big data and cognitive computing

MDPI

Review

Privacy-Enhancing Digital Contact Tracing with Machine Learning for Pandemic Response: A Comprehensive Review

Ching-Nam Hang [1], Yi-Zhen Tsai [2], Pei-Duo Yu [3], Jiasi Chen [2] and Chee-Wei Tan [4,*]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong; cnhang3-c@my.cityu.edu.hk
[2] Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA; ytsai036@ucr.edu (Y.-Z.T.); jiasi@cs.ucr.edu (J.C.)
[3] Department of Applied Mathematics, Chung Yuan Christian University, Taoyuan City 320314, Taiwan; peiduoyu@cycu.edu.tw
[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
* Correspondence: cheewei.tan@ntu.edu.sg

Abstract: The rapid global spread of the coronavirus disease (COVID-19) has severely impacted daily life worldwide. As potential solutions, various digital contact tracing (DCT) strategies have emerged to mitigate the virus's spread while maintaining economic and social activities. The computational epidemiology problems of DCT often involve parameter optimization through learning processes, making it crucial to understand how to apply machine learning techniques for effective DCT optimization. While numerous research studies on DCT have emerged recently, most existing reviews primarily focus on DCT application design and implementation. This paper offers a comprehensive overview of privacy-preserving machine learning-based DCT in preparation for future pandemics. We propose a new taxonomy to classify existing DCT strategies into forward, backward, and proactive contact tracing. We then categorize several DCT apps developed during the COVID-19 pandemic based on their tracing strategies. Furthermore, we derive three research questions related to computational epidemiology for DCT and provide a detailed description of machine learning techniques to address these problems. We discuss the challenges of learning-based DCT and suggest potential solutions. Additionally, we include a case study demonstrating the review's insights into the pandemic response. Finally, we summarize the study's limitations and highlight promising future research directions in DCT.

Keywords: digital contact tracing; COVID-19; computational epidemiology; machine learning

Citation: Hang, C.-N.; Tsai, Y.-Z.; Yu, P.-D.; Chen, J.; Tan, C.-W. Privacy-Enhancing Digital Contact Tracing with Machine Learning for Pandemic Response: A Comprehensive Review. *Big Data Cogn. Comput.* **2023**, *7*, 108. https://doi.org/10.3390/bdcc7020108

Academic Editor: Carson K. Leung

Received: 25 April 2023
Revised: 10 May 2023
Accepted: 16 May 2023
Published: 1 June 2023

1. Introduction

The highly-transmissible coronavirus disease 2019 (COVID-19) is regarded as the biggest global crisis since World War II [1] and has caused more than 6.9 million deaths among 765 million confirmed infection cases (up to 20 April 2023) [2]. During the initial phase of the pandemic, when pharmaceutical interventions like vaccines were unavailable, non-pharmaceutical interventions such as social distancing and compulsory quarantine have been used to curb the spread of the pandemic in the short-term [3,4]. However, these restrictive measures can severely affect mental health, social life, and the economy [5–9]. As the COVID-19 pandemic transits to an endemic phase, the catastrophic impact of the COVID-19 pandemic is a sobering reminder that implementing effective non-pharmaceutical intervention strategies is a critical first step to the next pandemic.

Contact tracing, a public health strategy to trace the spread of the disease among individuals in a community, is a form of non-pharmaceutical intervention strategy to keep the number of confirmed cases low while allowing more social and economic activities to resume [10–12]. Manual contact tracing requires health workers to manually warn the direct contacts of an infected person about the risk of acquiring the virus via telephone

or interview [13–16]. Digital contact tracing (DCT) has emerged as a way to digitize and accelerate this process using mobile applications (apps), tackling the inefficiency and limited human resource issues in traditional manual contact tracing [17–22]. Although there have been rapid research contributions in the field of DCT towards the fight against pandemics, existing research or related reviews have mainly focused on DCT app design and implementation such as sensing methods (e.g., Bluetooth and GPS) [23–27], privacy issues [28–30], and infrastructure (e.g., centralized or decentralized) [31–33]. Indeed, in the early stages of the COVID-19 pandemic, most countries embraced DCT apps as a form of social intervention, but the uptake of DCT apps faced limitations due in part to privacy concerns and a lack of trust among the populace. DCT, however, remains essential to the early detection of a disease outbreak and to offer rapid response to raise awareness, reducing the induced costs while curbing the pandemic spread. This will necessitate DCT strategies to reach a critical mass of users to be effective as well as mitigating uncertainties due to false positives and negatives generated by DCT apps, which could lead to people being unnecessarily quarantined. There is a need for research on how emerging artificial intelligence (AI) techniques can be applied to various DCT strategies.

Different DCT approaches lead to various computational epidemiology problems, which often involve networks arising from social relationships. For instance, a contact tracing network can be constructed by knowing who is in close contact with whom in DCT by modeling individuals and social contacts between individuals as vertices and edges, respectively. Given the contact graph, the problem of finding the source case can be modeled as a maximum-likelihood estimation problem whose optimal solution depends on the graph type [34,35]. However, noisy or missing information in DCT can significantly affect the underlying network topology, degrading the overall estimation performance [36]. Another computational epidemiology problem of DCT is quantifying and predicting the risk of infection in a pandemic. For example, infectiousness can be estimated using deep learning with individual-level features (e.g., age and symptoms) [37,38].

In general, the computational epidemiology problems of DCT are usually formulated as optimization problems, where the parameter optimization process is called learning. Thus, it is important to understand how to leverage machine learning techniques to optimize DCT. More importantly, a comprehensive overview of machine learning-enabled DCT can serve to prepare for the next pandemic. To summarize, this paper posits the following contributions:

- We provide a general overview and propose a new taxonomy of DCT strategies. DCT strategies are categorized into three groups: Forward contact tracing, backward contact tracing, and proactive contact tracing.
- We overview several DCT apps specifically targeted toward mitigating the impacts of the COVID-19 pandemic and categorize them based on their tracing methods.
- We formulate three computational epidemiology subproblems related to DCT and present a comprehensive review of machine learning techniques to address the subproblems. We provide a detailed theoretical and empirical analysis for each learning method and summarise the corresponding contact tracing datasets.
- We highlight specific challenges of current machine learning techniques for DCT and provide potential solutions for how to overcome these challenges.

This research paper aims to provide a comprehensive review of machine learning-enabled DCT by examining how various DCT strategies can benefit from automation and computation. We decompose the specific DCT data-driven process into subproblems that can be addressed with a variety of machine learning techniques. It is important to note that the focus of this review is confined to machine learning-based DCT strategies. Table 1 summarizes the machine learning techniques used for DCT in this paper. In Section 2, we provide a comprehensive overview of related literature, introduce DCT strategies, review DCT apps used in response to COVID-19, and discuss related work focused on addressing DCT strategies. Section 3 outlines the methods used in conducting the review for this study. Privacy-preserving machine learning techniques for DCT strategies are discussed

in Section 4. Section 5 presents the results and analysis of our proposed methodology. Section 6 explores the data challenges of machine learning-based DCT. A case study for this review is provided in Section 7. Section 8 highlights the study's limitations and suggests future research directions. Finally, we conclude the paper in Section 9.

Table 1. A summary of machine learning techniques used for digital contact tracing in the paper.

Tracing Strategy	Computational Epidemiology Subproblem/Challenge	Machine Learning Technique	Model Type	Task Type	Reference
Forward	Contact Graph Construction	Boosted Decision Trees and Convolutional Neural Networks	Discriminative	Upstream	Section 4.2
Backward	Infection Source Estimation (Source Attribution)	Graph Neural Network	Discriminative	Downstream	Section 4.3.2
Proactive	Risk of Infectious Exposure Prediction	Set Transformer	Discriminative	Downstream	Section 4.4.1
		Deep Graph Infomax	Generative	Upstream	Section 4.4.2
		Graph Transformer	Generative	Upstream	Section 4.4.3
Forward, Backward, and Proactive	Privacy and Security	Federated Graph Learning with Differential Privacy	Discriminative	Downstream	Section 6.1
Forward, Backward, and Proactive	Data Availability	Generative Adversarial Network	Generative	Upstream	Section 6.2

2. Background and Related Work

In this section, we begin by examining existing survey literature on DCT. Following that, we introduce the three primary DCT strategies and analyze several popular contact tracing apps developed during the COVID-19 pandemic. Finally, we offer a comprehensive review of the relevant studies associated with these DCT approaches.

2.1. Related Reviews

The novelty of our review paper lies in its specific focus on the role of machine learning in optimizing DCT strategies for pandemic response, setting it apart from previous reviews. For instance, the study in [39] offers a scoping review of contact tracing strategies for COVID-19 prevention and containment, but it does not concentrate on machine learning applications. Likewise, the research in [40] systematically reviews the use of AI, including machine learning and deep learning techniques, in combating COVID-19's effects, but it does not explicitly address DCT. In [41], the authors examine how big data, AI, and nature-inspired computing models can be employed in detecting COVID-19 cases and provide an overview of their use in contact tracing. However, this work lacks the depth and specificity of our paper regarding machine learning-enabled DCT. The review in [42] delves into the broader topic of integrating emerging technologies into COVID-19 contact tracing, covering opportunities, challenges, and pitfalls. While it contributes valuable insights into contact tracing and technology, it primarily provides a general overview without focusing on the particular machine learning techniques for optimizing DCT. Similarly, the work in [43] reviews AI and machine learning applications in combating COVID-19 but lacks an in-depth analysis of machine learning models for contact tracing. Other reviews focus solely on the designs and technologies of contact tracing mobile apps, specifically for COVID-19 [44–46]. Our review paper not only stands out due to its comprehensive analysis of machine learning approaches investigated in the study but also presents a problem formulation and mathematical analysis. In contrast, previous reviews tend to

provide only a general understanding of how relevant approaches function. Unlike most prior reviews that leave identified DCT challenges as open questions for future research, our study presents potential machine learning-based solutions, aiming to inspire readers to explore further options by building upon our approach. Additionally, the majority of earlier reviews do not offer or summarize relevant contact tracing datasets. While the work in [40] does provide COVID-19-related datasets, they are not explicitly associated with DCT. As a result, our review paper delivers a more focused and detailed examination of machine learning's role in DCT for pandemic response, which is not provided in previous reviews. This makes our paper a better resource for researchers and practitioners interested in optimizing DCT strategies using machine learning techniques. Table 2 summarizes highly cited and relevant review papers examined in our study.

Table 2. A summary of related reviews discussed in this paper. The research domain specifies whether a study focuses on DCT and COVID-19. The technological aspect denotes the study's emphasis on big data (data-driven learning), AI (including machine learning, deep learning, and automated processes), and mobile apps (reviews of DCT apps). The dataset indicates whether a study provides relevant contact tracing data for further analysis.

Reference	Year	Research Domain		Technological Aspect			Dataset
		DCT	COVID-19	Big Data	AI	Mobile App	
Lalmuanawma et al. [43]	2020	✓	✓	✗	✓	✓	✗
Agbehadji et al. [41]	2020	✓	✓	✓	✓	✗	✗
Mbunge [42]	2020	✓	✓	✓	✓	✓	✗
Altmann et al. [44]	2020	✓	✓	✗	✗	✓	✗
Ahmed et al. [45]	2020	✓	✓	✗	✗	✓	✗
Mondal et al. [40]	2021	✗	✓	✓	✓	✗	✗
Alanzi [46]	2021	✓	✓	✗	✗	✓	✗
Ojokoh et al. [39]	2022	✓	✓	✗	✓	✓	✗
This Study	2023	✓	✓	✓	✓	✓	✓

2.2. Strategies of Digital Contact Tracing

In this subsection, we outline the three primary DCT strategies employed in pandemic response and use an illustrative example to demonstrate their distinct approaches, facilitating a better understanding of their differences.

Forward Contact Tracing (FCT) finds all the contacts who could have been infected after the index case is identified. It sends quarantine recommendations to all recent direct contacts of an infected individual only after a positive virus test result is obtained. This method is also called binary contact tracing as it uses binary information (positive or negative virus test result) to send binary recommendations (at-risk or no-risk). A contact tracing network is usually constructed from a given index case.

Backward Contact Tracing (BCT) traces the source of infection and identifies sibling cases (i.e., individuals infected by the same source) in superspreading events. This method requires infection network data obtained by FCT. It is generally regarded that the effectiveness of BCT is higher than FCT for outbreak control.

Proactive Contact Tracing (PCT) evaluates and predicts the risk level of individuals to send early warning signals. This approach uses a numerical index to assess the potential risk of infection for both symptomatic (or susceptible) and healthy individuals based on non-binary clues (e.g., symptoms, geodesic distance from confirmed cases, and pre-existing medical conditions), which can be obtained earlier than the virus test results. The proactive

estimates of expected infectiousness provide personalized health recommendations and share an evaluated health status with the close contacts of an individual.

In Figure 1, we use a contact tracing network to illustrate how the three DCT strategies are used to notify individuals. Person *A* reports a positive case, which is considered the index case. Then, FCT finds all the direct contacts of Person *A* (i.e., Person *B*–*F*) and informs them with health advice. If these contacts have later tested positive for the disease, then FCT continues to trace the close contacts of those infected individuals until no more positive cases are reported. In this scenario, FCT is not able to trace the disease transmission between Person *N* and Person *A*. Using BCT, a connection between Person *N* and Person *A* can be established based on factors like the date of infection or virus variant. To find the source case, an infection network that consists of only infected individuals is evaluated. For example, as Person *N* has the most infected neighbors, Person *N* is most contagious. It is deduced that Person *N* is more likely to be the source case and a superspreader in this infection network. Conversely, since Person *L* is not in close contact with any infected individuals, neither FCT nor BCT can provide any exposure notifications to Person *L*. Using proximity-based PCT, Person *L* receives an early warning signal that at least one positive case is 2 degrees of separation away from him.

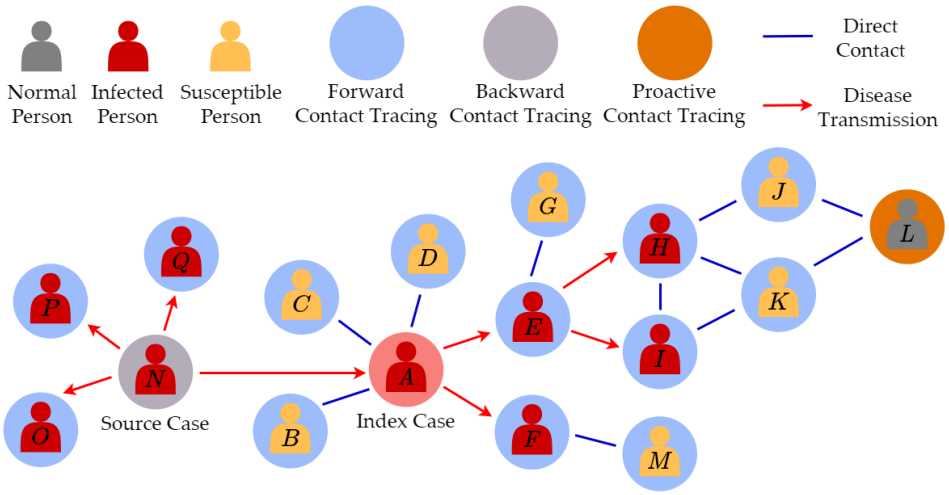


Figure 1. The contact tracing network showing how various DCT strategies can be applied. After identifying the index case (represented by the red colored circle), health agencies can then use different tracing strategies to find all the possible contacts and continually expand the network to prevent superspreading events.

2.3. Existing COVID-19 Digital Contact Tracing Applications

In the following, we review eight different DCT apps developed to mitigate the impact of the COVID-19 pandemic and categorize them using the concept of the above DCT strategies. We choose these eight apps for review as they represent a diverse range of strategies and techniques for contact tracing and have been widely adopted by users, making them an appropriate sample for analyzing real-world DCT apps. Table 3 summarizes the information of the eight considered DCT apps.

How We Feel App [47] is a voluntary crowdsourcing DCT app that gathers real-time individual-level data on the spread of the COVID-19 disease such as health conditions, COVID-19 test results, and pertinent lifestyle data. Most importantly, users need to self-report how they feel (well or not well) on a daily basis. If they are not feeling well, more details about symptoms are required. The app then applies statistical and machine learning methods with the aggregated data to estimate the prevalence of the disease in different regions of the United States and predict the percentage of having the disease for those who have provided their symptoms. The generated results and information provide insights and early warning signals for policymakers, medical professionals, and public health officials to take precautions for COVID-19 pandemic response.

Table 3. An overview of different DCT apps for COVID-19 pandemic response. Since the epidemic situation has gradually stabilized worldwide, most apps are no longer mandatory for citizens. Therefore, we only consider if an app had been made mandatory during the early outbreak period.

Tracing App	Location	Government App	Mandatory Use	Tracing Strategy
How We Feel App	United States	✗	✗	Proactive
LeaveHomeSafe App	Hong Kong	✓	✓	Forward
NHS COVID-19 App	England and Wales	✓	✗	Proactive
NOVID App	United States	✗	✗	Proactive
Outbreaks Near Me App	United States, Canada, and Mexico	✗	✗	Proactive
Taiwan Social Distancing App	Taiwan	✓	✗	Forward
TousAntiCovid App	France	✓	✗	Forward
TraceTogether App	Singapore	✓	✓	Forward

LeaveHomeSafe App, launched by the Hong Kong Government, is mandatory for all Hong Kong citizens as part of the disease prevention measures. Users must scan a venue's QR code with the app before entering designated premises such as theaters, restaurants, and government agencies to record travel histories. If a COVID-19-positive case is reported, users who visited the same place at about the same time as the recently infected patient will receive notification about when and where the exposure may have occurred.

NHS COVID-19 App [48], launched by the United Kingdom Government, is a voluntary DCT app for reducing the spread of the COVID-19 virus. It uses Bluetooth to exchange randomly-generated codes with nearby devices to record close contacts who have been around for more than 15 min within a distance of 2 m. The app not only sends alerts to users who have been in close contact with COVID-19-infected patients but also provides a symptoms checker. If a user reports COVID-19 symptoms, a risk score is computed for all the direct contacts to estimate the risk of infection as an early warning sign. The risk score for an interaction is calculated based on the distance between users, duration of the contact, and infectiousness of the COVID-19 carrier at the time of the interaction. If the risk score is higher than a given threshold, the app sends additional health advice for virus prevention.

NOVID App [49] is a voluntary contact tracing app that uses a network-based approach to model users and social contacts between users as vertices and edges, respectively, in contact tracing networks. For each positive case, instead of just finding direct contacts via Bluetooth and ultrasound, it uses network distance to send pre-exposure notifications, telling everyone how far they were away from the disease that just happened. The network distance between two vertices in a network is the length of the shortest path required to traverse from one vertex to the other along the network. The hop is the unit of network distance. For example, for a newly infected vertex v, instead of only notifying the vertices at network distance 1, it will notify all other vertices that are within 12 hops from v. This approach uses the shortest-path network distance to measure the risk level of each user rather than geographic distance. By continually providing this network distance information and growing the contact tracing networks over time, every individual user can foresee the potential risk (infection approaching or receding) in their own network.

Outbreaks Near Me App [50] is a crowdsourcing DCT app that allows users from the general public to voluntarily report their symptoms and virus test results for COVID-19. It verifies and analyzes the crowdsourced data through cross-validation with other sources (e.g., HealthMap) to ensure the quality of information. Reports are gathered and mapped to provide early warning signals so that users can search and browse real-time COVID-19 outbreak information on an interactive map, thereby knowing when COVID-19 is approaching their community. National public health agencies of the United States can view and analyze the data to search for potential outbreaks and predict where COVID-19 will impact next.

Taiwan Social Distancing App, launched by the Taiwan Artificial Intelligence Laboratory, is a voluntary DCT app for reducing the risk of COVID-19 transmission. It exchanges

a random hashed ID with other devices using Bluetooth to detect and record nearby users. When a user tests positive for COVID-19, the app shares the health status and sends COVID-19 exposure notifications to all the close contacts, who have passed the infected user for at least 2 min within a radius of 2 m, based on the contact history.

TousAntiCovid App is a voluntary DCT app launched by the French national research institution to monitor the spread of the COVID-19 pandemic. It uses Bluetooth to exchange user IDs and record the history of nearby contacts. If a user has tested positive for COVID-19, the direct contacts who have stayed with the recently confirmed COVID-19 patient for at least 5 min within a radius of 2 m will receive an exposure notification.

TraceTogether App, launched by the Government of Singapore, is mandated for all Singaporeans attempting to enter all places, such as shopping malls, workplaces, and schools, for community-driven contact tracing purposes. It uses Bluetooth to communicate with nearby devices and finds direct contacts when they are in close proximity. If a user tests positive for COVID-19, the app identifies and contacts people who are potentially exposed to the virus based on the direct contacts of the infected user.

All of these apps facilitate community-based disease surveillance. FCT apps are usually deployed by the government as an anti-epidemic measure to quickly find the close contacts of infected patients. These apps seek to provide users with the same primary value: After receiving a positive virus test result, the app traces the users who were directly exposed to an infected individual and notifies them to take precautions in order to protect society. In contrast, BCT apps rely on the contact tracing information collected by FCT to make further inferences on the most likely spreading source, whose recent movement history, once publicized as a form of alert, can warn the public. Instead of passively waiting for the post-exposure signals, PCT apps tend to predict and evaluate the potential risk of infection for non-close contacts based on different quantitative factors. The early warning signals provided by the PCT apps can help prevent and reduce the number and impact of superspreading events in the community.

2.4. Related Work

In this subsection, we examine the existing literature pertaining to the three previously mentioned DCT strategies. Table 4 summarizes the related work considered in the paper.

Table 4. A summary of the related studies discussed in this paper for the three DCT strategies.

Tracing Strategy	Year	Related Work
Forward	2020	Hellewell et al. [51], Aleta et al. [52]
	2021	Hinch et al. [53], Grantz et al. [54]
	2022	Yu et al. [34], Tan et al. [55]
Backward	2020	Endo et al. [56]
	2021	Müller et al. [57], Kojaku et al. [58]
	2022	Tan et al. [55], Raymenants et al. [59]
Proactive	2009	Ginsberg et al. [60]
	2020	Gupta et al. [38], Gallotti et al. [61], Briers et al. [62], Herbrich et al. [63]
	2021	Leung et al. [64], Bengio et al. [37], Baker et al. [65], Murphy et al. [66], Fenton et al. [67]
	2022	Lorch et al. [68]
	2023	Rivest et al. [69], Gupta et al. [70], Feng et al. [71]

Focusing on FCT, several research papers have been dedicated to evaluating and enhancing its effectiveness. The study in [53] introduces an agent-based simulation model to assess the impact of manual and digital contact tracing, alongside other interventions, on COVID-19 transmission. The authors in [54] present a mathematical modeling framework that evaluates the influence of test-trace-isolate programs on reducing the reproductive number of COVID-19, contributing to a comprehensive public health response. In [51], a stochastic transmission model is employed to evaluate the potential effectiveness of contact tracing and case isolation in controlling a COVID-19-like pathogen outbreak, quantifying the maximum number of cases traced weekly to assess the feasibility of public health efforts. In [52], an agent-based model is proposed to evaluate the impact of an enhanced testing and contact tracing response system on COVID-19 transmission in the Boston metropolitan area, using integrated mobility and demographic data. Building contact tracing networks is essential in FCT to track disease transmission. The authors in [34] propose a novel contact tracing network model for social contacts, representing each vertex as either an infected person or a visited location, with edges indicating a connection between a person and a place or two places sharing a common visitor. The study in [55] applies graph traversal algorithms, such as breadth-first search (BFS) and depth-first search (DFS), for constructing contact tracing networks in FCT scenarios.

For BCT, the work in [57] proposes bidirectional tracing using percolation theory to estimate eradication probability and identify super-spreaders, despite practical challenges. In [55], the authors introduce DeepTrace, a Graph Neural Network (GNN) framework for BCT, which outperforms prior heuristics. The study in [58] demonstrates the effectiveness of backward tracing compared to forward tracing and suggests revising strategies to include deep tracing while preserving privacy. In [56], the authors explore combining backward and forward contact tracing for COVID-19 control using a branching process model and highlight the increased potential for outbreak control. The work in [59] extends the tracing window to improve case identification and supports the implementation of BCT for rigorous suppression of viral transmission.

For PCT, the work in [69] focuses on designing exposure-detection functions for personal devices to reduce COVID-19 exposure risk while preserving privacy. In [64], the authors use digital proxies of human mobility to monitor viral transmissibility and assess social distancing effectiveness. Various methods have been proposed to quantify the risk. The authors in [37,38,70] leverage the rich suite of individual-level features such as age and lifestyle habits as inputs to deep learning models [72] trained to predict the COVID-19 infectiousness of susceptible individuals. The work in [68] uses a sampling algorithm with Bayesian optimization and longitudinal case data to estimate the transmission rate of infected individuals in their households and at the locations they visited. Google Flu Trends [60] applies a linear model to Google search query data to estimate the epidemic risk of influenza. In [61], the authors use the follower counts of tweet authors to evaluate the COVID-19 infodemic risks of news on Twitter (i.e., risk of exposure to fake news). Some probabilistic estimation methods have been developed to evaluate the risk. For example, the authors in [65] develop Bayesian inference methods to estimate the probabilistic infection risk for epidemic control. IDRLECA [71] employs a GNN model with deep reinforcement learning to compute the current infection probability of each individual. The work in [62] proposes a risk-scoring algorithm and associates the risk score with the probability of infection for a contact tracing app. In [66], the authors use the features collected by the GAEN system to compute the risk score and estimate the probability of infection. CRISP [63] is a probabilistic model that uses Gibbs sampling and SEIR model to predict the individual-level COVID-19 infection risk. The authors in [67] apply a Bayesian network model to compute the probability of a user contracting COVID-19.

The existing techniques in the related work have made significant strides in the three primary DCT strategies. However, several technical gaps can be observed, which have inspired the exploration of alternative methodologies in this study.

- Many existing techniques in FCT primarily focus on identifying direct contacts, often overlooking other potential transmission routes. Utilizing a novel network model to represent social contacts typically relies on existing datasets, which may be impractical in certain situations. Additionally, even when employing graph traversal algorithms, there is still a need for scalable, efficient, and accurate algorithm designs to determine contacts effectively.
- In BCT, numerous existing methods utilize approximate algorithms for infection source identification, which could lead to a less-than-ideal accuracy performance. Certain approaches require supplementary personal private information or hard-to-obtain data, such as precise timestamps of transmission events, limiting the feasibility and practicality in real-world applications.
- For risk estimation in PCT, most existing methods focus on specialized applications or are heavily dependent on specific feature types. This may constrain the adaptability and efficacy across a broad range of infectious diseases and diverse populations.

By targeting the observed technical gaps, our methodology seeks to develop more efficient, accurate, and adaptable approaches for the three primary DCT strategies.

3. Methods

This section outlines the review methods employed in this study, which followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.

3.1. Scope Identification

To determine the focus of this study, we formulated relevant research questions on machine learning in DCT. In view of our objective to model the social interactions among individuals, the terms social graph, contact tracing network, and contact graph were used interchangeably as we explored the following questions:

- *Contact Graph Construction*: As the underlying contact tracing network in FCT is usually unknown, how to construct such a network efficiently starting from a given index case?
- *Infection Source Estimation (Source Attribution)*: Given the contact tracing network data from FCT, how to accurately and efficiently identify the best source estimator in BCT?
- *Risk of Infectious Exposure Prediction*: How to reasonably quantify and estimate the risk of infectious exposure for non-close contacts in PCT?

Studying these problems in DCT is crucial as they can have a significant impact on controlling and managing the spread of infectious diseases. Contact graph construction is essential in understanding the transmission dynamics of an infectious disease by identifying the network of individuals that have been in contact with an infected person. Accurate and efficient infection source estimation is critical for identifying the origin of an outbreak, tracing the transmission routes, and implementing effective control measures. By estimating the risk of infectious exposure, individuals can be alerted and take necessary preventive measures to avoid getting infected. Therefore, addressing these problems can lead to the development of effective DCT systems, providing appropriate control measures.

3.2. Search Strategy

We devised a comprehensive search strategy to identify recent and relevant research articles centered on machine learning in DCT. Our scoping searches encompassed a variety of databases, including Institute of Electrical and Electronics Engineers (IEEE) Xplore, Association for Computing Machinery (ACM) Digital Library, Nature, arXiv, Elsevier (ScienceDirect), Multidisciplinary Digital Publishing Institute (MDPI), PLoS, medRxiv, Science Magazine, Frontiers, PubMed, Springer, The New England Journal of Medicine (NEJM), National Academy of Sciences, Massachusetts Institute of Technology (MIT) Libraries, MIT Press, BMJ, Bentham Science Publishers, USENIX, Stanford InfoLab Publication Server, Taylor & Francis, Journal of Open Source Software, Journal of Medical Internet Research (JMIR) Publications, Proceedings of Machine Learning Research, Wellcome Open Research,

Social Science Research Network, The Royal Society, Cochrane Library, EDP Sciences, and Oxford Academic. In addition to the mentioned databases, we also utilized Google Scholar and Google Search as supplementary search tools. Our primary search period spanned from 2019 to 2023, capturing the most recent advancements in machine learning applied to DCT. However, to ensure a comprehensive understanding and provide context, we also sought out select papers published before this time-frame that focused on foundational machine learning models and algorithms relevant to our study.

The essential keywords were identified using Boolean operators AND and OR. The keywords used for the searches through databases were contact tracing, coronavirus, COVID-19, epidemic, pandemic, computational epidemiology, artificial intelligence, machine learning, deep learning, generative learning, discriminative learning, and federated learning. To refine our search strategy and align it more closely with the proposed taxonomy, we specifically incorporated additional keywords, including digital contact tracing, forward contact tracing, backward contact tracing, and proactive contact tracing. This targeted approach allowed us to identify articles that were particularly relevant to each of the distinct contact tracing strategies under examination. We focused on articles related to constructing contact graphs using localization techniques with machine learning, estimating infection sources by machine learning, and predicting the risk of infectious exposure by deep learning. Additionally, factors such as article titles, abstracts, and keywords were considered during the search process. We synthesized the information by extracting details such as study objectives and methodology, available datasets, types of artificial intelligence, machine learning, and deep learning techniques used, and performance outcomes.

We conducted a comprehensive search for relevant literature, identifying a total of 841 papers from various reputable publishers and preprint servers using the search criteria and keywords described earlier. Subsequently, a two-stage screening process was carried out to further refine the selection of relevant articles. First, we performed a manual screening of titles and abstracts to evaluate their relevance to the application of machine learning in DCT. During this initial screening phase, we excluded 427 articles that did not meet our defined criteria, mainly due to the absence of machine learning techniques in the context of DCT or insufficient relevance to the topic. Next, we conducted a full-text review of the remaining 414 articles to assess their eligibility for inclusion in our study, as further elaborated in the subsequent paragraph. This in-depth review allowed us to evaluate each article's methodology, results, and overall contribution to the field of machine learning-enabled DCT strategies. Following this detailed assessment, we selected 106 articles that met our inclusion criteria and provided valuable insights into the application of machine learning techniques for DCT. These 106 articles were then included in our study and cited throughout the paper, forming the basis of our comprehensive review. Figure 2 presents an overview of the review process utilized in this paper.

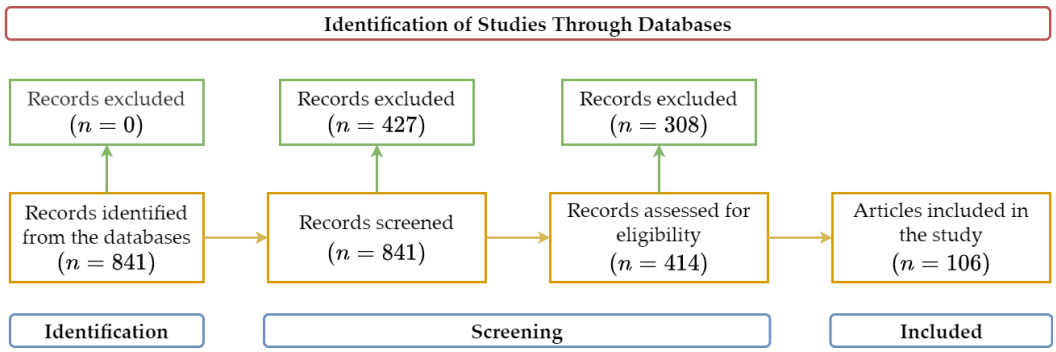


Figure 2. The review process of the study followed the PRISMA guidelines [73].

During the full-text review, we incorporated a quality assessment step to evaluate the methodological rigor and reliability of the 414 identified studies. This assessment was

conducted using a set of predefined criteria to ensure consistency and objectivity across all articles. The criteria considered the following aspects:

- Study design: Was the study designed and conducted using appropriate methods that align with its objectives?
- Sample size and data quality: Were the datasets employed in the study sufficiently large to draw meaningful conclusions?
- Machine learning techniques: Were the machine learning algorithms and techniques used in the study clearly described and justified with respect to the DCT strategy being investigated, and was the mathematical basis of the models presented in a comprehensible and evaluable manner?
- Performance evaluation: Were the performance metrics used to evaluate the machine learning techniques relevant, valid, and clearly reported?
- Validation and generalizability: Were the study findings validated using external datasets or cross-validation techniques, and do the results have the potential for generalization to broader contexts?

Each article was carefully evaluated based on the established criteria. Studies that did not meet the desired quality thresholds were subjected to further examination, and if the methodological rigor or reliability was found to be inadequate, the articles were excluded from the review. This quality assessment step ensured that our review was grounded in reliable and high-quality research findings, which in turn allowed us to offer a more precise and robust assessment of the current state of the art in machine learning-enabled DCT strategies. Additionally, this process aided in identifying potential limitations and areas for improvement in the existing literature, serving as a guide for recommendations for future research and development in this domain.

4. Privacy-Preserving Machine Learning-Based Digital Contact Tracing

This section introduces privacy-preserving machine learning techniques to solve the computational epidemiology problems of the three DCT strategies mentioned in Section 3.1. Figure 3 illustrates the proposed approaches for optimizing DCT strategies by incorporating big data and machine learning methods while considering privacy specifications.

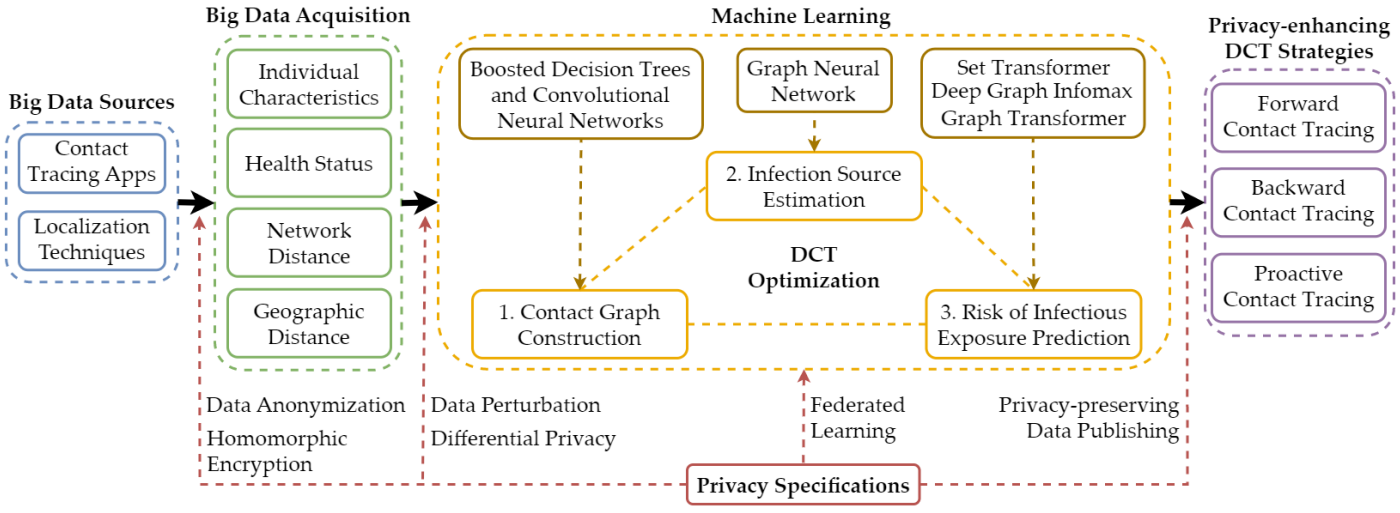


Figure 3. A high-level overview of the methodology applied to optimize DCT strategies, incorporating privacy specifications along with big data and machine learning techniques.

4.1. Privacy Specifications

The incorporation of privacy specifications is critical in the design and implementation of machine learning-based DCT systems. Given the sensitive nature of contact tracing data, a variety of methods can be utilized to enhance privacy and security. In the following, we elaborate on four main privacy specifications.

- Privacy-preserving data acquisition: The acquisition of large-scale contact tracing data should adhere to strict privacy standards. Data anonymization and homomorphic encryption are two methods that uphold these protocols. Anonymization eliminates personal identifiers from data, while homomorphic encryption enables calculations on encrypted data without decryption, thus maintaining privacy.
- Differential privacy and data perturbation in model training: During the machine learning model training phase, differential privacy and data perturbation techniques can be used to ensure privacy. Differential privacy adds a calculated amount of noise to the data or the queries, offering a mathematical guarantee of privacy by ensuring that the removal or addition of a single database item does not significantly affect the outcome of any analysis. Data perturbation, on the other hand, modifies the data slightly so that individual's private data cannot be identified or inferred, yet the overall statistical characteristics of the data remain accurate for model training.
- Federated learning: This technique offers a decentralized approach to machine learning where the model is trained across multiple decentralized devices or servers holding local data samples without exchanging their data. It ensures data privacy as all the raw data remains on the local device, and only model updates are communicated back to a central server for aggregation (see details in Section 6.1).
- Privacy-preserving data publishing: This privacy specification is aimed at protecting privacy when disseminating information derived from the DCT system. Techniques under this specification ensure that the data released for public consumption (e.g., statistics, graphs, reports) does not allow the re-identification of individuals or reveal any sensitive information.

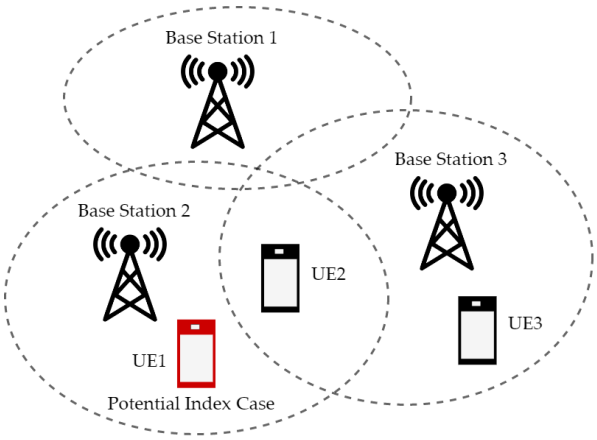
The importance of privacy specifications is not just about optimizing DCT strategies; they also help build public trust. Privacy concerns are one of the major reasons that individuals hesitate to use DCT apps. Fear of personal data misuse or leakage may lead to reluctance in adoption, thereby reducing the effectiveness of these apps in pandemic response. As such, providing robust privacy protections not only maintains the highest degree of user privacy but also fosters user trust in DCT systems. The more transparent and privacy-conscious a DCT system is, the higher its adoption rate is likely to be, amplifying its impact on managing pandemics.

4.2. Contact Graph Construction

Given the index case(s), we can generate a contact tracing network by identifying contacts or proximity with the index case(s). Contact tracing apps can be used to determine contacts. This approach, however, relies on the user's voluntary adoption, which can be low. In contrast, several localization techniques can be utilized to estimate contacts and construct the contact tracing network, as listed below.

- Wireless signal signatures such as indoor WiFi signal angle of arrival (AoA) measurements with respect to multiple nearby access points have been studied to localize user equipment (UE) [74].
- LTE signal fingerprinting measures the channel signatures in different locations to form a database. A UE is localized by matching the user's fingerprints (channel signatures measurements) to the database [75].

However, with the broad coverage of the cellular subscription service and the rich channel signatures information captured by Network Service Providers (NSPs), it is more promising to estimate users' proximity (rather than ground truth location) in order to build a concrete contact tracing network. An example is shown in Figure 4, where a potential index case UE1 is more likely to share a similarity in cellular signal signatures received from the three base stations with neighboring UE2 than with further UE3. In this sense, cellular signal signatures can serve as a representation of the UEs and be utilized as features for identifying the proximity between different UEs.



Cellular Channel Signatures

UE	RSRP_1	RSRP_2	RSRP_3
UE1	−120	−95	−100
UE2	−120	−100	−95
UE3	−125	−125	−85

Features

UE1: $\mathbf{x}_i = \{-120, -95, -100\}$
UE2: $\mathbf{x}_i = \{-120, -100, -95\}$
UE3: $\mathbf{x}_i = \{-125, -125, -85\}$

Figure 4. An example of the UEs within three base stations [76]. The sets of cellular channel signatures (features $\mathbf{x}_i$, where RSRP indicates the Reference Signal Received Power) can represent the UEs, and a proximal pair of UEs are likely to have similar signal signatures.

With the features that represent the UEs, machine learning techniques can serve as the estimator to identify contacts among UEs and generate the contact tracing graph. Specifically, for a pair of UEs at a specific timestamp, their channel signatures at this time are fed as input features to the estimator. The estimator, therefore, performs as a binary classifier to predict whether there is a contact between this pair of UEs at this given time as the output. All the outputs over time are used to construct the contact tracing network, with UEs being the vertices and estimated contacts being the edges. In [76], boosted decision trees [77–79] and convolutional neural networks (CNNs) [80–82] are used as the estimator to determine proximity among UEs using 5G mmWave channel signatures as the features. Figure 5 illustrates the architecture of the CNN model.

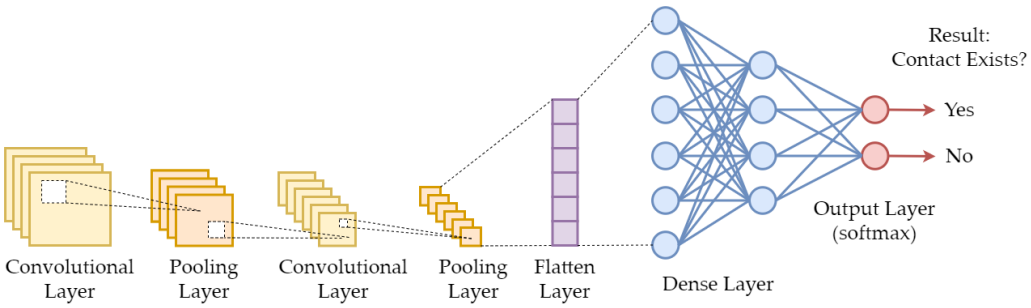


Figure 5. The CNN architecture [80] for estimating contacts among UEs with binary classification consists of 2 convolutional layers and a softmax layer.

Formally, we can model the contact tracing graph as a time-varying proximity graph with the sets of UEs (V) and the interconnections among UEs ($G[t] = (V, E[t])$, which is a time-evolving undirected graph). We first model the proximity problem (the estimation of close proximity between a pair of UEs at a given time) as a binary classification problem. To be more specific, given a pair of UEs, $v_i \in V$ and $v_j \in V$ at time t, the edge between this pair of UEs ($w_{ij}[t] \in E[t]$) indicates the contact. Each UE can then be represented by a set of features $\mathbf{x}_i$ and $\mathbf{x}_j$, which can be obtained from WiFi or Bluetooth signal signatures, LTE signal fingerprints, or 5G mmWave channel signatures provided by the NSP or measured by the UE. Our goal is to train a classifier to solve:

$$\underset{g}{\text{minimize}} \quad \sum_{u,v \in V} \mathcal{L}(w_{ij} - g(\mathbf{x}_i, \mathbf{x}_j)),$$

where $g(\cdot)$ is the estimator, and $\mathcal{L}$ is a generic loss function, e.g., L2 norm. Once the estimator has been learned, the contact tracing graphs can be constructed by estimating contacts $\tilde{w}_{ij}[t] = g(\mathbf{x}_i[t], \mathbf{x}_j[t])$ as the edges among UEs.

4.3. Infection Source Estimation (Source Attribution)

After constructing a contact tracing network using the FCT method, we may ask who is the index case that leads to this contact graph. That is, given a contact tracing graph, which vertex is the most probable infection source? In general, without further information, such as infection time or an infection spreading order, we can only construct the source estimator based on the topology of the contact graph. In the following, we provide an overview of solving the infection source estimation problem by the network centrality approach or graph neural network (GNN) approach.

The problem of estimating the infection source can be formulated as a maximum likelihood estimation problem on a graph structure [83,84]. As in the previous subsection, we first define a real-world social graph G as a set of vertices V and edges E, where each vertex represents an individual, and each edge represents a social connection between two individuals. We assume that the outbreak of the disease on G starts from a random vertex in G selected uniformly, i.e., the source v^*. We further assume the transmission of the disease follows the susceptible-infectious (SI) model considered in [34,83]. The SI model assumes that once an individual is infected, it stays infected and can transmit the disease to its susceptible neighbors through the incident edges. Hence, all infected individuals form a connected subgraph of G denoted as G_N, where N is the total number of infected vertices. Given a snapshot observation of the infected graph $G_N \subset G$, our goal is to find the vertex $\hat{v}$ that maximizes the likelihood $P(v = \text{source}|G_N)$. By Bayes' theorem, we can mathematically represent the source detection problem as an optimization problem on graph constrained:

$$\begin{aligned} \underset{v \in G_N}{\text{maximize}} \quad & P(G_N | v = \text{ source}) \\ \text{subject to} \quad & G_N \subset G. \end{aligned} \tag{1}$$

For brevity, in the following analysis, we denote $P(G_N|v = \text{ source})$ as $P(G_N|v)$.

4.3.1. Network Centrality Approach

Solving Problem (1) is NP-hard when G is a general graph. However, it can be solved efficiently when G is an infinite-size degree regular tree. Let σ_i denote a possible spreading order starting from v that leads to the infected graph G_N, and $M(v, G_N)$ be the collection of all σ_i. Then, we have

$$P(G_N|v = \text{source}) = \sum_{\sigma_i \in M(v, G_N)} P(\sigma_i|v). \tag{2}$$

Had G been an infinite-size d-regular tree, then the probability $P(\sigma_i|v)$ is constant for all i. In particular, we have

$$P(\sigma|v) = \prod_{i=2}^{N} \frac{1}{\sum_{j=1}^{i-1} d(v_j) - 2(i-2)}, \tag{3}$$

where $d(v)$ is the degree of v. Therefore, combining Equations (2) and (3) results in $P(G_N|v) \propto |M(v, G_N)|$. The value $|M(v, G_N)|$ is called the rumor centrality of v in G_N, and the vertex with the maximum rumor centrality among all vertices in G_N is called the rumor center of G_N. In general, evaluating the rumor centrality of a vertex in an arbitrary graph is an NP-hard problem; however, we can compute $|M(v, G_N)|$ in $O(N)$ time when G_N is a tree [83,85]. As a result, Problem (1) can be solved optimally within $O(N^2)$ time complexity as G is a degree regular tree with an infinite size.

When G is a finite graph or a general graph with cycles, the probability $P(\sigma_i|v)$ is not a constant anymore. The computation of Equation (2) becomes even harder than

evaluating $|M(v, G_N)|$ on a general graph since Equation (2) requires listing out all possible σ_i and tracking its corresponding probability $P(\sigma_i|v)$. The work in [34] generalizes the results to the case when G is a bounded-size regular graph containing a single cycle. Each "irregular vertex", such as a degree-1 vertex or a vertex on a cycle, is assigned a weight to quantify its influence on the probability. A statistical distance centrality is then proposed to approximate the maximum likelihood estimator of the source. Compared to the NP-hardness of Problem (1), the statistical distance centrality of each vertex can be computed in $O(N^3)$ time complexity. Beyond rumor and distance-based centrality [34,86], other centrality measures, such as eccentricity-based centrality (e.g., Jordan centrality [87]), can also be utilized for estimating an infection source. These metrics can be adapted for time-varying graphs by computing the centrality measure on the graph at each timestep, then outputting the most likely source across all time. This can be achieved in $O(N^3T)$ [76].

4.3.2. Graph Neural Network Approach

The probability of being the source can be seen as a function defined on each vertex in a given network topology. Hence, we can view Problem (1) as a learning task such that the goal is to learn to maximize $P(G_N|v)$ for a given G_N. We introduce a learning framework for learning $P(G_N|v)$ based on the GNN. There are several types of GNNs, each with its own unique architecture and method for aggregating information from neighbors. GraphSAGE [88] is one of the most popular GNN inductive frameworks for generating vertex embeddings. In the prediction stage, the computational complexity of GraphSAGE is $O(k|E|)$, where k is the number of graph convolutional layers. As the prediction stage is highly parallelizable, a major advantage of the GNN-based approach lies in the low computational complexity. In particular, GNNs are designed to consider the graph structure and vertex relationships when optimizing learning parameters. GNNs operate on the vertices of a graph, where each vertex v is assigned a vector of features $\mathbf{h}_v^0$ that describes its properties. We denote the set of the neighbors of v in G_N as $N_{G_N}(v)$. Let $\mathbf{h}_{N_{G_N}(v)}^{(l)}$ and $\mathbf{h}_v^{(l)}$ be the neighborhood representation of vertex v and the representation of v itself in the lth layer, respectively. We first construct $\mathbf{h}_{N_{G_N}(v)}^{(l)}$ by aggregating information from neighboring vertices, then we concatenate $\mathbf{h}_{N_{G_N}(v)}^{(l)}$ and $\mathbf{h}_v^{(l-1)}$ to compute $\mathbf{h}_v^{(l)}$ through a non-linear function σ. This process is repeated iteratively for multiple rounds, allowing the network to incorporate information from multiple layers of neighbors. In general, this iterative process can be formulated as:

$$\mathbf{h}_{N_{G_N}(v)}^{(l)} = \text{AGGREGATE}(\{\mathbf{h}_u^{(l-1)} : u \in N_{G_N}(v)\});$$
$$\mathbf{h}_v^{(l)} = \sigma(\mathbf{W}^{(l)} \cdot \text{CONCAT}(\mathbf{h}_v^{(l-1)}, \mathbf{h}_{N_{G_N}(v)}^{(l)})).$$

Lastly, we can optimize the learning parameters $\{\mathbf{W}^{(l)} : \forall l > 0\}$ using graph-based loss functions, which aim to minimize the difference between nearby vertices.

The choice of feature representation can have a significant impact on the performance of the prediction. To select interpretable features, we can utilize prior research on a probabilistic analysis of Problem (1) [34,83,89]. We begin by introducing a vertex feature, the boundary distance ratio, linked to the graph boundary. The influence of the graph boundary on the likelihood has been studied in [34,89], which is crucial when the size of the network is finitely large. We first define the boundary distance $b(v_i)$ of a vertex v_i as the distance from v_i to the most distant leaf vertex at the boundary of the given network. Then, the boundary distance ratio of v_i is defined as $\check{r}(v_i) = \frac{b(v_i)}{\max_{v_j \in G_N} b(v_j)}$. Another vertex feature that directly affects the likelihood is the vertex degree. For example, the formulation in Equation (3) includes the vertices' degree [34,83]. Hence, we construct a vertex feature, the degree ratio, based on the vertex degree. We define the degree ratio of v_i as $r(v_i) = \frac{d(v_i)}{\sum_{j=1}^{N} d(v_j)}$,

where $d(v_i)$ is the degree of v_i. Intuitively, a vertex with more infected neighbors than other vertices has a higher probability of being infected. The generalized spreading model considered in [34] captures this property and leads to a vertex feature called the infected portion [55]. The infected portion of vertex v_i is defined as $\hat{r}(v_i) = \frac{\hat{d}(v_i)}{d(v_i)}$, where $\hat{d}(v_i)$ is the number of infected neighbors of v_i. It is worth noting that the computation of these vertex features has a linear computational complexity with respect to N. We can first pre-train the GNN model on networks with a heuristic solution to Problem (1) and then fine-tune the model using small networks with accurate labels. The flexibility and generalizability of GNN models allow us to apply the trained model to large-scale networks. The architecture shown in Figure 6 is a GNN specifically designed to predict the probability of each infected vertex being the source within the contact tracing network illustrated in Figure 1.

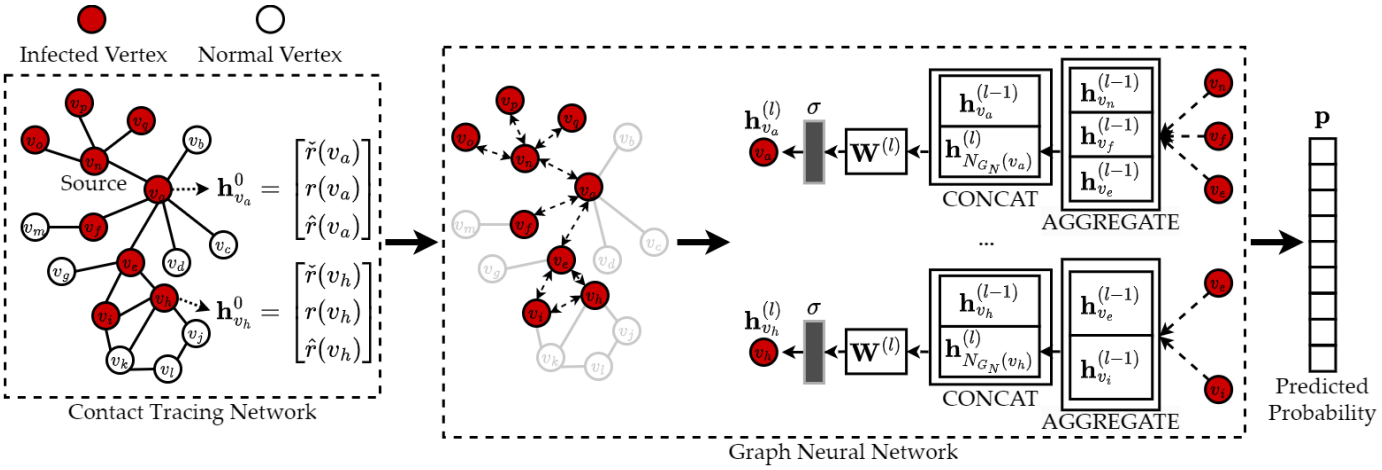


Figure 6. The GNN architecture [88] for predicting the probability of being the source for each infected vertex of a contact tracing network.

4.4. Risk of Infectious Exposure Prediction

In the following, we introduce three different approaches (feature-based, network-based, and rank-based) to quantify and predict the risk of infectious exposure for non-close contacts in PCT.

4.4.1. Feature-Based Approach

This approach aims to identify a diverse set of features that have the potential to impact an individual's risk of infection. By utilizing these features, this approach can estimate the likelihood of infection and generate personalized disease prevention recommendations. In contrast to other approaches, this method circumvents the need for centralized storage of the contact graph, which aligns with privacy constraints prevalent in many societies.

The locally observable features can be obtained through self-reporting by individuals for predicting the risk of infection [38,90]. We divide these features into two groups: Individual characteristics, which are static and unlikely to change significantly in the short term, and health status, which includes features that can vary on a daily basis and can be updated frequently. Below are the features that can be considered for each group.

- Individual characteristics:
  - Age: Older individuals are often considered to be more vulnerable to infections, and they may need tailored recommendations for disease prevention.
  - Gender: Certain risk factors may be specific to gender and could affect an individual's likelihood of infection.
  - Pre-existing health conditions: The health conditions, such as asthma or diabetes, can increase an individual's risk of developing severe illness if infected.

  - Lifestyle habits: Bad habits, such as smoking, can have various negative impacts on an individual's health, including a weakened immune system and a higher risk of developing respiratory infections.
- Health status:
  - Symptoms: The symptoms, such as fever, cough, and loss of taste or smell, are common indicators for estimating an individual's risk of infection.
  - Test results: The results of virus tests can indicate if an individual has an active infection and can help estimate their risk of spreading the virus to others.

In addition to these two feature types, the work in [37] suggests we should also consider daily encounters to estimate the potential infectiousness of encountered individuals such that these contacts can estimate their past infectiousness a posteriori. Let y_i^d be the infectiousness of an individual i on day d and $\mathbf{X}_i^d$ be the feature set over the past d_{max} days ($d \geq d_{max}$). The goal is to model the history of the individual's infectiousness in the last d_{max} days. This can be expressed as $P(\mathbf{y}_i^d | \mathbf{X}_i^d)$, where $\mathbf{y}_i^d = (y_i^d, y_i^{d-1}, \ldots, y_i^{d-d_{max}})$ is a vector that represents the current and past infectiousness of i. As both $\mathbf{y}_i^d$ and $\mathbf{X}_i^d$ are of the set data type, the work in [37] guarantees that any neural network that maps between sets can be used to model the correlation between the features and the infectiousness in this setting [91–93]. The architecture of the Set Transformer utilized to estimate the infectiousness of an individual based on the feature set is depicted in Figure 7.

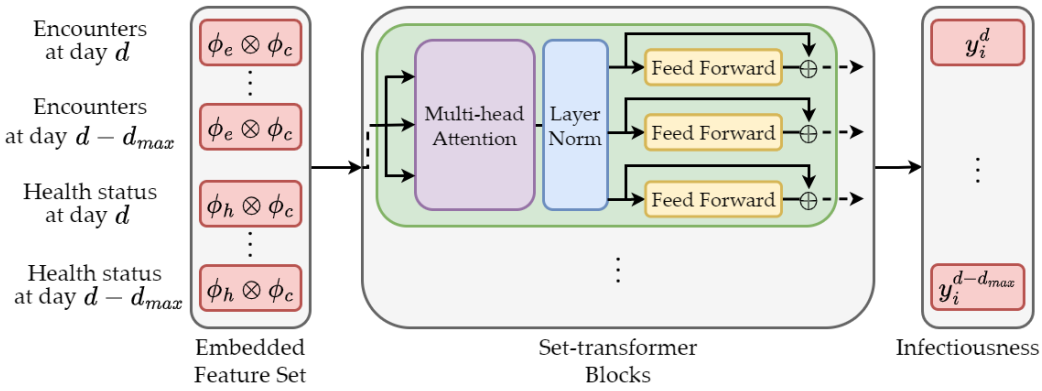


Figure 7. The Set Transformer architecture [92] for predicting the current and past infectiousness of individual i with the feature set. The embedding Multi-Layer Perceptron (MLP) modules for encounters, health status, and individual characteristics are denoted as ϕ_e, ϕ_h, and ϕ_c, respectively. The concatenation operation is denoted by $\otimes$, and the addition operation by $\oplus$.

The neural network model can undergo supervised learning by being trained on a dataset consisting of feature vectors and their corresponding labels that indicate the infectiousness level [72]. The model's weights and biases are iteratively adjusted during the offline training process to reduce the difference between its predicted output and the true output for each training example. It can then make online predictions on new, unseen data by taking the feature set as input and outputting the infectiousness level.

4.4.2. Network-Based Approach

The proximity-based approach relies on determining the shortest distance between infected patients and healthy individuals, which can be computed through network-based analysis in contact graphs [49,94] or estimated through machine learning or deep learning techniques to establish their proximity [95,96]. Figure 8 illustrates an example of how NOVID App uses the shortest network distance to send an early warning signal. User A receives a pre-exposure notification indicating that the closest infected user, User B, is at a distance of 2 from him. However, this method may overlook other potential risks of User A, such as the infected cluster at degree 3. Focusing solely on proximity may also disregard other nontrivial network properties. By leveraging the contact graph, one can

employ vertex embedding to capture both the local neighbor properties and the global graph structure.

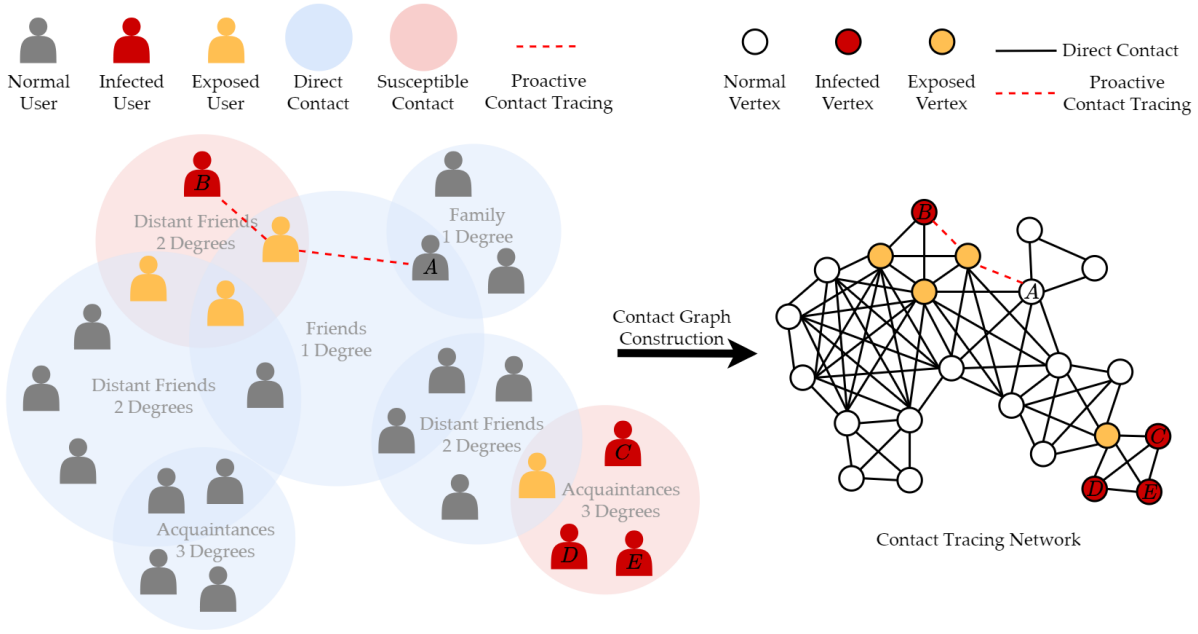


Figure 8. An example of how NOVID App [49] applies the proximity-based proactive contact tracing to estimate the risk of infectious exposure of User A.

We first present a method for leveraging the network properties to estimate the probability of infection. Under the SI model, vertices can either be susceptible or infected. We only focus on the infection likelihood of susceptible vertices in this subsection. Let $S_i(t)$ be the health status of a vertex v_i at time t such that $S_i(t) = 1$ if v_i is infected at time t, and $S_i(t) = 0$ otherwise. Our goal is to find the probability that v_i becomes infected at $t+1$ given that it is currently susceptible and its neighbors in the set $N_G(v_i)$ have a different health status in the contact graph G, i.e., $P(S_i(t+1) = 1 \mid S_i(t) = 0; \{S_j(t) \text{ for all } v_j \in N_G(v_i)\})$. We define $r_i = \frac{1}{|N_G(v_i)|} \sum_{v_j \in N_G(v_i)} (w_{ij} r_j)$ as the pandemic risk index of v_i, where w_{ij} is the edge weight between v_i and v_j, and r_j is the pandemic risk index of v_j. Assume the spreading rate of a disease is λ, then the infection probability of v_i is given as: $p_i = 1 - \exp(-\lambda r_i)$.

Deep Graph Infomax (DGI) [97] is a graph representation learning technique that learns vertex embeddings. It relies on maximizing the mutual information between vertex-level (local) and graph-level (global) representations for learning an encoder. DGI employs a discriminator to identify actual vertex representations from the negative samples obtained from a stochastic corruption function based on the graph-level representation. Let $G = (V, E)$ be a contact tracing network with N vertices. We compute the features of each vertex $v_i \in V$, $\mathcal{F} : \mathbb{R} \to \mathbb{R}^F$, such that $\mathcal{F}(v_i) = \mathbf{x}_i$, where $\mathbf{x}_i \in \mathbb{R}^F$ is the vertex features of v_i. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a set of vertex features and $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the adjacency matrix of G. The goal of DGI is to learn a graph convolutional encoder, $\mathcal{E} : \mathbb{R}^{N \times F} \times \mathbb{R}^{N \times N} \to \mathbb{R}^{N \times F'}$, such that $\mathcal{E}(\mathbf{X}, \mathbf{A}) = \mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$ represents the high-level patch representations $\mathbf{h}_i \in \mathbb{R}^{F'}$ for each v_i. Using a readout function, $\mathcal{R} : \mathbb{R}^{N \times F} \to \mathbb{R}^F$, we can summarize the patch representations into a graph-level representation, $\mathbf{s} = \mathcal{R}(\mathcal{E}(\mathbf{X}, \mathbf{A}))$. The discriminator, $\mathcal{D} : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$, then assigns the patch-summary pair a probability score, $\mathcal{D}(\mathbf{h}_i, \mathbf{s})$, to determine if the given pair is a positive or negative sample.

In DGI, the negative samples are produced with a stochastic corruption function, $\mathcal{C} : \mathbb{R}^{N\times F} \times \mathbb{R}^{N\times N} \rightarrow \mathbb{R}^{M\times F} \times \mathbb{R}^{M\times M}$. There is no explicit constraint on creating a negative sample using the corruption function (i.e., it can be a graph with any number of vertices and edges). Let $\tilde{G} = (\tilde{V}, \tilde{E})$ be a negative sample with a feature matrix $\tilde{\mathbf{X}}$ and adjacency matrix $\tilde{\mathbf{A}}$ generated using the corruption function. Based on the Jensen-Shannon divergence [98–101], the loss function of maximizing mutual information between $\mathbf{h}_i$ and $\mathbf{s}$ is given by:

$$\frac{1}{N+M}\left(\sum_{i=1}^{N}\mathbb{E}_{(\mathbf{X},\mathbf{A})}[\log \mathcal{D}(\mathbf{h}_i,\mathbf{s})] + \sum_{j=1}^{M}\mathbb{E}_{(\tilde{\mathbf{X}},\tilde{\mathbf{A}})}\left[\log\left(1-\mathcal{D}(\tilde{\mathbf{h}}_j,\mathbf{s})\right)\right]\right).$$

Figure 9 shows how DGI generates vertex embeddings using the contact graph as input. The vertex-level features such as the clustering coefficient [102,103] and centrality measures [104,105] can be computed and used as part of the feature vector of each vertex.

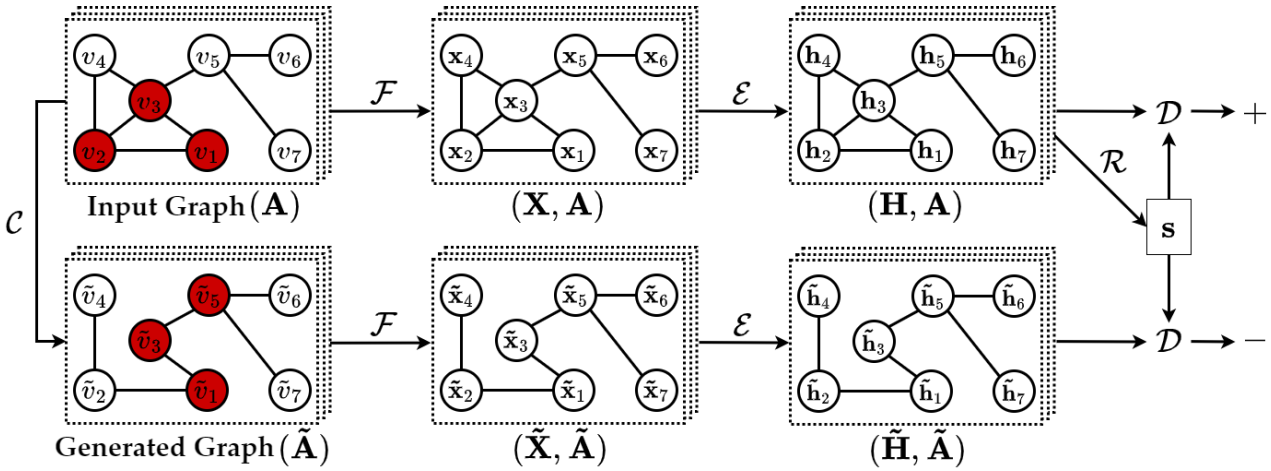


Figure 9. The DGI architecture [97] for learning vertex embeddings of a contact tracing network. The infected vertices are colored in red, and this feature is captured in $\mathbf{X}$ and $\tilde{\mathbf{X}}$.

To estimate the risk of infection for each susceptible vertex in the contact tracing network, we can assign labels to the risk level based on the computed infection probability, where values greater than 0.7 are labeled as high-risk, values between 0.3 and 0.7 are labeled as moderate-risk, and values less than 0.3 are labeled as low-risk. A supervised logistic regression model can then be trained to predict the risk of infection for the susceptible vertices using the learned embeddings as input and the corresponding risk level as output.

4.4.3. Rank-Based Approach

Bipartite graphs [106–108] can be used to model epidemics by representing the interactions between two types of vertices, such as people and venues. While much of the focus in epidemic modeling is on person-to-person transmission, it is important to consider the risk of transmission from people to venues and then back to people. For instance, a venue can become a hotspot for transmission if infected individuals have visited it or if large numbers of people have gathered there. By using a bipartite graph, we can model both types of interactions and capture the full extent of the risk of infection. In this model, people and venues are represented as vertices in the graph, and the edges between them indicate visits from people to venues. The PageRank algorithm [109] can be applied to the bipartite graph model to rank the risk of infection for individuals. This ranking considers the venues that people have visited and the number of other people who have also visited those venues. The more people who have visited a venue, the higher the risk of infection for all individuals who have visited that venue. Based on the venue visitation patterns, we can identify high-risk individuals who have visited high-risk venues. This information can be used to target public health interventions and reduce the spread of disease. It is worth

noting that the approach of using crowdsourced data and learning models to rank and assess a collection of objects has already been well-studied [110–112].

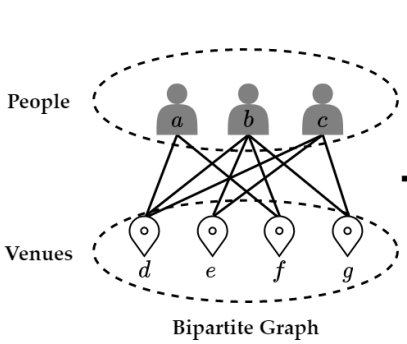
To apply the PageRank algorithm to a bipartite graph [113–115], we can represent the graph as an adjacency matrix, where the rows represent the people, and the columns represent the venues. The matrix entries indicate whether a person has visited a particular venue or not. Let the vertex sets of people and venues be V_P and V_S, respectively. Then, we denote the adjacency matrix of the bipartite graph as $\mathbf{A} \in \mathbb{R}^{|V_P| \times |V_S|}$, where

$$\mathbf{A}_{ij} = \begin{cases} w_{ij} & \text{if person } i \text{ visited venue } j \\ 0 & \text{otherwise.} \end{cases}$$

The edge weight w_{ij} represents the strength of the relationship between the connected vertices v_i and v_j. The diagonal matrix $\mathbf{D}_P$ for all vertices in set V_P is defined such that $(\mathbf{D}_P)_{ii} = d(v_i)$, where $d(v_i)$ is the sum of the weights of all edges connected to v_i. Similarly, the diagonal matrix $\mathbf{D}_S$ for all vertices in set V_S is defined in the same way. Let the ranking vectors of people and venues be $\mathbf{p} \in \mathbb{R}^{|V_P|}$ and $\mathbf{s} \in \mathbb{R}^{|V_S|}$, respectively. Then, the PageRank algorithm solves the following system of equations:

$$\begin{aligned} \mathbf{p} &= \alpha \mathbf{B}\mathbf{s} + (1-\alpha)\mathbf{p}_0; \\ \mathbf{s} &= \beta \mathbf{B}^T \mathbf{p} + (1-\beta)\mathbf{s}_0, \end{aligned} \tag{4}$$

where α and β are the damping factors, and $\mathbf{B} = \mathbf{D}_P^{-\frac{1}{2}} \mathbf{A} \mathbf{D}_S^{-\frac{1}{2}}$ is a symmetric normalization of $\mathbf{A}$. The query vectors $\mathbf{p}_0$ and $\mathbf{s}_0$ are typically determined based on prior knowledge or assumptions about the relative importance of different vertices in the graph. The example shown in Figure 10 demonstrates the application of PageRank to calculate the risk rankings of people and venues in a bipartite graph.



$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{D}_P = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{D}_S = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

People	Iteration 0	Iteration 1	⋯	Convergency	Rank
a	0.333	0.834	⋯	0.855	3
b	0.333	1.156	⋯	1.141	1
c	0.333	1.010	⋯	1.005	2

Venue	Iteration 0	Iteration 1	⋯	Convergency	Rank
d	0.25	1.18	⋯	1.144	1
e	0.25	0.909	⋯	0.946	3
f	0.25	1.002	⋯	0.963	2
g	0.25	0.909	⋯	0.946	3

Figure 10. An example of applying Equation (4) to compute the risk rankings of three people and four venues in a bipartite graph, with a damping factor of 0.85 (i.e., $\alpha = 0.85$ and $\beta = 0.85$) and edge weights set to 1. We initialize the elements in $\mathbf{p}_0$ with $\frac{1}{|V_P|}$ (i.e., $\frac{1}{3}$) and $\mathbf{s}_0$ with $\frac{1}{|V_S|}$ (i.e., $\frac{1}{4}$).

The Transformer [116] has become popular due to its proven efficacy in numerous applications, including pandemic response [43,117–119]. We can define the edge weights in the bipartite graph using learned embeddings from the Graph Transformer [120–125]. The embeddings can capture the historical behavior of people and venues, modeling the likelihood of a person visiting a particular venue based on their previous visits. To achieve this goal, we can consider different vertex features for people and venues. For instance, for people, we can use features like age, gender, occupation, or previous venues visited, whereas, for venues, we can use features like location, capacity, type of venue, or historical attendance to model the interactions between people and venues more accurately. Figure 11 illustrates the architecture of the Graph Transformer that employs scaled dot-product attention with Laplacian positional encodings to learn the vertex embeddings of a bipartite graph. We opt for the Graph Transformer over the GNN or DGI as it excels in

managing complex relationships and long-term dependencies. Once the embeddings of all vertices have been learned via the Graph Transformer, we can compute the edge weight between two vertices v_i and v_j as: $w_{ij} = \mathbf{h}_i^T \cdot \mathbf{h}_j$, where $\mathbf{h}_i$ and $\mathbf{h}_j$ are the learned embeddings of v_i and v_j, respectively. Note that if the embeddings have different dimensions, we can use a learned linear projection to map the embeddings to a common vector space before computing their dot product. This edge weight intuitively measures the similarity or relatedness between the two vertices based on their learned embeddings. Stronger edge weights indicate a greater similarity, while weaker edge weights indicate less similarity. This approach allows us to leverage the power of the Graph Transformer to learn rich, meaningful representations of the vertices, which in turn enables more accurate modeling of the underlying relationships between them.

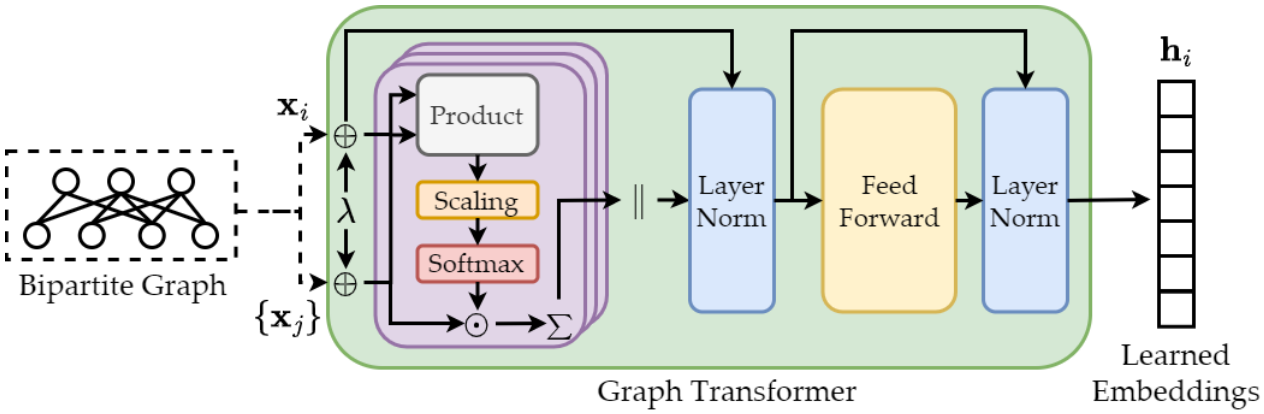


Figure 11. The Graph Transformer architecture [122] for learning vertex embeddings of a bipartite graph. We use the notation $\mathbf{x}_i$ to denote the features of vertex v_i, and $\{\mathbf{x}_j\}$ to denote the vertex features of its neighbors $\forall v_j \in N_G(v_i)$ in the contact graph G. The addition operation and Laplacian positional encodings are represented by $\oplus$ and λ, respectively. The symbol $\odot$ represents multiplication, and $\sum$ denotes the summation of all neighbor features. The attention block, which is colored purple, captures the attention mechanism used in the model. We use the symbol $\|$ to represent the concatenation of all the attention blocks. The learned embedding of v_i is denoted as $\mathbf{h}_i$.

5. Results and Analysis

In this section, we discuss the results and analysis of the suggested methodology. It is important to note that not all methods have undergone experimentation; however, those that have are derived from the studies reviewed herein.

The approach selected for contact graph construction is the first to employ machine learning methods for contact detection, setting it apart from traditional strategies. In [76], the performance of this machine learning-based approach was compared to Bluetooth-based contact tracing app baselines, with decision trees having a maximum depth of 6 and 100 trees, and a CNN with 2 convolutional layers containing 64 neurons each and a softmax layer. The results indicated that the chosen approach surpasses the contact tracing app baselines, achieving higher accuracy in detecting contacts. A notable advantage of this method is that it does not depend on user adoption to improve accuracy, unlike contact tracing apps. This autonomy from user adoption enables more efficient and consistent contact tracing, further emphasizing the potential of machine learning-based strategies in optimizing DCT. For infection source estimation, the preferred GNN approach in [55] utilized small-sized graphs as training data and underwent training for 150 epochs. The trained model was then evaluated on large-scale datasets, showcasing high accuracy and effectiveness in tackling the problem at hand. The pre-training and fine-tuning process revealed that supervised training accuracy does not decline significantly in infection source estimation as the number of vertices in the infection networks increases. This observation can be beneficial for developing data-driven models using small-sized networks as training data, anticipating that the models can still function for larger-sized networks. Employing smaller-sized networks

for training can also reduce training time and provide the opportunity to leverage transfer learning, enabling model adaptation for superspreader detection within more extensive networks. For risk of infectious exposure prediction using the feature-based approach, in [37], the preferred Set Transformer with 160 epochs and a batch size of 1024 required fewer computational resources for early and accurate detection of potential infection cases. This method demonstrated a notable advantage over the no-tracing baseline and substantially reduced the number of false quarantine recommendations compared to all existing DCT approaches. For the network-based approach, the chosen DGI model in [97] provided a stable and robust performance in unsupervised learning of vertex embeddings. The DGI loss function demonstrated the benefits of employing wider models rather than deeper ones. For the rank-based approach, the selected Graph Transformer in [122] featured a straightforward and adaptable architecture that significantly outperformed baseline isotropic and anisotropic GNNs in implementing vertex attention. Additionally, standard GNNs were unable to manage complex relationships and long-term dependencies effectively, unlike Transformer networks. In Table 5, we list GitHub links for the machine learning models, particularly for those selected studies that have open-sourced code, enabling researchers to conduct a more comprehensive implementation and evaluation of these models.

Table 5. A summary of available GitHub links for the machine learning models considered in this study, with all links accessed as of 20 April 2023.

Machine Learning Model	Year	GitHub Repository
DGI [97]	2019	https://github.com/PetarV-/DGI
Graph Transformer [122]	2020	https://github.com/graphdeeplearning/graphtransformer
Set Transformer [37]	2020	https://github.com/mila-iqia/COVI-ML
GNN [55]	2022	https://github.com/convexsoft/deeptrace

Most of the approaches presented in this review are pioneering and innovative solutions that have addressed the proposed research problems. As some of these methods are based on theoretical foundations and expert opinions, it can be challenging to provide immediate, measurable results, leading to certain missing experimental outcomes in our review. As a comprehensive review, our primary objective is to introduce inventive ideas for optimizing DCT strategies, inspiring researchers to delve deeper into these concepts and conduct experiments within their own studies. This review primarily focuses on providing insights and motivation for DCT optimization rather than performing extensive experiments for each approach. By emphasizing novel approaches and their potential implications, this study aims to contribute to the advancement of DCT optimization and encourage further research in this domain.

Contact tracing data is essential for pandemic response, as it helps identify patterns and trends that might otherwise remain unnoticed. By utilizing machine learning techniques, researchers can analyze and model complex relationships between individuals, their contacts, and disease transmission, leading to more efficient and effective DCT strategies. In Table 6, we provide a summary of available contact tracing datasets that can be employed for further evaluation of the proposed methodology. This enables researchers and practitioners to develop and refine machine learning approaches specifically for optimizing DCT. The application of machine learning in this area can significantly enhance our understanding of disease transmission dynamics and inform decision-making for public health officials, ultimately improving our ability to control infectious disease outbreaks.

Table 6. A summary of the available contact tracing datasets relevant to this study, with all links accessed as of 20 April 2023.

Reference	Year	Data Category	Dataset	Link
Xu et al. [126]	2020	Health Profile	Individual-level Epidemiological Data for COVID-19 Outbreak	https://github.com/beoutbreakprepared/nCoV2019
Gupta et al. [38]	2020	Health Profile	COVID-19 Mobility and Characteristics Simulation Dataset	https://github.com/mila-iqia/COVI-AgentSim
Firth et al. [10]	2020	Contact Graph	COVID-19 Infectious Disease Social Interaction Dataset	https://github.com/skissler/haslemere
Adam et al. [127]	2020	Contact Graph	COVID-19 Superspreading in Hong Kong	https://github.com/dcadam/covid-19-sse
Serafino et al. [128]	2022	Contact Graph	COVID-19 Digital Contact Tracing Geolocalized Human Mobility Dataset	https://github.com/makselab/COVID19
Moosa et al. [129]	2023	Contact Graph	COVID-19 Contact Tracing Networks	https://ieee-dataport.org/documents/covid-19-contact-tracing-networks
This Study	2023	Contact Graph	Digital Contact Tracing Dataset	https://dctracing.shinyapps.io/DCTracing/

6. Challenges of Digital Contact Tracing

Several challenges affect the implementation of machine learning-based DCT. This section summarizes some of these challenges and describes possible solutions.

6.1. Privacy and Security

There are fundamental data challenges associated with privacy and security in digital contact tracing. Most traditional contact tracing methods adopt a single trusted centralized server to aggregate information, but this leads to a single point of failure, i.e., a cyberattack can cause a massive data breach, which is of particular concern in healthcare applications such as DCT. Privacy concerns are also of paramount importance as the single server can learn a lot about the mobile users' local sensitive data from multiple rounds of two-way interactions that may occur during the training of the neural networks. Some users who have fewer training data will inadvertently need the single server to crowdsource more data, as accessing more data can lead to models with improved generalization capabilities in comparison with centralized supervised learning. Crowdsourcing more data in DCT can help to create a more diverse and representative sample of the population, which in turn can lead to a more robust and accurate model. This approach can be superior to centralized supervised learning, where models are trained on fixed datasets that may not capture the full range of variation in the population, often resulting in overfitting and poor generalization performance. However, this method comes with privacy concerns as the single server can access sensitive local data from multiple users. Thus, maintaining the balance between the need for crowdsourced data and the privacy of individuals is crucial.

Federated learning (FL) [130–134] is a machine learning approach to allow multiple devices to collaboratively train a single model without the need to exchange training data (i.e., each device only has access to its own local data). As contact tracing apps collect individuals' movements, FL can be applied to train a machine learning model to predict the likelihood of this particular user contracting an infectious disease based on their movement patterns. This model can also be trained using data from multiple devices without the need to transfer any raw data to a centralized server. Instead, the contact tracing apps can download a partially-trained model issued by the health authority on traits of the disease

(e.g., reproduction number and infection distance) and further refine this model using its own local data, and only sends model updates to the central server. This approach ensures that users' privacy is preserved while still enabling the development of effective contact tracing models. Additionally, FL can help ensure that the contact tracing model remains up-to-date and effective, as it can be continuously updated with new data from other users' devices. For real-time predictions, lightweight metadata may be optionally exchanged between devices to improve the predictions made by local federated models, trading off with privacy.

Differential privacy (DP) [135–137] is another approach to protect individual privacy by aggregating data in a way that preserves the anonymity of individuals while still allowing for the analysis of large DCT datasets using machine learning. The basic idea behind DP is to add noise to the data in a way that preserves the overall statistical properties of the data but prevents individual data points from being accurately reconstructed. This means that the output of an analysis or model will be largely the same whether or not a specific individual's data is included. However, the amount of noise added to the data must be carefully controlled in order to balance privacy and accuracy.

DP techniques can ensure that machine learning models are not biased or influenced by individual data points while still enabling accurate predictions and insights to be generated from the data. Machine learning techniques can be used to implement DP [138–149]. For example, a machine learning algorithm can be trained to generate synthetic data that is statistically similar to the original data but with added noise to prevent individual data points from being reconstructed. Another approach is the Laplace mechanism which adds Laplacian-distributed noise directly to the model's parameters during training. Yet another method to remove the need for a trusted centralized server for aggregation is to have a local DP mechanism applied to the updates from each distributed source before aggregating the high-dimensional data at a centralized server. This, however, requires a sufficiently large pool of distributed sources and thus poses potential problems in information leakage vulnerability and computation latency.

A framework of FL with formal DP guarantees proposed in [150] used traditional verification techniques to ensure that the FL process satisfied certain privacy and security properties while also using DP to further protect the privacy of users' data. Figure 12 shows the framework of FL with DP for DCT. By combining these two approaches, it is possible to achieve a higher level of privacy and security than what could be achieved with either approach alone [150–152]. The idea is to amplify privacy protection by utilizing the randomness in sampling training examples ("amplification-via-sampling") [153]. Since restricting the contributions of individual users and adding noise may affect model accuracy, it is crucial to maintain model quality while providing robust DP guarantees. However, in scenarios where users can provide multiple training examples, algorithms that ensure user-level DP necessitate that the output distribution of models remains unchanged even when the training examples from any individual user (or all examples from any particular device in the application) are modified. As FL aggregates all training data from a user into a single model update, federated algorithms are suitable for providing user-level DP guarantees. The Differentially Private Federated Averaging (DP-FedAvg) algorithm [154] is an example of a federated algorithm that extends the DP-SGD approach to the federated setting. This algorithm provides user-level DP guarantees and has been successfully deployed on mobile devices. Recent studies [155–157] presented the first consumer-scale next-word prediction (NWP) model trained with FL while leveraging the DP-FedAvg technique.

Other research on the development of practical FL infrastructure included training language models on mobile devices within wireless networks. The feasibility of training NWP models using the DP-FedAvg algorithm with user-level differential privacy has also been shown (in simulations on a public corpus). However, training production-quality NWP models using DP-FedAvg in a real-world production environment on a heterogeneous fleet of smartphones presents several challenges. For example, the central server must keep track of the available devices at the beginning of each round and randomly sample them

while ensuring the secrecy of the sample. Therefore, it is important to deploy contact tracing algorithms that incorporate a differentially private mechanism for training a production neural network in FL. It is also crucial to design the production training infrastructure to aggregate contact tracing data in a manner that guarantees the training mechanism is not overly sensitive to any particular user's data.

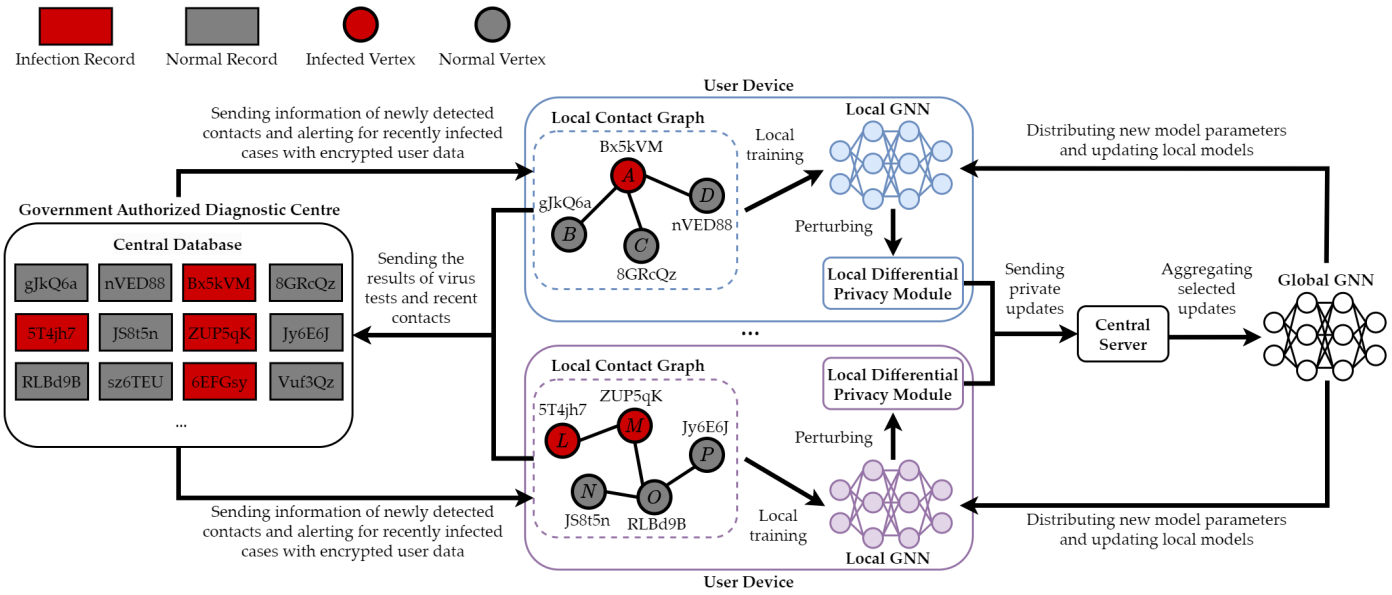


Figure 12. The architecture of the federated graph learning framework [151] with differential privacy [152] for digital contact tracing. The left-hand side illustrates privacy-preserving contact tracing network expansion, while the right-hand side demonstrates privacy-preserving model update.

Finally, secure multi-party computation and homomorphic encryption are two cryptographic techniques for ensuring privacy while computing joint functions of the inputs (e.g., the probability of contact). In the homomorphic encryption case, a central server that receives encrypted data from multiple users could infer the contact probability through specially designed functions and return the results to the users without directly decrypting the data. In the secure multi-party computation case, the assumption is that there is no central server, and the users could jointly compute their contact probabilities through a specially designed protocol while keeping their inputs private. Designing such techniques for learning models [158] or federated learning in particular [159,160] are open research problems. Applying such cryptographic methods to the contact tracing domain introduces further challenges, such as lightweight execution on local devices and computational complexity due to a large number of users.

6.2. Data Availability

During the early phase of a pandemic, such as the COVID-19 pandemic, the data available for contact tracing is often insufficient. It is difficult to determine the rate of infection and the scope of the outbreak as the knowledge about the disease and the number of confirmed cases is limited. The available data for contact tracing in this phase may not be complete or accurate, especially in areas with limited resources or low levels of digitalization. Under these conditions, the speed at which the pandemic spreads may outpace the ability to collect and process data, making it difficult to make informed decisions about how to respond to the outbreak. This lack of data can limit the effectiveness of contact tracing efforts and make it difficult to control the spread of the disease. For instance, if the data used to train the aforementioned GNN are insufficient, the performance of the GNN may be less than expected, as it requires a large amount of data to learn complex patterns in the network. Therefore, it is essential for public health authorities and researchers to

address the challenge of data availability during the early phase of a pandemic to inform effective response strategies.

Generative Adversarial Networks (GANs) [161–170] can be considered one of the solutions to the data scarcity problem in the early days of a pandemic. As a type of deep learning model, GANs can generate new synthetic data that resembles real-world data by learning patterns and relationships in existing data. In the context of contact tracing network data, a GAN could be used to generate synthetic graphs representing infection networks, given limited actual data as input. For example, NetGAN in [171] generated graphs that captured the underlying structural patterns of the real-world graph distributions by using random walks with a GAN-based approach. This can enable the creation of infection networks with similar characteristics observed in real-world outbreaks, even when real-world data is limited. The GAN consists of two parts: A generator that creates synthetic graphs and a discriminator that evaluates the quality of the generated graphs and provides feedback to improve the generator. The generator and discriminator would be trained together in an adversarial manner, with the generator trying to create graphs that are as close as possible to real infection networks and the discriminator trying to correctly identify which graphs are real and which are generated. The goal of this training process is to produce a generator capable of creating synthetic graphs similar to real graphs to be useful for analysis, despite the limited availability of real data.

To generate infection graphs using GAN, we use a generator to fool the discriminator by producing deceptive graphs and a discriminator to identify if an input graph is real or generated accurately. Both the generator and discriminator are neural networks that are used to solve a minimax two-player game. Let $\mathbf{x}$ be a graph drawn from a distribution p_{data} and $\mathbf{z}$ be a latent vector from some probability distribution p_z. The loss function of the GAN is

$$\min_{\mathcal{G}_{\text{GAN}}} \max_{\mathcal{D}_{\text{GAN}}} \mathbb{E}_{\mathbf{x}\sim p_{data}}[\log(\mathcal{D}_{\text{GAN}}(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_z}[\log(1 - \mathcal{D}_{\text{GAN}}(\mathcal{G}_{\text{GAN}}(\mathbf{z})))],$$

where $\mathcal{G}_{\text{GAN}}$ is the generator, and $\mathcal{D}_{\text{GAN}}$ is the discriminator. In the GAN architecture for generating infection networks, p_{data} indicates the distribution of the real-world data (i.e., infection networks, represented by the adjacency matrix) that the generator tries to imitate. The latent vector $\mathbf{z}$ is a low-dimensional representation of the input data, and it is randomly sampled from a probability distribution p_z. The purpose of the latent vector is to provide randomness and diversity in the generated data. In GANs, p_z is usually set to a simple distribution such as a Gaussian distribution. The choice of p_z affects the generated data and its diversity. Figure 13 illustrates a typical GAN architecture designed to simulate different infection networks based on the spreading rate λ of an SI model.

The GAN-based approach offers a promising solution to the data scarcity issue during the early stages of a pandemic. As more real-world data becomes available, the GAN model can be fine-tuned and continually adapt to improve its performance, generating even more accurate synthetic graphs representing infection networks. This contributes to a more comprehensive understanding of the pandemic and helps inform more effective response strategies. The iterative fine-tuning process not only enhances the accuracy and relevance of the generated graphs but also demonstrates the potential of GANs to support decision-making in the face of limited data availability. Ultimately, the GAN-based approach can play a crucial role in guiding public health efforts and informing pandemic response strategies, even when faced with the challenges of data scarcity and rapidly evolving situations. Therefore, exploring whether such GAN models can work in the contact tracing domain is a promising area for future research.

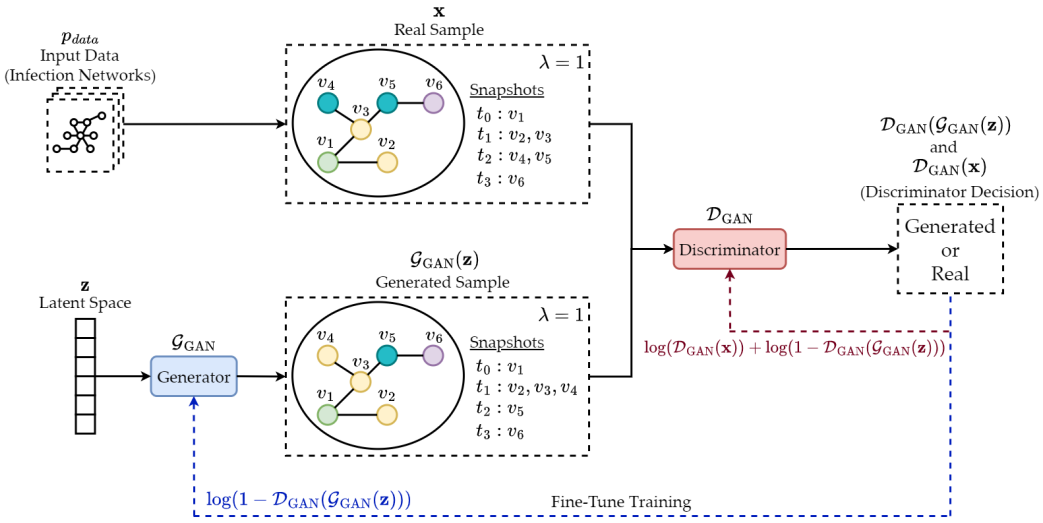


Figure 13. The GAN architecture [161] for generating infection networks based on the spreading rate of a SI model. For each time period t, we can identify which vertex has newly become infected.

7. Case Study

In this section, we present a hypothetical case study that demonstrates the application of machine learning-enabled DCT strategies for pandemic response. We explore how the approaches presented in this review can be integrated to develop an effective DCT system to enhance the ability of public health officials to manage infectious disease outbreaks.

- Background: A new infectious disease outbreak has occurred in a densely populated urban area. Public health authorities are in need of implementing a DCT system that can aid in mitigating the spread of the infection.
- DCT strategy selection and app development: Authorities decide to employ a combination of FCT, BCT, and PCT strategies to maximize the efficacy of the DCT system. An app is developed that incorporates these DCT strategies and leverages machine learning techniques to optimize the system's performance.
- Privacy considerations: To address privacy concerns, the DCT system adopts privacy-preserving techniques, such as data anonymization, encryption, and federated learning with differential privacy. These approaches enable the system to learn from decentralized data sources while preserving individual privacy.
- Data collection and management: The DCT system collects and manages data, including contact tracing data and supplementary information (e.g., demographic factors and health status), to improve the accuracy of machine learning models. Privacy-preserving techniques are employed to protect users' personal information.
- Data availability enhancement: Data scarcity may arise during a pandemic due to factors such as insufficient testing, reporting delays, or inconsistent data collection. To mitigate this issue, advanced data augmentation techniques, such as synthetic data generation, can help improve the available dataset for DCT systems. By combining synthetic data with existing data, more effective learning can occur, leading to improved performance of the DCT system, even in cases where data availability is limited.
- Implementation and evaluation: The DCT app is deployed in the affected area, and its effectiveness is evaluated using various performance metrics, such as infection detection rates, contact tracing efficiency, and false quarantine recommendations. The evaluation process also assesses the system's ability to maintain user privacy.
- Continuous improvement: As the pandemic situation evolves, researchers continue to investigate novel machine learning techniques and strategies to enhance the DCT system's performance. These efforts aim to improve the accuracy, efficiency, and adaptability of the DCT system in response to the changing dynamics of infectious disease outbreaks.

This case study demonstrates the potential of machine learning-enabled DCT strategies in controlling pandemics. By integrating various DCT approaches with advanced machine learning techniques, a comprehensive and effective DCT system can be developed to support public health officials in managing infectious disease outbreaks. This case study emphasizes the importance of continued research and innovation in the field of machine learning-based DCT to prepare for and combat future pandemics.

8. Discussion

In this section, we discuss the limitations of our study and identify potential areas for future exploration in the domain of machine learning-based DCT.

8.1. Limitations

While our review provides valuable insights into the application of machine learning in DCT strategies for pandemic response, it is important to acknowledge some limitations.

- Machine learning techniques: The machine learning techniques emphasized in this study are primarily graph-based learning models, as the focus is on contact tracing network data. This may limit the scope of the review, as other machine learning techniques might be relevant and applicable to DCT systems.
- Experimental evaluations: The proposed approaches in the study lack detailed experimental evaluations to validate their effectiveness and applicability in real-world DCT systems. Future work should include thorough experimental evaluations to ensure the viability of these approaches.
- Pandemic response strategies: Our review mainly focuses on applying machine learning techniques to optimize DCT strategies, which might exclude other important public health interventions and strategies that can be employed during a pandemic.

Despite the mentioned limitations, our study provides a comprehensive analysis and a significant contribution to the field.

8.2. Future Research Directions

In this subsection, we highlight future research directions in machine learning-based DCT, offering promising opportunities for continued exploration and advancement.

- Blockchain technology: The integration of blockchain technology in DCT systems can provide enhanced security, privacy, and trust [172–174]. Blockchain's decentralized and tamper-proof nature could offer a reliable means to store and share contact tracing data while preserving user privacy and ensuring data integrity. Future research could investigate novel approaches to combine blockchain with machine learning techniques for more secure and efficient DCT systems.
- Large language models: Advanced large language models [175], such as ChatGPT, can be leveraged to improve communication and information dissemination in DCT applications [176]. These models can potentially be used to develop user-friendly interfaces, provide personalized risk information, and answer user queries regarding contact tracing or health recommendations [177]. Future work could focus on adapting and fine-tuning these models specifically for DCT applications to enhance their effectiveness and user experience.
- Obfuscation techniques: The incorporation of obfuscation techniques in DCT systems can further enhance security and privacy. Obfuscation methods, such as data perturbation or anonymization, can help protect sensitive user information by adding noise or altering data in a controlled manner. This approach can make it difficult for adversaries to re-identify individuals or infer sensitive information from the shared data. Future research could explore the development of advanced obfuscation techniques in machine learning-based DCT systems, aiming to strike a balance between data utility and privacy protection.
- Adversarial learning methods: Investigating the application of adversarial learning methods in DCT systems can potentially improve the robustness and generalizability

of machine learning models. By training models to withstand adversarial attacks, such as crafted input perturbations designed to mislead the model, they may become more resilient and effective in real-world scenarios. Future research could focus on developing advanced adversarial training techniques tailored to the unique challenges of machine learning-based DCT systems, enhancing their performance and security.

- Cross-disciplinary collaboration: DCT is a complex field that requires expertise from multiple disciplines, such as public health, computer science, and social science. Future research should promote cross-disciplinary collaboration to develop more effective DCT solutions that consider the technical, ethical, and social aspects of the problem.

By exploring future research directions, researchers can contribute to the development of more robust, efficient, and user-friendly machine learning-based DCT systems better prepared to handle potential infectious disease outbreaks in the future.

9. Conclusions

In this paper, we have provided a comprehensive overview of machine learning-based digital contact tracing strategies with privacy-enhancing considerations for pandemic response. We have proposed a novel taxonomy to classify existing DCT strategies into forward contact tracing, backward contact tracing, and proactive contact tracing. We have analyzed and categorized a range of COVID-19-era DCT apps based on their tracing methods and conducted a thorough review of related literature in the field of learning-based DCT. To optimize the DCT strategies, this study has addressed key computational epidemiology problems of DCT and delivered an extensive examination of machine learning techniques to tackle these issues. The study has highlighted the potential advantages of machine learning-based approaches in optimizing DCT, stressing the importance of further research for more robust and user-friendly DCT systems. We have also investigated the data challenges associated with machine learning-based DCT and proposed potential solutions to overcome them. To demonstrate the relevance of our review, we have included a case study. Lastly, we have outlined the study's limitations and suggested promising future research directions in the field of machine learning-based DCT. In fact, before the next pandemic hits, there is much for machines to learn from the COVID-19 pandemic data. Scalable data-driven methodologies will represent a promising step forward to optimize digital contact tracing for future pandemic preparedness.

Author Contributions: Conceptualization, C.-N.H. and C.-W.T.; methodology, C.-N.H., Y.-Z.T., P.-D.Y., J.C. and C.-W.T.; validation, C.-N.H., J.C. and C.-W.T.; supervision, J.C. and C.-W.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Ministry of Science and Technology of Taiwan under Grant 110-2115-M-033-001-MY2, the Ministry of Education, Singapore, under its Academic Research Fund (No. 022307 and AcRF RG91/22), a grant from the NTU World Health Organization Collaborating Centre for Digital Health and Health Education, and the Hong Kong Innovation and Technology Fund (Project No. ITS/188/20).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://dctracing.shinyapps.io/DCTracing/ (accessed on 20 April 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gates, B. Responding to Covid-19—A once-in-a-century pandemic? *N. Engl. J. Med.* **2020**, *382*, 1677–1679. [CrossRef]
2. World Health Organization. WHO Coronavirus (COVID-19) Dashboard. Available online: https://covid19.who.int/ (accessed on 20 April 2023).

3. Chinazzi, M.; Davis, J.T.; Ajelli, M.; Gioannini, C.; Litvinova, M.; Merler, S.; Piontti, A.P.Y.; Mu, K.; Rossi, L.; Sun, K.; et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **2020**, *368*, 395–400. [CrossRef]
4. Flaxman, S.; Mishra, S.; Gandy, A.; Unwin, H.J.T.; Mellan, T.A.; Coupland, H.; Whittaker, C.; Zhu, H.; Berah, T.; Eaton, J.W.; et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **2020**, *584*, 257–261. [CrossRef]
5. Ma, Q.; Liu, Y.Y.; Olshevsky, A. Optimal lockdown for pandemic control. *arXiv* **2020**, arXiv:2010.12923.
6. Fernandes, N. Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy. IESE Business School Working Paper No. WP-1240-E; 2020. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3557504 (accessed on 20 April 2023).
7. Anderson, R.M.; Heesterbeek, H.; Klinkenberg, D.; Hollingsworth, T.D. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *Lancet* **2020**, *395*, 931–934. [CrossRef] [PubMed]
8. Rossi, R.; Socci, V.; Talevi, D.; Mensi, S.; Niolu, C.; Pacitti, F.; Di Marco, A.; Rossi, A.; Siracusano, A.; Di Lorenzo, G. COVID-19 pandemic and lockdown measures impact on mental health among the general population in Italy. *Front. Psychiatry* **2020**, *11*, 790. [CrossRef]
9. Mandel, A.; Veetil, V. The economic cost of COVID lockdowns: An out-of-equilibrium analysis. *Econ. Disasters Clim. Chang.* **2020**, *4*, 431–451. [CrossRef]
10. Firth, J.A.; Hellewell, J.; Klepac, P.; Kissler, S.; Kucharski, A.J.; Spurgin, L.G. Using a real-world network to model localized COVID-19 control strategies. *Nat. Med.* **2020**, *26*, 1616–1622. [CrossRef]
11. Kwok, K.O.; Tang, A.; Wei, V.W.; Park, W.H.; Yeoh, E.K.; Riley, S. Epidemic models of contact tracing: Systematic review of transmission studies of severe acute respiratory syndrome and Middle East respiratory syndrome. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 186–194. [CrossRef] [PubMed]
12. Müller, J.; Kretzschmar, M. Contact tracing—Old models and new challenges. *Infect. Dis. Model.* **2021**, *6*, 222–231. [CrossRef] [PubMed]
13. He, X.; Lau, E.H.; Wu, P.; Deng, X.; Wang, J.; Hao, X.; Lau, Y.C.; Wong, J.Y.; Guan, Y.; Tan, X.; et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **2020**, *26*, 672–675. [CrossRef]
14. Ferretti, L.; Wymant, C.; Kendall, M.; Zhao, L.; Nurtay, A.; Abeler-Dörner, L.; Parker, M.; Bonsall, D.; Fraser, C. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* **2020**, *368*, eabb6936. [CrossRef]
15. Barrat, A.; Cattuto, C.; Kivelä, M.; Lehmann, S.; Saramäki, J. Effect of manual and digital contact tracing on COVID-19 outbreaks: A study on empirical contact data. *J. R. Soc. Interface* **2021**, *18*, 20201000. [CrossRef]
16. Braithwaite, I.; Callender, T.; Bullock, M.; Aldridge, R.W. Automated and partly automated contact tracing: A systematic review to inform the control of COVID-19. *Lancet Digit. Health* **2020**, *2*, e607–e621. [CrossRef] [PubMed]
17. Meister, M.; Kleinberg, J. Optimizing the order of actions in a model of contact tracing. *PNAS Nexus* **2023**, *2*, pgad003. [CrossRef]
18. Landau, S. *People Count: Contact-Tracing Apps and Public Health*; MIT Press: Cambridge, MA, USA, 2021.
19. Rodríguez, P.; Graña, S.; Alvarez-León, E.E.; Battaglini, M.; Darias, F.J.; Hernán, M.A.; López, R.; Llaneza, P.; Martín, M.C.; Ramirez-Rubio, O.; et al. A population-based controlled experiment assessing the epidemiological impact of digital contact tracing. *Nat. Commun.* **2021**, *12*, 587. [CrossRef] [PubMed]
20. Anglemyer, A.; Moore, T.H.; Parker, L.; Chambers, T.; Grady, A.; Chiu, K.; Parry, M.; Wilczynska, M.; Flemyng, E.; Bero, L. Digital contact tracing technologies in epidemics: A rapid review. *Cochrane Database Syst. Rev.* **2020**, *8*, CD013699.
21. Kleinman, R.A.; Merkel, C. Digital contact tracing for COVID-19. *CMAJ* **2020**, *192*, E653–E656. [CrossRef] [PubMed]
22. Trivedi, A.; Vasisht, D. Digital contact tracing: Technologies, shortcomings, and the path forward. *ACM SIGCOMM Comput. Commun. Rev.* **2020**, *50*, 75–81. [CrossRef]
23. Loh, P.S.; Bershteyn, A.; Yee, S.K. Lessons learned in piloting a digital personalized COVID-19 "Radar" on a university campus. *Public Health Rep.* **2022**, *137*, 76S–82S. [CrossRef]
24. Trivedi, A.; Zakaria, C.; Balan, R.; Becker, A.; Corey, G.; Shenoy, P. WiFiTrace: Network-based contact tracing for infectious diseases using passive WiFi sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 1–26. [CrossRef] [PubMed]
25. Wang, S.; Ding, S.; Xiong, L. A new system for surveillance and digital contact tracing for COVID-19: Spatiotemporal reporting over network and GPS. *JMIR mHealth uHealth* **2020**, *8*, e19457.
26. Zhao, Q.; Wen, H.; Lin, Z.; Xuan, D.; Shroff, N. On the accuracy of measured proximity of Bluetooth-based contact tracing apps. In Proceedings of the International Conference on Security and Privacy in Communication Systems, Washington, DC, USA, 21–23 October 2020; pp. 49–60.
27. Hatke, G.F.; Montanari, M.; Appadwedula, S.; Wentz, M.; Meklenburg, J.; Ivers, L.; Watson, J.; Fiore, P. Using Bluetooth Low Energy (BLE) signal strength estimation to facilitate contact tracing for COVID-19. *arXiv* **2020**, arXiv:2006.15711. [CrossRef]
28. Bengio, Y.; Janda, R.; Yu, Y.W.; Ippolito, D.; Jarvie, M.; Pilat, D.; Struck, B.; Krastev, S.; Sharma, A. The need for privacy with public digital contact tracing during the COVID-19 pandemic. *Lancet Digit. Health* **2020**, *2*, e342–e344.
29. Alsdurf, H.; Belliveau, E.; Bengio, Y.; Deleu, T.; Gupta, P.; Ippolito, D.; Janda, R.; Jarvie, M.; Kolody, T.; Krastev, S.; et al. COVI white paper. *arXiv* **2020**, arXiv:2005.08502. [CrossRef]
30. Xu, H.; Zhang, L.; Onireti, O.; Fang, Y.; Buchanan, W. J.; Imran, M. A. BeepTrace: Blockchain-enabled privacy-preserving contact tracing for COVID-19 pandemic and beyond. *IEEE Internet Things J.* **2020**, *8*, 3915–3929.

31. Troncoso, C.; Payer, M.; Hubaux, J.P.; Salathé, M.; Larus, J.; Bugnion, E.; Lueks, W.; Stadler, T.; Pyrgelis, A.; Antonioli, D.; et al. Decentralized privacy-preserving proximity tracing. *arXiv* **2020**, arXiv:2005.12273. [CrossRef]
32. Troncoso, C.; Bogdanov, D.; Bugnion, E.; Chatel, S.; Cremers, C.; Gürses, S.; Hubaux, J.P.; Jackson, D.; Larus, J.R.; Lueks, W.; et al. Deploying decentralized, privacy-preserving proximity tracing. *Commun. ACM* **2022**, *65*, 48–57.
33. Li, J.; Guo, X. COVID-19 contact-tracing apps: A survey on the global deployment and challenges. *arXiv* **2020**, arXiv:2005.03599. [CrossRef]
34. Yu, P.D.; Tan, C.W.; Fu, H.L. Epidemic source detection in contact tracing networks: Epidemic centrality in graphs and message-passing algorithms. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 234–249.
35. Tan, C.W.; Yu, P.D. Contagion source detection in epidemic and infodemic outbreaks: Mathematical analysis and network algorithms. *Found. Trends® Netw.* **2023**, *13*, 107–251. [CrossRef]
36. Fei, Z.; Ryeznik, Y.; Sverdlov, A.; Tan, C.W.; Wong, W.K. An overview of healthcare data analytics with applications to the COVID-19 pandemic. *IEEE Trans. Big Data* **2021**, *8*, 1463–1480.
37. Bengio, Y.; Gupta, P.; Maharaj, T.; Rahaman, N.; Weiss, M.; Deleu, T.; Muller, E.B.; Qu, M.; Schmidt, V.; St-Charles, P.-L.; et al. Predicting infectiousness for proactive contact tracing. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
38. Gupta, P.; Maharaj, T.; Weiss, M.; Rahaman, N.; Alsdurf, H.; Sharma, A.; Minoyan, N.; Harnois-Leblanc, S.; Schmidt, V.; Charles, P.L.S.; et al. COVI-AgentSim: An agent-based model for evaluating methods of digital contact tracing. *arXiv* **2020**, arXiv:2010.16004. [CrossRef]
39. Ojokoh, B.A.; Aribisala, B.; Sarumi, O.A.; Gabriel, A.J.; Omisore, O.; Taiwo, A.E.; Igbe, T.; Chukwuocha, U.M.; Yusuf, T.; Afolayan, A.; et al. Contact tracing strategies for COVID-19 prevention and containment: A scoping review. *Big Data Cogn. Comput.* **2022**, *6*, 111. [PubMed]
40. Mondal, M.R.H.; Bharati, S.; Podder, P. Diagnosis of COVID-19 using machine learning and deep learning: A review. *Curr. Med Imaging* **2021**, *17*, 1403–1418. [CrossRef] [PubMed]
41. Agbehadji, I.E.; Awuzie, B.O.; Ngowi, A.B.; Millham, R.C. Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5330. [CrossRef]
42. Mbunge, E. Integrating emerging technologies into COVID-19 contact tracing: Opportunities, challenges and pitfalls. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2020**, *14*, 1631–1636. [CrossRef] [PubMed]
43. Lalmuanawma, S.; Hussain, J.; Chhakchhuak, L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* **2020**, *139*, 110059. [CrossRef]
44. Altmann, S.; Milsom, L.; Zillessen, H.; Blasone, R.; Gerdon, F.; Bach, R.; Kreuter, F.; Nosenzo, D.; Toussaert, S.; Abeler, J. Acceptability of app-based contact tracing for COVID-19: Cross-country survey study. *JMIR mHealth uHealth* **2020**, *8*, e19857. [CrossRef]
45. Ahmed, N.; Michelin, R.A.; Xue, W.; Ruj, S.; Malaney, R.; Kanhere, S.S.; Seneviratne, A.; Hu, W.; Janicke, H.; Jha, S.K. A survey of COVID-19 contact tracing apps. *IEEE Access* **2020**, *8*, 134577–134601. [CrossRef]
46. Alanzi, T. A review of mobile applications available in the App and Google Play stores used during the COVID-19 outbreak. *J. Multidiscip. Healthc.* **2021**, *14*, 45–57. [CrossRef] [PubMed]
47. Allen, W.E.; Altae-Tran, H.; Briggs, J.; Jin, X.; McGee, G.; Shi, A.; Raghavan, R.; Kamariza, M.; Nova, N.; Pereta, A.; et al. Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. *Nat. Hum. Behav.* **2020**, *4*, 972–982.
48. NHSx. Risk-Scoring Algorithm (Interim): Technical Information. 2020. Available online: https://www.gov.uk/government/collections/nhs-covid-19-app (accessed on 20 April 2023).
49. Loh, P.S. Flipping the perspective in contact tracing. *arXiv* **2020**, arXiv:2010.03806. [CrossRef] [PubMed]
50. Freifeld, C.C.; Chunara, R.; Mekaru, S.R.; Chan, E.H.; Kass-Hout, T.; Ayala Iacucci, A.; Brownstein, J.S. Participatory epidemiology: Use of mobile phones for community-based health reporting. *PLoS Med.* **2010**, *7*, e1000376. [CrossRef]
51. Hellewell, J.; Abbott, S.; Gimma, A.; Bosse, N.I.; Jarvis, C.I.; Russell, T.W.; Munday, J.D.; Kucharski, A.J.; Edmunds, W.J.; Sun, F.; et al. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. *Lancet Glob. Health* **2020**, *8*, e488–e496. [CrossRef]
52. Aleta, A.; Martin-Corral, D.; y Piontti, A.P.; Ajelli, M.; Litvinova, M.; Chinazzi, M.; Dean, N.E.; Halloran, M.E.; Longini, I.M., Jr.; Merler, S.; et al. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nat. Hum. Behav.* **2020**, *4*, 964–971. [CrossRef]
53. Hinch, R.; Probert, W.J.M.; Nurtay, A.; Kendall, M.; Wymant, C.; Hall, M.; Lythgoe, K.; Bulas Cruz, A.; Zhao, L.; Stewart, A.; et al. OpenABM-Covid19—An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLoS Comput. Biol.* **2021**, *17*, 1–26. [CrossRef]
54. Grantz, K.H.; Lee, E.C.; D'Agostino McGowan, L.; Lee, K.H.; Metcalf, C.J.E.; Gurley, E.S.; Lessler, J. Maximizing and evaluating the impact of test-trace-isolate programs: A modeling study. *PLoS Med.* **2021**, *18*, 1–16.
55. Tan, C.W.; Yu, P.D.; Chen, S.; Poor, H.V. DeepTrace: Learning to optimize contact tracing in epidemic networks with Graph Neural Networks. *arXiv* **2022**, arXiv:2211.00880. [CrossRef]
56. Endo, A.; Leclerc, Q.J.; Knight, G.M.; Medley, G.F.; Atkins, K.E.; Funk, S.; Kucharski, A.J. Implication of backward contact tracing in the presence of overdispersed transmission in COVID-19 outbreaks. *Wellcome Open Res.* **2020**, *5*, 239. [CrossRef]

57. Müller, J.; Kretzschmar, M. Forward thinking on backward tracing. *Nat. Phys.* **2021**, *17*, 555–556. [CrossRef]
58. Kojaku, S.; Hébert-Dufresne, L.; Mones, E.; Lehmann, S.; Ahn, Y.Y. The effectiveness of backward contact tracing in networks. *Nat. Phys.* **2021**, *17*, 652–658. [CrossRef]
59. Raymenants, J.; Geenen, C.; Thibaut, J.; Nelissen, K.; Gorissen, S.; Andre, E. Empirical evidence on the efficiency of backward contact tracing in COVID-19. *Nat. Commun.* **2022**, *13*, 4750. [CrossRef]
60. Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature* **2009**, *457*, 1012–1014. [CrossRef] [PubMed]
61. Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; De Domenico, M. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nat. Hum. Behav.* **2020**, *4*, 1285–1293.
62. Briers, M.; Charalambides, M.; Holmes, C. Risk scoring calculation for the current NHSx contact tracing app. *arXiv* **2020**, arXiv:2005.11057.
63. Herbrich, R.; Rastogi, R.; Vollgraf, R. CRISP: A probabilistic model for individual-level COVID-19 infection risk estimation based on contact data. *arXiv* **2020**, arXiv:2006.04942. [CrossRef]
64. Leung, K.; Wu, J.T.; Leung, G.M. Real-time tracking and prediction of COVID-19 infection using digital proxies of population mobility and mixing. *Nat. Commun.* **2021**, *12*, 1501. [CrossRef]
65. Baker, A.; Biazzo, I.; Braunstein, A.; Catania, G.; Dall'Asta, L.; Ingrosso, A.; Krzakala, F.; Mazza, F.; Mézard, M.; Muntoni, A.P.; et al. Epidemic mitigation by statistical inference from contact tracing data. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2106548118.
66. Murphy, K.; Kumar, A.; Serghiou, S. Risk score learning for COVID-19 contact tracing apps. In Proceedings of the 6th Machine Learning for Healthcare Conference, Virtual, 6–7 August 2021; pp. 373–390. [CrossRef]
67. Fenton, N.E.; McLachlan, S.; Lucas, P.; Dube, K.; Hitman, G.A.; Osman, M.; Kyrimi, E.; Neil, M. A Bayesian network model for personalised COVID19 risk assessment and contact tracing. *medRxiv* **2021**. [CrossRef]
68. Lorch, L.; Kremer, H.; Trouleau, W.; Tsirtsis, S.; Szanto, A.; Schölkopf, B.; Gomez-Rodriguez, M. Quantifying the effects of contact tracing, testing, and containment measures in the presence of infection hotspots. *ACM Trans. Spat. Algorithms Syst.* **2022**, *8*, 1–28.
69. Rivest, R.; Schiefelbein, M.C.; Zissman, M.A.; Bay, J.; Bugnion, E.; Finnerty, J.; Liccardi, I.; Nelson, B.; Norige, A.S.; Shen, E.H.; et al. *Automated Exposure Notification for COVID-19*; Lincoln Laboratory Technical Report; TR-1288; MIT: Cambridge, MA, USA, 2023. [CrossRef] [PubMed]
70. Gupta, P.; Maharaj, T.; Weiss, M.; Rahaman, N.; Alsdurf, H.; Minoyan, N.; Harnois-Leblanc, S.; Merckx, J.; Williams, A.; Schmidt, V.; et al. Proactive contact tracing. *PLoS Digit. Health* **2023**, *2*, e0000199. [CrossRef]
71. Feng, T.; Song, S.; Xia, T.; Li, Y. Contact tracing and epidemic intervention via deep reinforcement learning. *ACM Trans. Knowl. Discov. Data* **2023**, *17*, 1–24.
72. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. [CrossRef]
73. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Int. J. Surg.* **2021**, *88*, 105906.
74. Kotaru, M.; Joshi, K.; Bharadia, D.; Katti, S. SpotFi: Decimeter level localization using WiFi. In Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, London, UK, 17–21 August 2015; pp. 269–282.
75. Margolies, R.; Becker, R.; Byers, S.; Deb, S.; Jana, R.; Urbanek, S.; Volinsky, C. Can you find me now? Evaluation of network-based localization in a 4G LTE network. In Proceedings of the IEEE INFOCOM 2017—IEEE Conference on Computer Communications, Atlanta, GA, USA, 1–4 May 2017; pp. 1–9.
76. Tsai, Y.Z.; Chen, J. Network-side 5G mmWave channel signatures for pandemic contact tracing. In Proceedings of the ICC 2022—IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022; pp. 3598–3603.
77. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
78. Ye, J.; Chow, J.H.; Chen, J.; Zheng, Z. Stochastic gradient boosted distributed decision trees. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 2061–2064.
79. Coadou, Y. Boosted decision trees and applications. In Proceedings of the EPJ Web of Conferences, Autrans, France, 28 May–2 June 2012. [CrossRef]
80. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377.
81. O'Shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
82. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [CrossRef]
83. Shah, D.; Zaman, T. Rumors in a network: Who's the culprit? *IEEE Trans. Inf. Theory* **2011**, *57*, 5163–5181.
84. Gomez-Rodriguez, M.; Leskovec, J.; Krause, A. Inferring networks of diffusion and influence. *ACM Trans. Knowl. Discov. Data* **2012**, *5*, 1–37. [CrossRef]
85. Yu, P.D.; Tan, C.W.; Fu, H.L. Averting cascading failures in networked infrastructures: Poset-constrained graph algorithms. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 733–748.

86. Shah, D.; Zaman, T. Detecting sources of computer viruses in networks: Theory and experiment. In Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, NY, USA, 14–18 June 2010; pp. 203–214. [CrossRef]
87. Zhu, K.; Ying, L. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Trans. Netw.* **2016**, *24*, 408–421.
88. Hamilton, W.L.; Ying, Z.; Leskovec, J. Inductive representation learning on large graphs. In Proceedings of the International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1025–1035.
89. Zheng, L.; Tan, C.W. A probabilistic characterization of the rumor graph boundary in rumor source detection. In Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; pp. 765–769. [CrossRef] [PubMed]
90. Eagle, N.; Pentland, A.; Lazer, D. Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15274–15278.
91. Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R.R.; Smola, A.J. Deep Sets. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3391–3401.
92. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set Transformer: A framework for attention-based permutation-invariant neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 3744–3753.
93. Edwards, H.; Storkey, A. Towards a neural statistician. *arXiv* **2016**, arXiv:1606.02185. [CrossRef]
94. Xu, D. Modeling of network based digital contact tracing and testing strategies, including the pre-exposure notification system, for the COVID-19 pandemic. *Math. Biosci.* **2021**, *338*, 108645. [CrossRef]
95. Zhang, B.; Wang, R.; Xu, H.; Zhang, X.; Zhang, L. DISTERNING: Distance estimation using machine learning approach for COVID-19 contact tracing and beyond. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 3207–3223. [CrossRef]
96. Yi, F.; Xie, Y.; Jamieson, K. Cellular-Assisted, Deep learning based COVID-19 contact tracing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 1–27.
97. Veličković, P.; Fedus, W.; Hamilton, W.L.; Liò, P.; Bengio, Y.; Hjelm, R.D. Deep Graph Infomax. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019. [CrossRef]
98. Wong, A.K.; You, M. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1985**, *PAMI-7*, 599–609. [CrossRef]
99. Chen, H.; Friedman, J.H. A new graph-based two-sample test for multivariate and object data. *J. Am. Stat. Assoc.* **2017**, *112*, 397–409.
100. Lopez-Paz, D.; Oquab, M. Revisiting classifier two-sample tests. *arXiv* **2016**, arXiv:1610.06545. [CrossRef]
101. Liu, X.; Fu, L.; Wang, X.; Zhou, C. On the similarity between von Neumann graph entropy and structural information: Interpretation, computation, and applications. *IEEE Trans. Inf. Theory* **2022**, *68*, 2182–2202. [CrossRef]
102. Wu, X.; Liu, Z. How community structure influences epidemic spread in social networks. *Phys. A Stat. Mech. Its Appl.* **2008**, *387*, 623–630.
103. Kuo, C.Y.; Hang, C.N.; Yu, P.D.; Tan, C.W. Parallel counting of triangles in large graphs: Pruning and hierarchical clustering algorithms. In Proceedings of the 2018 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 25–27 September 2018; pp. 1–6.
104. Preciado, V.M.; Zargham, M.; Enyioha, C.; Jadbabaie, A.; Pappas, G. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In Proceedings of the 52nd IEEE Conference on Decision and Control, Firenze, Italy, 10–13 December 2013; pp. 7486–7491. [CrossRef]
105. Jalili, M.; Perc, M. Information cascades in complex networks. *J. Complex Netw.* **2017**, *5*, 665–693. [CrossRef]
106. Eubank, S.; Guclu, H.; Anil Kumar, V.; Marathe, M.V.; Srinivasan, A.; Toroczkai, Z.; Wang, N. Modelling disease outbreaks in realistic urban social networks. *Nature* **2004**, *429*, 180–184. [CrossRef]
107. Newman, M.E. Spread of epidemic disease on networks. *Phys. Rev. E* **2002**, *66*, 016128.
108. Bhapkar, H.; Mahalle, P.N.; Dhotre, P.S. Virus graph and COVID-19 pandemic: A graph theory approach. In *Big Data Analytics and Artificial Intelligence against COVID-19: Innovation Vision and Approach*; Springer: Cham, Switzerland, 2020; pp. 15–34.
109. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford, CA, USA, 1999.
110. Negahban, S.; Oh, S.; Shah, D. Iterative ranking from pair-wise comparisons. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2474–2482.
111. Karger, D.; Oh, S.; Shah, D. Iterative learning for reliable crowdsourcing systems. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1953–1961. [CrossRef] [PubMed]
112. Sun, K.; Chen, J.; Viboud, C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study. *Lancet Digit. Health* **2020**, *2*, e201–e208. [CrossRef]
113. He, X.; Gao, M.; Kan, M.Y.; Wang, D. BiRank: Towards ranking on bipartite graphs. *IEEE Trans. Knowl. Data Eng.* **2016**, *29*, 57–71. [CrossRef] [PubMed]
114. Yang, K.C.; Aronson, B.; Ahn, Y.Y. BiRank: Fast and flexible ranking on bipartite networks with R and Python. *J. Open Source Softw.* **2020**, *5*, 2315. [CrossRef]
115. Chung, F.; Horn, P.; Tsiatas, A. Distributing antidote using PageRank vectors. *Internet Math.* **2009**, *6*, 237–254.

116. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef] [PubMed]
117. Haug, N.; Geyrhofer, L.; Londei, A.; Dervic, E.; Desvars-Larrive, A.; Loreto, V.; Pinior, B.; Thurner, S.; Klimek, P. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat. Hum. Behav.* **2020**, *4*, 1303–1312.
118. Tang, S.; Hu, X.; Atlas, L.; Khanzada, A.; Pilanci, M. Hierarchical multi-modal transformer for automatic detection of COVID-19. In Proceedings of the 2022 5th International Conference on Signal Processing and Machine Learning, Dalian, China, 4–6 August 2022; pp. 197–202. [CrossRef]
119. Ahmad, K.; Alam, F.; Qadir, J.; Qolomany, B.; Khan, I.; Khan, T.; Suleman, M.; Said, N.; Hassan, S.Z.; Gul, A.; et al. Global user-level perception of COVID-19 contact tracing applications: Data-driven approach using natural language processing. *JMIR Form. Res.* **2022**, *6*, e36238.
120. Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.Y. Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28877–28888.
121. Zhang, J.; Zhang, H.; Xia, C.; Sun, L. GRAPH-BERT: Only attention is needed for learning graph representations. *arXiv* **2020**, arXiv:2001.05140.
122. Dwivedi, V.P.; Bresson, X. A generalization of transformer networks to graphs. *arXiv* **2020**, arXiv:2012.09699.
123. Yun, S.; Jeong, M.; Kim, R.; Kang, J.; Kim, H.J. Graph transformer networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 11983–11993.
124. Rampášek, L.; Galkin, M.; Dwivedi, V.P.; Luu, A.T.; Wolf, G.; Beaini, D. Recipe for a general, powerful, scalable graph transformer. *arXiv* **2022**, arXiv:2205.12454.
125. Joshi, C. Transformers are graph neural networks. *The Gradient* 2020. Available online: https://graphdeeplearning.github.io/post/transformers-are-gnns/ (accessed on 20 April 2023). [CrossRef]
126. Xu, B.; Gutierrez, B.; Mekaru, S.; Sewalk, K.; Goodwin, L.; Loskill, A.; Cohn, E.L.; Hswen, Y.; Hill, S.C.; Cobo, M.M.; et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **2020**, *7*, 106. [CrossRef]
127. Adam, D.C.; Wu, P.; Wong, J.Y.; Lau, E.H.; Tsang, T.K.; Cauchemez, S.; Leung, G.M.; Cowling, B.J. Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* **2020**, *26*, 1714–1719. [CrossRef] [PubMed]
128. Serafino, M.; Monteiro, H.S.; Luo, S.; Reis, S.D.; Igual, C.; Lima Neto, A.S.; Travizano, M.; Andrade, J.S., Jr.; Makse, H.A. Digital contact tracing and network theory to stop the spread of COVID-19 using big-data on human mobility geolocalization. *PLoS Comput. Biol.* **2022**, *18*, e1009865.
129. Moosa, J.; Awad, W.; Kalganova, T. COVID-19 contact-tracing networks datasets. In Proceedings of the 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 8–9 March 2023; pp. 1–4.
130. Konečnỳ, J.; McMahan, H.B.; Yu, F.X.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv* **2016**, arXiv:1610.05492. [CrossRef]
131. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* **2019**, *10*, 1–19. [CrossRef]
132. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [CrossRef]
133. Nguyen, H.T.; Sehwag, V.; Hosseinalipour, S.; Brinton, C.G.; Chiang, M.; Poor, H.V. Fast-convergent federated learning. *IEEE J. Sel. Areas Commun.* **2020**, *39*, 201–218. [CrossRef]
134. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends® Mach. Learn.* **2021**, *14*, 1–210.
135. Dwork, C. Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming (ICALP), Venice, Italy, 10–14 July 2006; pp. 1–12.
136. Dwork, C. Differential privacy: A survey of results. In Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC), Xi'an, China, 25–29 April 2008; pp. 1–19. [CrossRef]
137. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends® Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
138. Hsu, H.; Martinez, N.; Bertran, M.; Sapiro, G.; Calmon, F.P. A survey on statistical, information, and estimation—Theoretic views on privacy. *IEEE BITS Inf. Theory Mag.* **2021**, *1*, 45–56.
139. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318. [CrossRef]
140. Sarwate, A.D.; Chaudhuri, K. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Process. Mag.* **2013**, *30*, 86–94.
141. Ji, Z.; Lipton, Z.C.; Elkan, C. Differential privacy and machine learning: A survey and review. *arXiv* **2014**, arXiv:1412.7584. [CrossRef]
142. Zhu, T.; Ye, D.; Wang, W.; Zhou, W.; Philip, S.Y. More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 2824–2843. [CrossRef]
143. Blanco-Justicia, A.; Sánchez, D.; Domingo-Ferrer, J.; Muralidhar, K. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Comput. Surv.* **2022**, *55*, 1–16.

144. Bun, M.; Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Proceedings of the 14th Theory of Cryptography Conference (TCC), Beijing, China, 31 October–3 November 2016; pp. 635–658.
145. Kairouz, P.; McMahan, B.; Song, S.; Thakkar, O.; Thakurta, A.; Xu, Z. Practical and private (deep) learning without sampling or shuffling. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 5213–5225.
146. Konečnỳ, J.; McMahan, H.B.; Ramage, D.; Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv* **2016**, arXiv:1610.02527.
147. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282. [CrossRef]
148. Arachchige, P.C.M.; Bertok, P.; Khalil, I.; Liu, D.; Camtepe, S.; Atiquzzaman, M. Local differential privacy for deep learning. *IEEE Internet Things J.* **2019**, *7*, 5827–5842. [CrossRef]
149. Bonawitz, K.; Kairouz, P.; McMahan, B.; Ramage, D. Federated learning and privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. *Queue* **2021**, *19*, 87–114.
150. McMahan, B.; Thakurta, A. Federated Learning with Formal Differential Privacy Guarantees. Available online: https://ai.googleblog.com/2022/02/federated-learning-with-formal.html (accessed on 20 April 2023). [CrossRef]
151. Wu, C.; Wu, F.; Lyu, L.; Qi, T.; Huang, Y.; Xie, X. A federated graph neural network framework for privacy-preserving personalization. *Nat. Commun.* **2022**, *13*, 3091.
152. Truex, S.; Liu, L.; Chow, K.H.; Gursoy, M.E.; Wei, W. LDP-Fed: Federated learning with local differential privacy. In Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking, Heraklion, Greece, 27 April 2020; pp. 61–66.
153. Balle, B.; Kairouz, P.; McMahan, B.; Thakkar, O.; Guha Thakurta, A. Privacy amplification via random check-ins. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4623–4634.
154. McMahan, H.B.; Ramage, D.; Talwar, K.; Zhang, L. Learning differentially private recurrent language models. *arXiv* **2017**, arXiv:1710.06963.
155. Ramaswamy, S.; Thakkar, O.; Mathews, R.; Andrew, G.; McMahan, H.B.; Beaufays, F. Training production language models without memorizing user data. *arXiv* **2020**, arXiv:2009.10031.
156. Thakkar, O.; Ramaswamy, S.; Mathews, R.; Beaufays, F. Understanding unintended memorization in federated learning. *arXiv* **2020**, arXiv:2006.07490.
157. Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečný, J.; Mazzocchi, S.; McMahan, B.; et al. Towards federated learning at scale: System design. *Proc. Mach. Learn. Syst.* **2019**, *1*, 374–388.
158. Knott, B.; Venkataraman, S.; Hannun, A.; Sengupta, S.; Ibrahim, M.; van der Maaten, L. Crypten: Secure multi-party computation meets machine learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4961–4973.
159. Kanagavelu, R.; Li, Z.; Samsudin, J.; Yang, Y.; Yang, F.; Goh, R.S.M.; Cheah, M.; Wiwatphonthana, P.; Akkarajitsakul, K.; Wang, S. Two-phase multi-party computation enabled privacy-preserving federated learning. In Proceedings of the 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), Melbourne, Australia, 11–14 May 2020; pp. 410–419.
160. Zhang, C.; Li, S.; Xia, J.; Wang, W.; Yan, F.; Liu, Y. BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning. In Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC 2020), Virtual, 15–17 July 2020. [CrossRef]
161. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
162. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2234–2242.
163. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 7354–7363.
164. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
165. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
166. Wang, Y. A mathematical introduction to generative adversarial nets (GAN). *arXiv* **2020**, arXiv:2009.00169.
167. Tao, C.; Chen, L.; Henao, R.; Feng, J.; Duke, L.C. Chi-square generative adversarial network. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 4887–4896.
168. Ho, J.; Ermon, S. Generative adversarial imitation learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4565–4573.
169. Guo, X.; Hong, J.; Lin, T.; Yang, N. Relaxed Wasserstein with applications to GANs. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3325–3329.
170. Yoon, J.; Jarrett, D.; Van der Schaar, M. Time-series generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5508–5518.
171. Bojchevski, A.; Shchur, O.; Zügner, D.; Günnemann, S. NetGAN: Generating graphs via random walks. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 610–619. [CrossRef]

172. Klaine, P.V.; Zhang, L.; Zhou, B.; Sun, Y.; Xu, H.; Imran, M. Privacy-preserving contact tracing and public risk assessment using blockchain for COVID-19 pandemic. *IEEE Internet Things Mag.* **2020**, *3*, 58–63.
173. Peng, Z.; Xu, C.; Wang, H.; Huang, J.; Xu, J.; Chu, X. P2B-Trace: Privacy-preserving blockchain-based contact tracing to combat pandemics. In Proceedings of the 2021 International Conference on Management of Data, Virtual, 20–25 June 2021; pp. 2389–2393. [CrossRef] [PubMed]
174. Idrees, S.M.; Nowostawski, M.; Jameel, R. Blockchain-based digital contact tracing apps for COVID-19 pandemic management: Issues, challenges, solutions, and future directions. *JMIR Med. Inform.* **2021**, *9*, e25245. [CrossRef] [PubMed]
175. Sallam, M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
176. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198.
177. Oniani, D.; Wang, Y. A qualitative evaluation of language models on automatic question-answering for COVID-19. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Virtual, 21–24 September 2020; pp. 1–9.