touch) of the newly emerged pathogen become the basis for our next steps in pandemic outbreak prevention and will also populate the database of unknown infectious diseases.

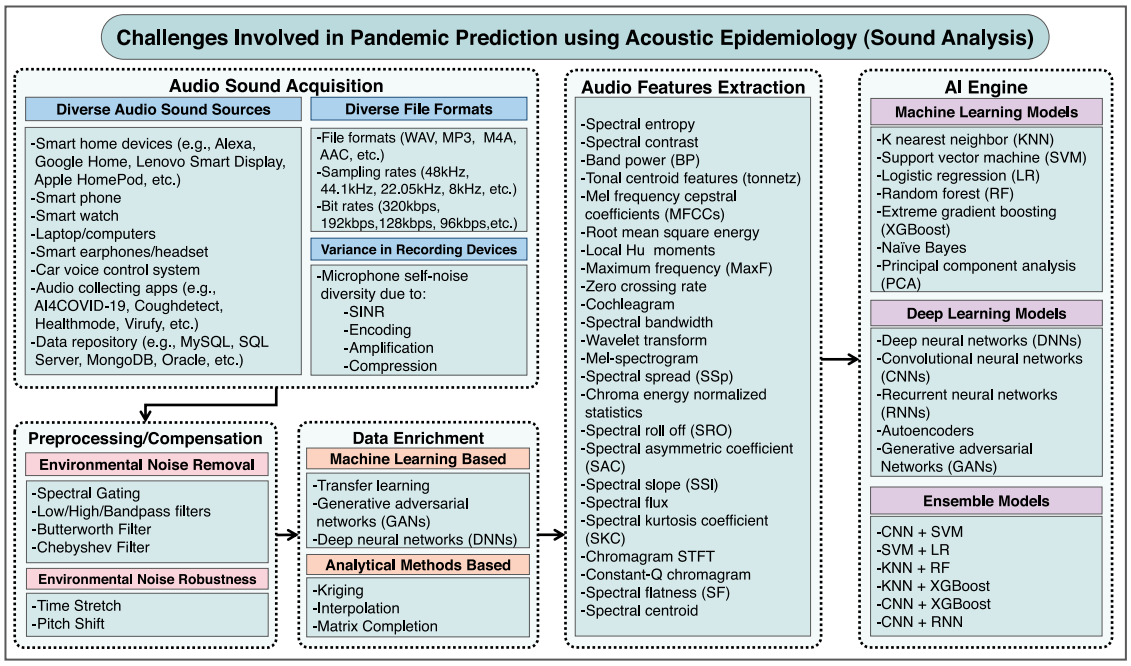


Fig. 5. A representation of engineering challenges associated with the implementation of the pandemic prediction framework.

2.7. Data privacy challenges and proposed mitigation methods

The collection of population-level biomarkers data for iPREDICT introduces inherent challenges related to data privacy. Given the sensitive nature of health-related information, ensuring the confidentiality and privacy of individuals participating in biomarkers data collection is paramount. We discuss some of the key challenges and respective mitigation methods as follows. We anticipate that this paper will lead to further research that will address the privacy concerns in more detail.

1. Informed Consent and Participant Awareness: Ensuring comprehensive informed consent and participant awareness about the potential risks and uses of biomarkers data can be challenging at the population level. To alleviate this challenge we develop clear and easy-to-understand materials to inform participants about the purpose of biomarkers data collection, the potential risks, and the privacy safeguards in place. We developed user-friendly interfaces to enhance participant understanding during the consent process for our previous work for cough data collection [23].
2. Individual Identifiability: The risk of individual identifiability using biomarkers data poses a significant concern. The individual biomarker profiles may carry unique signatures, increasing the risk of re-identification. This challenge can be mitigated by leveraging anonymization techniques, such as differential privacy [87], to protect against the re-identification of individuals. The introduction of controlled noise during the creation of individual biomarker profiles 2.2 ensures individual data points cannot be isolated.
3. Secure Data Transmission and Storage: The transmission and storage of biomarkers data demand robust security measures to prevent unauthorized access and data breaches. To mitigate this challenge we propose two approaches that can be used based on the nature of the biomarker data and the need to store it in a centralized database or not. One approach is the use of encryption protocols for secure data transmission and the implementation of stringent access controls and encryption for data storage. The second and more decentralized approach is using federated learning [88] which will eliminate the need to transfer data, instead distributed individual models can be trained and their respective learning will used for decision making. However, leveraging federated learning will lead to further challenges and researchers are finding ways to overcome them [89].

Addressing data privacy challenges in population-level biomarkers data collection requires a multifaceted approach. Given iPREDICT is a concept framework we anticipate that it will lead to in-depth research on each of its components which will address respective potential privacy challenges and mitigation strategies.

3. Engineering challenges in implementation of iPREDICT

A variety of challenges in the medical and engineering domains will be faced when implementing iPREDICT. The challenges in engineering arise from devices used to record a variety of biomarkers using biosensing technology, see Fig. 1. The recording devices involved vary in terms of their biomarker-capturing mechanisms, hardware components, and software capabilities (operating system, middleware, etc.) that introduce variability and randomness in the data collection process. Moreover, environmental factors such as ambient noise can also impact the performance of iPREDICT. Therefore, in this study, we identified and quantitatively assessed several key engineering challenges encountered in diagnosing COVID-19 based on cough sound biomarker, recorded using a smartphone microphone. The following engineering challenges included questions related to audio signal processing, such as the effects of contamination of cough sound with the environmental noise, variability of self-induced noise (by active circuitry, and Brownian movement of air particles) by a microphone [90], variation in sampling frequencies (in order to capture high-quality cough samples needed for AI-based pandemic prediction), and compression rates for efficient storage and transmission of cough samples.

3.1. Environmental noise and noise variations

Recording cough sounds via smartphones in a public setting induces environmental noise and reverberation which contaminate the

recording and consequently compromise the accuracy of the ML models for disease diagnosis. While reverberation can be characterized by sound propagation models in indoor and outdoor settings, [91] environmental noise can vary greatly based on the surroundings. The noise amplitude and frequency distribution along with the signal-to-noise ratio in each sound recording can be unique. Moreover, even in the same ambient setup, the microphone-to-mouth positioning in terms of distance and angle causes noise variations in recordings. Although noise can be reduced by leveraging filtering and smoothing algorithms, this results in the elimination of high-frequency components in the recording. However, these high-frequency components are not always noise-induced and can be crucial for accurate cough-based diagnosis and hence, cannot be completely removed. Moreover, ML models trained on noiseless data or data with limited noise scenarios tend to overfit and cannot be generalized to real-life settings where infinite types of environmental noises exist. Therefore, the challenge is to build noise-aware ML models that are robust to environmental distortions at training and inference [92].

3.2. Heterogeneity of microphones

Another factor that complicates ambulatory sound-based cough diagnostics is the heterogeneity of recording devices at three levels: (1) varying device types (e.g., cellphones, laptops, lapel microphones, and smartwatches); (2) devices of the same type from different manufacturers (Apple, Google, and Samsung, etc.) with varying specifications frequency response, phase response, sensitivity, noise level, sound pressure level, signal to noise ratio, and; (3) devices with the same specifications (same brand and same model) that exhibit electro-acoustic variations due to inherent manufacturing process uncertainties of microphone chips [90,93]. Our focus is on the differences in recording microphones from the same or different manufacturers. Sound recorded via different microphones is not identical and hence affects the performance of iPREDICT. Therefore, the challenge is to generalize the ML models beyond the electro-acoustic differences in microphones.

3.3. Diversity in audio sampling rate

In conjunction with the hardware dissimilarities of the microphones presented in the previous section, software characteristics such as the sampling rate (at which the audio is recorded) generate sound recordings with variable size and quality. Thus, poses the caveat of the trade-off between ML model accuracy and speed (audio with a higher sampling rate will take more time to process, which will compromise on efficiency of the ML model but yield more accurate results). We highlight the audio sampling rate diversity challenge by presenting the quantified impact of 4 different sampling rates in Section 4. The results present a comparative analysis of different sampling rates on the diagnosis of COVID-19 using cough audio data. The cough sounds can have frequency components up to 20 kHz [94]. Therefore we highlight the impact of 8 kHz, 22.05 kHz, 44.1 kHz, and 48 kHz sampling rates in the case study observing the Shannon–Nyquist sampling theorem (i.e., the sampling frequency must be more than double the highest frequency component) [95].

3.4. Diversity in audio file format

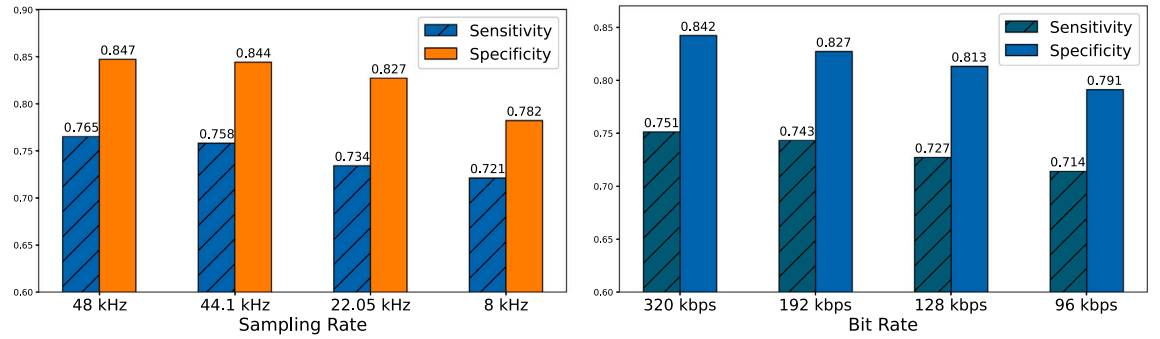
In addition to the sampling rate, the data is lost from the sound recordings through compression as they are stored on the recording devices using different file formats. Although lossless audio formats such as WAV, AIFF, ALAC, and FLAC exist, their larger storage size renders them inefficient for mobile transmission over the network, storage in recording devices and cloud, and consumption by the ML models at scale. Therefore, compression formats such as 3GP, WMA, AAC, M4A, and MP3, with MP3 being the most common [96] reduce the size of the audio file. Furthermore, the reduced file size is efficient to store as well as transmit over the network, to process the audio file on the cloud for biomarker profile signature creation and matching for the potential pandemic prediction. However, file compression does not only lose data but also, does so in favor of keeping human audibility while discarding human indiscernible components. Moreover, to discriminate the nuances of cough originating through closely related respiratory disorders, the frequency components beyond human audible interest can be significant. In addition to the data loss, compression formats also engender the challenge of portability over many codecs used by different recording software. A generalized cough-based diagnosis system must therefore be able to process different codecs or re-encode variable formats into one while being robust to the compression rates and mechanisms of varying formats. This challenge is further highlighted and quantified in Section 4.

4. Results and discussion: A case study to show feasibility of iPREDICT

We leverage our seminal work AI4COVID-19 [23] as a case study to show the feasibility of iPREDICT. By exploiting one biomarker (cough sound) and one biosensing device (smartphone microphone) we analyze and quantify the engineering challenges discussed in Section 3. These challenges come from: (1) audio data acquisition, (2) data transfer over the wireless network, (3) data pre-processing for noise removal and noise robustness, and, (4) the diversity of data acquisition devices. Interested readers can refer to [23] for the details regarding the multi-pronged data-driven AI model, cough sound features, and dataset used in AI4COVID-19.

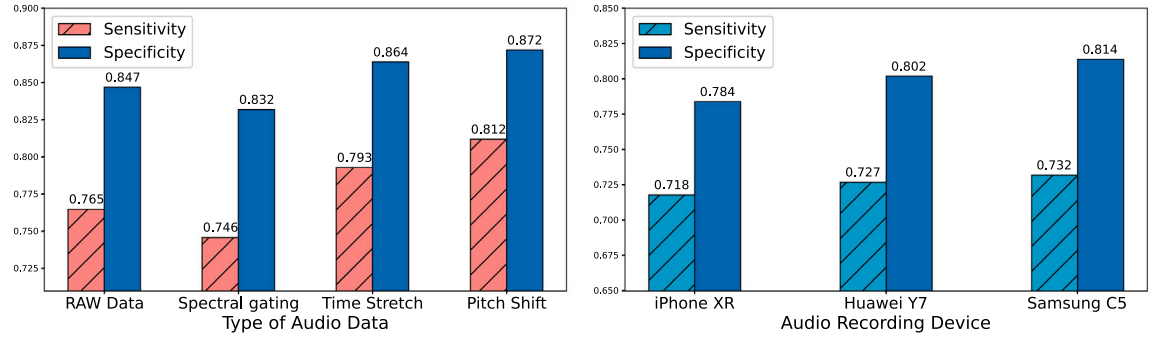
The first challenge is in the audio data acquisition phase due to the variable sampling rates discussed in Section 3 and highlighted in Fig. 5. We present an analysis of AI-based COVID-19 diagnosis using cough data acquired at 4 different sampling rates 48 kHz, 44.1 kHz, 20.05 kHz, and 8 kHz. Fig. 6(a) presents a comparison of COVID-19 diagnosis performance using a variation in true positive and true negative rates (sensitivity and specificity). The results in Fig. 6(a) show that the sensitivity of COVID-19 diagnosis decreases from 0.765 to 0.721 when the sampling rate of cough sounds used for investigation is decreased from 48 kHz to 8 kHz. Moreover, the specificity is also reduced from 0.847 to 0.844, 0.827, and 0.782 for the respective sampling rates. The drop in performance is a function of 3 factors that are involved in the process of up and downsampling of cough data. These factors include interpolation, anti-aliasing, and decimation [97]. This challenge can be further investigated as future work, as an optimization problem between diagnosis performance (diagnosis accuracy of the ML model) and efficiency (time taken by ML model for the inference) of the proposed framework.

Once the audio data is acquired by a biosensor, it needs to be transmitted over a wireless network that has different transmission capabilities. This requires audio data to be compressed, which has several compression formats and bit rates highlighted in Section 3 and Fig. 5. To further highlight this challenge we present MP3 file format compressed at 320 kbps, 192 kbps, 128 kbps, and 96 kbps bit rates, to see the effect of cough data compression on the COVID-19 diagnosis. Fig. 6(b) shows a performance deterioration in sensitivity from 0.751 to 0.714 when the compression bitrate of MP3 is changed from 320 kbps to 96 kbps. Also, the specificity is reduced from 0.842 to 0.79 for the respective bitrates. The drop in performance is a function of quantization error, as well as data encoding [98]. Both, the quantization error and data encoding compromise the quality of audio, which leads to losing some latent features such as MFCCs, band power, and energy, (a comprehensive list is given in Fig. 5) that are important for COVID-19 diagnosis. The detrimental impact of file compression can be further investigated using more sophisticated ensemble methods (e.g. CNN+XGBoost, CNN+LSTM, and CNN+SVM) that are robust to noise caused by data compression [99].



(a) Impact (in terms of sensitivity and specificity) of different sampling rates (used to record audio data) on COVID-19 diagnosis

(b) Impact of different compression rates (data stored in different file formats for efficient transfer over cellular network) on COVID-19 diagnosis



(c) The Impact (in terms of sensitivity and specificity) of different pre-processing methods (used for sound-based data robustness and environmental noise reduction) on COVID-19 diagnosis

(d) The Impact (in terms of sensitivity and specificity) of device diversity (model trained on data acquired using one device and tested on data acquired by a different device) on COVID-19 diagnosis

Fig. 6. Performance analysis of AI4COVID-19 based on various software and hardware related data diversities, different (a) sampling rates (b) bit rates (compression rates) (c) pre-processing methods (d) device diversity, used to acquire training data.

The third challenge arises before transferring the audio data to the ML model, due to the use of pre-processing methods to mitigate the impact of environmental noise in sound recording. There can be two approaches to reduce the effect of environmental noise on the efficiency of the ML method: (1) remove the environmental noise, (2) add more noise to the training data to make the ML model robust to learning from the noisy data. We present the results of spectral gating (SG) as a noise removal method and time stretch (TS) and pitch shift (PS) as noise robustness methods for this case study in Fig. 6(c), while more methods can be found in Fig. 5. Fig. 6(c) shows that TS and PS improve the sensitivity from 0.765 to 0.793 and 0.812 respectively, and also achieve an enhanced specificity for both TS and PS. In contrast, the noise reduction technique SG brings the sensitivity down to 0.746 from 0.765, the specificity is also decreased slightly. A major reason for the drop in performance can be the nature of cough data which is more like noise itself, and when a static noise removal technique like SG is applied it removes some of the latent frequency features that contribute towards the COVID-19 diagnosis, which leads to slightly poor performance. The slight change in the performance of the ML models can be attributed to the lack of variation in the environmental noise due to the controlled nature of the environment setting (hospital setting) used for the data collection for this feasibility of the case. With more diverse data gathering settings, it is expected to have more noise variations, and hence it remains crucial to further investigate the impact of noise on the performance of the ML models. The fourth challenge results from the variance in audio data recording devices. The devices can have different hardware (microphone, speakers, etc.) and software (operating system, middle-ware, etc.) based on brand, make, and model. In this case study, we focus on diversity based on the microphone because that is used to record the audio data. We present an analysis of a diverse set of devices consisting of an iPhone XR, Huawei Y7, and Samsung C5. We trained the AI4COVID-19 framework on data acquired using an Android device and tested on data from iPhone XR, Huawei Y7, and Samsung C5. The results in Fig. 6(d), show a decrease in sensitivity from 0.765 to 0.718 and specificity from 0.847 to 0.784, for the cough data that has an added noise signature of iPhone XR. In contrast, for the Android devices (Huawei Y7 and Samsung C5) the dip in performance is relatively lesser. The potential reasons can be (1) the software of Android devices differs from an iPhone device, and (2) the microphone chips differ for different mobile phone brands. These diversities contribute to the self-noise profiles of each device which leads to variation in the diagnosis performance.

5. Conclusion

The idea we presented in this paper is a concept framework based on the emergence of multi-disciplinary research enabled by advances in artificial intelligence, omnipresent wireless networks, and a surge in the use of smart biosensing devices for health and fitness purposes, can make the pandemic prediction goal attainable. iPREDICT—a holistic concept framework designed to forecast an epidemic based on crowd-sourced biomarkers acquired using biosensing wearable devices. iPREDICT performs real-time anomaly detection on biosensing profiles of humans in a spatio-temporal domain to alert the relevant authorities to take respective actions at the pre-emergence stage. The alert-worthy number of cases in any population is quantified through a threshold which is learned by graphically modeling the data of historical epidemics and training a GNN-based threshold predictor. The GNNs trained over graphs of various populations predict the epidemic alarm threshold at variable scales and resolution of any region. Timely actions based on the prediction output can prevent an epidemic from becoming a pandemic. We present our previous work AI4COVID-19 as a tool to show the feasibility of iPREDICT and the underlying engineering challenges to predict a COVID-19 like pandemic in the future. The

case study results provide an analysis of several software (sampling rates to record audio and compression rates to transfer audio efficiently over the network) and hardware (audio recording device diversity) related challenges, associated with pandemic prediction based on sound analysis. We present an extensive framework for real-time pandemic prediction based on several state-of-the-art emerging technologies and uncover several research questions with the hope of extricating humanity from another devastating pandemic. Based on our research we identified some research directions that still need to be explored such as data privacy challenges for a framework that works at the population level. Data quality and standardization (we showed the results deterioration due to low-quality audio in our case study), and engagement and participation.

CRediT authorship contribution statement

Muhammad Sajid Riaz: Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maria Shaukat:** Writing – review & editing, Writing – original draft, Visualization. **Tabish Saeed:** Writing – review & editing, Data curation. **Aneeqa Ijaz:** Writing – review & editing, Visualization, Data curation. **Haneya Naeem Qureshi:** Writing – review & editing, Supervision. **Iryna Posokhova:** Validation. **Ismail Sadiq:** Writing – review & editing. **Ali Rizwan:** Conceptualization. **Ali Imran:** Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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

References

[1] Žižek S. Pandemic!: COVID-19 shakes the world. John Wiley & Sons; 2020.
[2] Barro RJ, Ursúa JF, Weng J. The coronavirus and the great influenza pandemic: Lessons from the "Spanish flu" for the coronavirus's potential effects on mortality and economic activity. Tech. rep., National Bureau of Economic Research; 2020.
[3] UNCTAD. Global economy could lose over $4 trillion due to COVID-19 impact on tourism | UNCTAD. 2021, [Online] Available: https://unctad.org/news/global-economy-could-lose-over-4-trillion-due-covid-19-impact-tourism. [Accessed 10 August 2021].
[4] Wu W-K, Liou J-M, Hsu C-C, Lin Y-H, Wu M-S. Pandemic preparedness in Taiwan. Nature Biotechnol 2020;38(8):932–3.
[5] Brüssow H, Brüssow L. Clinical evidence that the pandemic from 1889 to 1891 commonly called the Russian flu might have been an earlier coronavirus pandemic. Microb Biotechnol 2021;14(5):1860–70.
[6] Breitnauer J. The Spanish Flu epidemic and its influence on history. Pen and Sword; 2020.
[7] Jackson C. History lessons: the Asian flu pandemic. Br J Gen Pract 2009;59(565):622–3.
[8] HIVgov. Symptoms of HIV. 2022, [Online]. Available: https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/symptoms-of-hiv. [Accessed 30 June 2022].
[9] MAYO Clinic. H1N1 flu (swine flu). 2022, [Online]. Available:https://www.mayoclinic.org/diseases-conditions/swine-flu/symptoms-causes/syc-20378103. [Accessed 30 June 2022].
[10] American Lung Association. Middle Eastern Respiratory Syndrome (MERS). 2022, [Online]. Available:https://www.lung.org/lung-health-diseases/lung-disease-lookup/middle-eastern-respiratory-syndrome-mers. [Accessed 30 June 2022].
[11] CDC. Middle east respiratory syndrome (MERS). 2022, [Online]. Available:https://www.cdc.gov/coronavirus/mers/about/symptoms.html. [Accessed 30 June 2022].
[12] CDC. Ebola (Ebola Virus Disease). 2022, [Online]. Available:https://www.cdc.gov/vhf/ebola/symptoms/index.html. [Accessed 30 June 2022].
[13] Stokes EK, Zambrano LD, Anderson KN, Marder EP, Raz KM, Felix SEB, et al. Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020. Morb Mortal Wkly Rep 2020;69(24):759.
[14] Van den Driessche P. Reproduction numbers of infectious disease models. Infect Dis Model 2017;2(3):288–303.
[15] M. H. A. Biswas Md. A SEIR model for control of infectious diseases with constraints. Math Biosci Eng 2014;11(4):761–84.
[16] Berestycki H, Roquejoffre J-M, Rossi L. Propagation of epidemics along lines with fast diffusion. Bull Math Biol 2021;83(1):1–34.
[17] Delamater P, Street E, Leslie T, Yang YT, Jacobsen K. Complexity of the basic reproduction number (R0). Emerg Infect Dis J 2019;25(1):1, [Online] Available: https://wwwnc.cdc.gov/eid/article/25/1/17-1901_article.
[18] Coccia M. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Sci Total Environ 2020;729:138474.
[19] Wahid MA, Bukhari SHR, Daud A, Awan SE, Raja MAZ. COVICT: an IoT based architecture for COVID-19 detection and contact tracing. J Ambient Intell Humaniz Comput 2023;14(6):7381–98.
[20] Coccia M. COVID-19 pandemic over 2020 (with lockdowns) and 2021 (with vaccinations): similar effects for seasonality and environmental factors. Environ Res 2022;208:112711.
[21] Bontempi E, Coccia M, Vergalli S, Zanoletti A. Can commercial trade represent the main indicator of the COVID-19 diffusion due to human-to-human interactions? A comparative analysis between Italy, France, and Spain. Environ Res 2021;201:111529.
[22] Benati I, Coccia M. Effective contact tracing system minimizes COVID-19 related infections and deaths: policy lessons to reduce the impact of future pandemic diseases. J Pub Adm Gov 2022;12(3).
[23] Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, John CN, Hussain MI, Nabeel M. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform Med Unlocked 2020;20:100378.
[24] Kummitha RKR. Smart technologies for fighting pandemics: The techno-and human-driven approaches in controlling the virus transmission. Gov Inf Q 2020;37(3):101481.
[25] Mujawar MA, Gohel H, Bhardwaj SK, Srinivasan S, Hickman N, Kaushik A. Nano-enabled biosensing systems for intelligent healthcare: towards COVID-19 management. Mater Today Chem 2020;17:100306.
[26] Erdem O, Es I, Saylan Y, Inci F. Unifying the efforts of medicine, chemistry, and engineering in biosensing technologies to tackle the challenges of the COVID-19 pandemic. Anal Chem 2021;94(1):3–25.
[27] Benati I, Coccia M. Global analysis of timely COVID-19 vaccinations: improving governance to reinforce response policies for pandemic crises. Int J Health Govern 2022;27(3):240–53.
[28] Coccia M. Improving preparedness for next pandemics: Max level of COVID-19 vaccinations without social impositions to design effective health policy and avoid flawed democracies. Environ Res 2022;213:113566.
[29] Coccia M. Optimal levels of vaccination to reduce COVID-19 infected individuals and deaths: A global analysis. Environ Res 2022;204:112314.
[30] Abdel-Ghani A, Abughazzah Z, Akhund M, Abualsaud K, Yaacoub E. Efficient pandemic infection detection using wearable sensors and machine learning. In: 2023 international wireless communications and mobile computing. IEEE; 2023, p. 1562–7.
[31] Coccia M. Effects of strict containment policies on COVID-19 pandemic crisis: lessons to cope with next pandemic impacts. Environ Sci Pollut Res 2023;30(1):2020–8.
[32] Coccia M. Preparedness of countries to face covid-19 pandemic crisis: Strategic positioning and underlying structural factors to support strategies of prevention of pandemic threats. Environ Res 2022;203(111678):10–1016.
[33] Marakhimov A, Joo J. Consumer adaptation and infusion of wearable devices for healthcare. Comput Hum Behav 2017;76:135–48.
[34] Tran V-T, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. NPJ Dig Med 2019;2(1):1–8.
[35] Lee SM, Lee D. Healthcare wearable devices: an analysis of key factors for continuous use intention. Serv Bus 2020;14(4):503–31.
[36] Chawla MN. AI, IoT and wearable technology for smart healthcare–A review. Int J Green Energy 2020;7(1):9–13.
[37] Iosa M, Picerno P, Paolucci S, Morone G. Wearable inertial sensors for human movement analysis. Expert Rev Med Dev 2016;13(7):641–59.
[38] Pacheco AG, Cabello FA, Fonoff AM, Rodrigues PG, Penatti OA, Pinto PR. Towards low-power heart rate estimation based on user's demographics and activity level for wearables. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing. IEEE; 2023, p. 1–5.
[39] Han DC, Shin HJ, Yeom SH, Lee W. Wearable human health-monitoring band using inkjet-printed flexible temperature sensor. J Sens Sci Technol 2017;26(5):301–5.
[40] Alsabek MB, Shahin I, Hassan A. Studying the Similarity of COVID-19 Sounds based on Correlation Analysis of MFCC. In: 2020 international conference on communications, computing, cybersecurity, and informatics. IEEE; 2020, p. 1–5.
[41] Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2022;1–28.

[42] Parsons TJ, Sartini C, Welsh P, Sattar N, Ash S, et al. Objectively measured physical activity and cardiac biomarkers: A cross sectional population based study in older men. Int J Cardiol 2018;254:322–7.

[43] Thomas D, Bouchard C, Church T, Slentz C, Kraus W, Redman L, et al. Why do individuals not lose more weight from an exercise intervention at a defined dose? An energy balance analysis. Obes Rev 2012;13(10):835–47.

[44] Abu Zaid M, Wu J, Wu C, Logan BR, Yu J, Cutler C, et al. Plasma biomarkers of risk for death in a multicenter phase 3 trial with uniform transplant characteristics post–allogeneic HCT. Blood J Am Soc Hematol 2017;129(2):162–70.

[45] Brasier N, Eckstein J. Sweat as a source of next-generation digital biomarkers. Dig Biomark 2019;3(3):155–65.

[46] Elhakeem A, Cooper R, Whincup P, Brage S, Kuh D, Hardy R. Physical activity, sedentary time, and cardiovascular disease biomarkers at age 60 to 64 years. J Am Heart Assoc 2018;7(16):e007459.

[47] Simonsick EM, Meier HC, Shaffer NC, Studenski SA, Ferrucci L. Basal body temperature as a biomarker of healthy aging. Age 2016;38:445–54.

[48] Levy J, Álvarez D, Rosenberg AA, Alexandrovich A, Del Campo F, Behar JA. Digital oximetry biomarkers for assessing respiratory function: standards of measurement, physiological interpretation, and clinical use. NPJ Dig Med 2021;4(1):1.

[49] Baliga S, Muglikar S, Kale R. Salivary pH: A diagnostic biomarker. J Indian Soc Periodontol 2013;17(4):461.

[50] Gazi AH, Gurel NZ, Richardson KL, Wittbrodt MT, Shah AJ, Vaccarino V, et al. Digital cardiovascular biomarker responses to transcutaneous cervical vagus nerve stimulation: state-space modeling, prediction, and simulation. JMIR mHealth uHealth 2020;8(9):e20488.

[51] Birch-Machin M, Russell E, Latimer J. Mitochondrial DNA damage as a biomarker for ultraviolet radiation exposure and oxidative stress. Br J Dermatol 2013;169(s2):9–14.

[52] Braithwaite JJ, Watson DG, Jones R, Rowe M. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology 2013;49(1):1017–34.

[53] Hagan S, Martin E, Enríquez-de Salamanca A. Tear fluid biomarkers in ocular and systemic disease: potential use for predictive, preventive and personalised medicine. Epma J 2016;7:1–20.

[54] Kozitsin V, Katser I, Lakontsev D. Online forecasting and anomaly detection based on the ARIMA model. Appl Sci 2021;11(7):3194.

[55] Canizo M, Triguero I, Conde A, Onieva E. Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. Neurocomputing 2019;363:246–60.

[56] Tziolas T, Papageorgiou K, Theodosiou T, Papageorgiou E, Mastos T, Papadopoulos A. Autoencoders for anomaly detection in an industrial multivariate time series dataset. Eng Proc 2022;18(1):23.

[57] Menda K, Laird L, Kochenderfer MJ, Caceres RS. Explaining COVID-19 outbreaks with reactive SEIRD models. Sci Rep 2021;11(1):1–12.

[58] Ramadan RH, Ramadan MS. Prediction of highly vulnerable areas to COVID-19 outbreaks using spatial model: Case study of Cairo Governorate, Egypt. Egypt J Remote Sens Space Sci 2022;25(1):233–47.

[59] Qureshi IH, Awais M, Awan SE, Abrar MN, Raja MAZ, Alharbi SO, et al. Influence of radially magnetic field properties in a peristaltic flow with internal heat generation: Numerical treatment. Case Stud Therm Eng 2021;26:101019.

[60] Awan SE, Raja MAZ, Gul F, Khan ZA, Mehmood A, Shoaib M. Numerical computing paradigm for investigation of micropolar nanofluid flow between parallel plates system with impact of electrical MHD and hall current. Arab J Sci Eng 2021;46:645–62.

[61] Hossain AD, Jarolimova J, Elnaiem A, Huang CX, Richterman A, Ivers LC. Effectiveness of contact tracing in the control of infectious diseases: a systematic review. Lancet Public Health 2022.

[62] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw 2009;20:61–80.

[63] Espinosa P, Quirola-Amores P, Teran E. Application of a susceptible, infectious, and/or recovered (SIR) model to the COVID-19 pandemic in Ecuador. Front Appl Math Statist 2020;6:55.

[64] Korolev I. Identification and estimation of the SEIRD epidemic model for COVID-19. J Econometrics 2021;220:63.

[65] Schlickeiser R, Kröger M. Analytical modeling of the temporal evolution of epidemics outbreaks accounting for vaccinations. Physics 2021;3(2):386–426.

[66] Yadav SK, Akhter Y. Statistical modeling for the prediction of infectious disease dissemination with special reference to COVID-19 spread. Front Public Health 2021;9:680.

[67] Marzouk M, Elshaboury N, Abdel-Latif A, Azab S. Deep learning model for forecasting COVID-19 outbreak in Egypt. Process Saf Environ Prot 2021;153:363.

[68] Kafieh R, Arian R, Saeedizadeh N, Amini Z, Serej ND, Minaee S, et al. COVID-19 in Iran: Forecasting pandemic using deep learning. Comput Math Methods Med 2021;2021.

[69] Wang L, Adiga A, Venkatramanan S, Chen J, Lewis B, Marathe M. Examining deep learning models with multiple data sources for COVID-19 forecasting. In: Proceedings - 2020 IEEE international conference on big data. Institute of Electrical and Electronics Engineers Inc.; 2020, p. 3846–55.

[70] Panja M, Chakraborty T, Nadim SS, Ghosh I, Kumar U, Liu N. An ensemble neural network approach to forecast Dengue outbreak based on climatic condition. Chaos Solitons Fractals 2023;167:113124.

[71] Liu W, Dai Q, Bao J, Shen W, Wu Y, Shi Y, et al. Influenza activity prediction using meteorological factors in a warm temperate to subtropical transitional zone, Eastern China. Epidemiol Infect 2019;147.

[72] Engebretsen S, Engø-Monsen K, Aleem MA, Gurley ES, Frigessi A, Blasio BFD. Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of Bangladesh. J R Soc Interface 2020;17.

[73] Wang L, Adiga A, Chen J, Sadilek A, Venkatramanan S, Marathe M. CausalGNN: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol. 36. 2022;12191–9.

[74] Fritz C, Dorigatti E, Rügamer D. Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany. Sci Rep 2022;12(1):1–18.

[75] Panagopoulos G, Nikolentzos G, Vazirgiannis M. Transfer graph neural networks for pandemic forecasting. In: Proceedings of the AAAI conference on artificial intelligence, vol. 35. Association for the Advancement of Artificial Intelligence; 2021, p. 4838–45.

[76] Yu S, Xia F, Li S, Hou M, Sheng QZ. Spatio-temporal graph learning for epidemic prediction. ACM Trans Intell Syst Technol 2023.

[77] Hy TS, Nguyen VB, Tran-Thanh L, Kondor R. Temporal multiresolution graph neural networks for epidemic prediction. PMLR; 2022, p. 21–32.

[78] Ma Y, Gerard P, Tian Y, Guo Z, Chawla NV. Hierarchical spatio-temporal graph neural networks for pandemic forecasting. In: International conference on information and knowledge management, proceedings. Association for Computing Machinery; 2022, p. 1481–90.

[79] CDC. Describing epidemiologic data | epidemic intelligence service | CDC. 2022, [Online]. Available:https://www.cdc.gov/eis/field-epi-manual/chapters/Describing-Epi-Data.html. [Accessed 30 June 2022].

[80] Waters N. Tobler's first law of geography. In: International encyclopedia of geography: People, the Earth, environment and technology. John Wiley & Sons, Ltd; 2017, p. 1–13.

[81] UnitedStatesNow. How many zip codes are in the United States? 2022, [Online]. Available:https://www.unitedstatesnow.org/how-many-zip-codes-are-in-the-united-states.htm. [Accessed 30 June 2022].

[82] US Census Bureau. Understanding geographic identifiers (GEOIDs). 2022, [Online]. Available:https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html. [Accessed 30 June 2022].

[83] Meta. Data for good tools and data. 2022, [Online]. Available: https://dataforgood.facebook.com/dfg/tools. [Accessed 30 June 2022].

[84] Li Y, Yu R, Shahabi C, Liu Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In: 6th international conference on learning representations, ICLR 2018 - conference track proceedings. International Conference on Learning Representations, ICLR; 2017.

[85] Pham HV, Thanh DH, Moore P. Hierarchical pooling in graph neural networks to enhance classification performance in large datasets. Sensors (Basel, Switzerland) 2021;21.

[86] Lee J, Lee I, Kang J. Self-attention graph pooling. In: 36th international conference on machine learning, vol. 2019-June. International Machine Learning Society (IMLS); 2019, p. 6661–70.

[87] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, editors. Theory of cryptography. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006, p. 265–84.

[88] McMahan B, Moore E, Ramage D, Hampson S, Arcas BAy. Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu J, editors. Proceedings of the 20th international conference on artificial intelligence and statistics. Proceedings of machine learning research, vol. 54, PMLR; 2017, p. 1273–82.

[89] Zhang T, Gao L, He C, Zhang M, Krishnamachari B, Avestimehr AS. Federated learning for the internet of things: Applications, challenges, and opportunities. IEEE Internet Things Mag 2022;5(1):24–9.

[90] Das A, Borisov N, Caesar M. Do you hear what i hear? fingerprinting smart devices through embedded acoustic components. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. 2014, p. 441–52.

[91] Wijayasingha L, Stankovic JA. Robustness to noise for speech emotion classification using CNNs and attention mechanisms. Smart Health 2021;19:100165.

[92] Kim S, Raj B, Lane I. Environmental noise embeddings for robust speech recognition. 2016, arXiv preprint arXiv:1601.02553.

[93] Mathur A, Zhang T, Bhattacharya S, Velickovic P, Joffe L, Lane ND, et al. Using deep data augmentation training to address software and hardware heterogeneities in wearable and smartphone sensing devices. In: 2018 17th ACM/IEEE international conference on information processing in sensor networks. IEEE; 2018, p. 200–11.

[94] Hall JI, Lozano M, Estrada-Petrocelli L, Birring S, Turner R. The present and future of cough counting tools. J Thorac Dis 2020;12(9):5207.

[95] Por E, Kooten Mv, Sarkovic V. Nyquist–Shannon sampling theorem. Leiden Univ 2019;1(1).
[96] Smirnova T. Comparative analysis of modern formats of lossy audio compression. 2020.
[97] Harris FJ. Multirate signal processing for communication systems. CRC Press; 2022.
[98] Sterne J. MP3: The meaning of a format. Duke University Press; 2012.
[99] Riaz MS, Qureshi HN, Masood U, Rizwan A, Abu-Dayya A, Imran A. A hybrid deep learning-based (HYDRA) framework for multifault diagnosis using sparse MDT reports. IEEE Access 2022;10:67140–51.