

Primary and Secondary Contact Tracing Using Machine Learning and Network Analysis

Sneha D P¹, Dr. Vasudev T²

¹Research Scholar, Department of Computer Science, University of Mysore, Mysuru, Karnataka, India

Email: sneha.p012@gmail.com

²Professor, Department of MCA, Maharaja Institute of Technology Mysore, Karnataka, India

Email: vasu@mitmysore.in

Received: 25th Aug, 2026 | Revised: 1st Sep, 2026 | Accepted: 5th Sep, 2026 | Available Online: 15th Sep, 2026

Abstract

Contact tracing is an important process for identifying individuals who may have been exposed to an infectious disease through direct or indirect interactions with an infected person. Traditional contact-tracing approaches primarily focus on identifying primary contacts, while secondary contacts may also play a significant role in the potential spread of infectious diseases. This paper proposes a computational framework for identifying and analyzing **primary and secondary contacts using contact network analysis and machine learning techniques**. In the proposed approach, individuals are represented as nodes in a contact network, and interactions between individuals are represented as edges. Primary contacts are identified based on direct interactions with an infected individual, whereas secondary contacts are determined through indirect interactions with the primary contacts. The framework incorporates contact duration, interaction frequency, proximity, and temporal information to estimate the relative risk associated with each contact. A machine-learning-based risk classification model can further categorize individuals into low-, medium-, and high-risk groups. The proposed system enables systematic identification and prioritization of potentially exposed individuals and provides a scalable approach for analyzing complex contact networks. Experimental evaluation can be performed using precision, recall, F1-score, accuracy, and network-based measures. The proposed framework can support public-health decision-making by improving the efficiency of contact identification and risk prioritization while maintaining appropriate privacy and data-protection considerations.

Keywords: Contact Tracing, Primary Contact, Secondary Contact, Machine Learning, Network Analysis, Risk Classification, Graph Theory, Public Health.

How to cite this article: Sneha D P¹ Dr. Vasudev T². Primary and Secondary Contact Tracing Using Machine Learning and Network Analysis. Int J Drug Deliv Technol. 2026;16(83s):498-513. DOI: 10.25258/ijddt.16.83s.63.

1. Introduction

Infectious diseases continue to pose significant challenges to public health systems because transmission can occur through complex interactions among individuals, communities, workplaces, educational institutions, and other social environments. The rapid identification of potentially exposed individuals is therefore essential for interrupting chains of transmission and supporting timely public-health interventions. **Contact tracing** is a systematic process used to identify, assess, manage, and support individuals who may have been exposed to an infectious person. Its primary objectives include preventing further infections, enabling early identification of cases, and improving understanding of disease-transmission patterns. ([World Health Organization](https://www.who.int))

Traditional contact tracing commonly involves identifying individuals who have had direct interactions with a confirmed or suspected case. However, disease transmission frequently occurs through multiple levels of social interaction, making it important to understand both **primary and secondary contacts**. In the context of the proposed computational framework, a primary contact refers to an individual directly connected to an index case through a recorded exposure or interaction, whereas a secondary contact refers to an individual connected indirectly through a primary contact. The identification of these multiple contact levels can help public-health authorities understand possible transmission pathways and prioritize follow-up activities. Importantly, actual definitions of exposure and risk must remain disease- and context-specific, considering factors such as transmission

mode, duration, distance, type of interaction, and timing of exposure. ([NCBI](#))

Conventional manual contact tracing relies heavily on interviews, questionnaires, telephone communication, and public-health personnel. Although such approaches can be effective, they may become difficult to scale during large outbreaks involving extensive social networks. Manual procedures can also be affected by incomplete recall of interactions, delays in identifying contacts, and the difficulty of analyzing complex relationships among a large number of individuals. Digital technologies have therefore emerged as useful tools for supporting the collection and analysis of contact information. The World Health Organization also recognizes that contact identification may involve interviews, electronic surveys, and digital proximity-tracing tools, while emphasizing that privacy, confidentiality, and appropriate use of collected information are essential. ([NCBI](#))

The increasing availability of digital interaction data has created opportunities to model contact relationships using **graph and network analysis**. In a contact network, individuals can be represented as nodes, while interactions or potential exposure events can be represented as edges. This representation enables direct contacts to be identified as first-level neighbors of an index case and indirect or secondary contacts to be identified through additional network connections. Graph-based approaches have been used in digital contact tracing to analyze complex relational data and identify contacts and high-risk locations more efficiently. Research has also explored graph-based machine learning and visualization techniques for identifying clusters and potential missing links within infection-contact networks. ([ScienceDirect](#))

Despite the advantages of digital contact tracing, simply identifying a connection between two individuals may not provide sufficient information for determining the relative importance of that contact. The likelihood of relevant exposure can depend on multiple factors, including **contact duration, interaction frequency, physical proximity, timing, and the characteristics of the interaction**. Therefore, an effective computational system should not only identify primary and secondary contacts but also provide a mechanism for analyzing and prioritizing contacts based on available evidence and disease-specific risk criteria. Current WHO guidance similarly emphasizes that contact definitions and risk assessments should consider multiple contextual factors rather than relying on a single universal interaction threshold. ([NCBI](#))

Machine learning provides an additional opportunity to support contact-network analysis by learning relationships between interaction features and predefined risk categories or outcomes. Features extracted from contact records, such as duration, frequency, proximity, temporal information, and network characteristics, can be used to develop computational models for **risk classification and contact prioritization**. When combined with graph-based identification, such models can support a structured workflow in which direct and indirect contacts are first identified and subsequently analyzed according to relevant exposure features. Recent research has also explored graph-based learning approaches for optimizing contact tracing in epidemic networks, demonstrating the potential of combining network structure with data-driven methods. ([Princeton University](#))

Motivated by these challenges, this paper proposes a **graph-based framework for primary and secondary contact tracing with machine-learning-based risk prioritization**. The proposed approach represents individuals and their interactions as a contact network. An index case is used as the starting point for identifying primary contacts through direct network connections, followed by the identification of secondary contacts through indirect connections associated with the primary-contact group. Relevant interaction and network features are subsequently extracted to calculate or predict a risk category, enabling contacts to be prioritized for further assessment according to the applicable public-health criteria.

The major objective of the proposed work is to provide a systematic computational framework that integrates **contact-network construction, multi-level contact identification, feature extraction, and risk-based classification**. Unlike approaches that focus only on direct contact identification, the proposed framework explicitly models both primary and secondary levels of contact to provide a broader representation of potential interaction pathways. The system can be evaluated using measures such as accuracy, precision, recall, F1-score, and appropriate network-based performance indicators. However, the proposed framework is intended as a **decision-support and analytical tool**, and its outputs should be interpreted alongside disease-specific guidance and professional public-health assessment.

The remainder of the paper is organized as follows. Section 2 presents the related work on traditional, digital, and graph-based contact tracing approaches. Section 3 describes the proposed methodology for primary and secondary contact identification and risk prioritization. Section 4 presents the experimental evaluation and comparative analysis,

followed by a discussion of the findings and limitations. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related Work

Contact tracing has evolved from conventional manual investigation to digital proximity tracing, temporal contact-network analysis, machine-learning-based prediction, and graph neural network approaches. Early research established the importance of identifying direct contacts of infected individuals, while subsequent studies demonstrated that digital technologies can improve the speed and scale of contact identification. More recent studies have investigated network structure, second- and third-degree contacts, privacy-preserving architectures, and graph-based learning. The following studies are particularly relevant to the proposed primary and secondary contact-tracing framework.

Ferretti et al. [1] investigated the transmission characteristics of SARS-CoV-2 and evaluated the feasibility of instantaneous digital contact tracing. Their mathematical analysis demonstrated that conventional manual tracing could be too slow for rapidly spreading infections, whereas smartphone-based tracing could provide immediate notification of potentially exposed individuals. The study established an important foundation for digital contact tracing and demonstrated the value of rapid identification and notification.

Keeling et al. [2] developed predictive models of contact tracing using social-contact information from more than 5,800 respondents. Their study examined the effectiveness of contact tracing and the number of individuals who would need to be traced for effective disease control. The results showed that contact tracing can substantially reduce transmission but can impose a considerable operational burden.

Grekousis and Liu [3] conducted a systematic review of digital contact tracing and proximity-awareness technologies. They reviewed 580 publications and included 19 studies in their qualitative synthesis. Their work identified digital contact tracing as a potentially rapid mechanism for collecting and analyzing proximity information while also highlighting privacy, community adoption, and effectiveness as important challenges.

Anglemyer et al. [4] performed a Cochrane rapid review of digital contact-tracing technologies. The review considered outbreak-response systems, proximity-tracing systems, and symptom-tracking technologies. The authors concluded that digital

approaches may improve the speed and reliability of contact identification, although evidence from real-world outbreak settings was limited. The study also emphasized privacy, accessibility, equity, and integration with manual tracing.

Huang et al. [5] evaluated digital contact-tracing tools in Singapore, comparing the Bluetooth-based TraceTogether application with a wearable real-time locating system. Their study demonstrated the practical use of digital proximity information for supporting contact tracing in a clinical environment. The results indicated that digital systems could complement conventional contact-tracing processes, although technical limitations and differences between technologies affected contact detection.

Cencetti et al. [6] proposed a modeling framework based on empirical high-resolution contact networks to evaluate digital proximity tracing. Their work considered realistic heterogeneity in human interactions and analyzed the effects of contact duration and physical distance. The study showed that contact-tracing effectiveness depends strongly on the structure of real-world contact networks and intervention delays. Interestingly, the authors found that tracing second-order contacts could involve significant privacy and intervention costs, demonstrating that simply expanding the tracing radius does not necessarily improve the overall outcome.

Xu [7] developed a network-based mathematical model for digital contact tracing and testing. The study used a high-resolution contact dataset and extended it into weighted temporal social networks. Importantly for the present research, the study investigated strategies involving **second- and third-degree contacts** in addition to direct contacts. The results indicated that testing and tracing beyond first-degree contacts can improve modeled epidemic control under certain conditions.

Rodríguez et al. [8] conducted a population-based controlled experiment to evaluate the Spanish Radar COVID digital contact-tracing application in La Gomera. Approximately one-third of the population adopted the technology during the experiment, and the application detected approximately 6.3 close contacts per simulated primary infection. This study provided empirical evidence that digital contact tracing can identify contacts, including contacts with people who may not otherwise be known to the index case.

Wymant et al. [9] investigated the effectiveness of a national digital contact-tracing system using real-world data from the United Kingdom. Their work demonstrated that digital tracing could identify

contacts rapidly and contribute to epidemic control when combined with other interventions. The study also highlighted the importance of user adoption and timely notification.

Kretzschmar et al. [10] analyzed the effect of delays in case isolation and contact tracing using mathematical modeling. Their work showed that rapid contact identification and notification are particularly important because delays can substantially reduce the effectiveness of tracing interventions. This supports the need for computational approaches capable of rapidly identifying multiple levels of contacts.

Braithwaite et al. [11] conducted a systematic review of digital contact-tracing technologies and evaluated the effectiveness of different technological approaches. Their analysis emphasized that digital tracing should be considered as a complement to conventional contact tracing rather than a complete replacement. The study also highlighted challenges associated with privacy, technology adoption, and heterogeneous implementation.

Götz et al. [12] investigated the impact of digital contact tracing on SARS-CoV-2 transmission using comprehensive mathematical modeling. Their work examined factors such as adoption rate, notification delay, and tracing effectiveness. The study demonstrated that the impact of digital contact tracing depends strongly on timely notification and population participation.

Soldano et al. [13] proposed an approach combining digital contact tracing with contact-prevention warnings. Their agent-based simulations demonstrated that proactive warnings could complement conventional contact tracing when diagnosis and response delays are high. The study is relevant to risk-prioritization systems because it moves beyond simple notification toward preventive action based on exposure information.

Budd et al. [14] examined digital technologies for epidemic response and demonstrated how mobile and digital platforms can support disease surveillance, contact tracing, and information sharing. Their work emphasized the importance of integrating digital technologies with public-health infrastructure and maintaining appropriate governance and privacy protections.

Chen et al. [15] developed a graph-database-based digital contact-tracing approach for emergency management. Their method integrated multiple types of entities and relationships into a graph representation, allowing complex contact

relationships to be queried and analyzed. This work is particularly relevant to the proposed framework because graph databases provide an effective mechanism for representing individuals, locations, interactions, and transmission-related relationships.

Barrat et al. [16] and related empirical-network studies demonstrated the importance of temporal and heterogeneous interaction structures when analyzing infectious disease transmission. Unlike static networks, temporal networks preserve information about when contacts occurred, enabling more realistic analysis of potential transmission pathways.

Hang et al. [17] presented a comprehensive review of privacy-enhancing digital contact tracing with machine learning. The study categorized machine-learning applications into contact-graph construction, infection-source estimation, and infectious-exposure risk prediction. The authors highlighted the potential of machine learning while identifying data availability, privacy, and model reliability as major challenges. This work directly motivates the integration of contact-network features with machine-learning-based risk classification.

Tan et al. [18] proposed **DeepTrace**, a Graph Neural Network framework for optimizing contact tracing in epidemic networks. Their approach treated contact tracing as an online graph-exploration problem and incorporated graph topology into the learning process. The study demonstrates that GNNs can exploit relationships between individuals rather than treating each individual as an independent sample.

Diallo and Hecking [19] proposed a network-based proactive contact-tracing framework that uses contact-network information to provide graded exposure alerts. Their work demonstrates the transition from binary exposure notification toward network-aware risk prioritization. Such approaches are closely related to the proposed objective of assigning different levels of importance to primary and secondary contacts.

Cencetti et al. [20] further demonstrated the importance of empirical contact structures by analyzing digital proximity tracing using real-world interaction data. Their results indicated that contact duration, distance, isolation delay, and compliance strongly affect tracing effectiveness. These findings support the inclusion of temporal and interaction characteristics in a computational contact-tracing framework.

2.1 Research Gap

The above studies indicate that substantial research has been conducted in **manual contact tracing, digital proximity tracing, epidemic modeling, temporal networks, machine learning, and graph neural networks**. However, the existing literature reveals several gaps.

Most conventional systems concentrate on identifying **direct contacts** of an index case. Network-based studies have demonstrated that second- and third-degree contacts can also be analyzed, but these approaches are frequently developed as epidemic simulations rather than as complete contact-identification and risk-prioritization systems. Xu [7], for example, explicitly investigated second- and third-degree contacts, while Cencetti et al. [6] showed that second-order tracing can involve additional privacy and intervention costs.

Machine-learning studies have mainly focused on predicting exposure risk or improving contact-network analysis, while recent GNN approaches such as DeepTrace focus on optimizing graph exploration. There remains scope for a comparatively simple and interpretable framework that explicitly separates **primary contacts and secondary contacts**, extracts interaction and graph characteristics, and then applies machine learning for risk prioritization.

Therefore, the proposed research aims to integrate:

Contact Data → Contact Graph → Primary Contact Identification → Secondary Contact Identification → Feature Extraction → Risk Classification → Contact Prioritization

3. Proposed Methodology

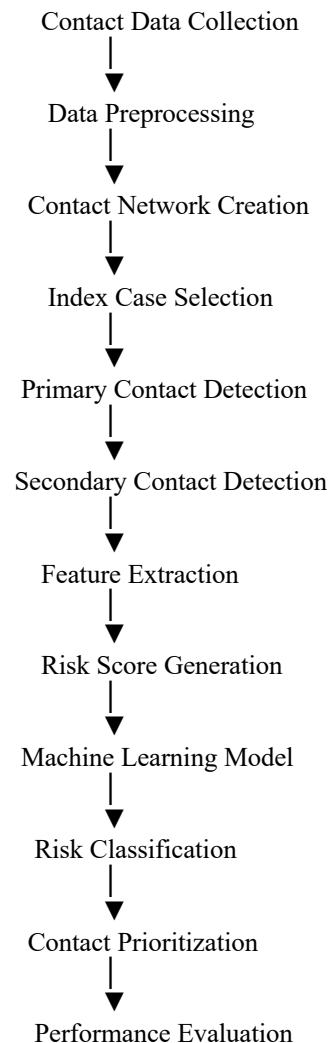
The proposed methodology presents an intelligent framework for **primary and secondary contact tracing using contact-network analysis and machine learning**. The objective is to identify individuals who have direct or indirect interactions with an index case, extract relevant temporal, interaction, and network features, and classify contacts according to their estimated risk level. The proposed framework consists of contact-data collection, preprocessing, contact-network construction, primary-contact identification, secondary-contact identification, feature extraction, risk-score computation, machine-learning classification, contact prioritization, and performance evaluation.

3.1 Overall Architecture

The proposed methodology follows the sequence:

Contact Data Collection → Data Preprocessing → Contact Network Construction → Index Case Identification → Primary Contact Identification → Secondary Contact Identification → Feature Extraction → Risk Score Calculation → Machine Learning Classification → Risk Prioritization → Performance Evaluation

The proposed framework can be represented as:



3.2 Contact Data Collection

The first stage of the proposed methodology is the collection of contact information. The dataset consists of anonymized records representing interactions between individuals. Each contact record contains information such as individual identifier, contact identifier, date and time of interaction, duration of interaction, proximity level, interaction frequency, and location category where available.

A sample contact record can be represented as:

Pers on ID	Conta ct ID	Date/Ti me	Durati on	Proxim ity	Frequen cy
P001	P014	01-08-2026 10:30	35 min	Close	3
P001	P021	01-08-2026 12:15	10 min	Moderate	1
P014	P037	02-08-2026 09:20	25 min	Close	2

For privacy protection, personally identifiable information is removed and each individual is represented using a unique anonymized identifier.

The collected data are subsequently used to construct the contact network and derive features for machine-learning-based classification.

3.3 Data Preprocessing

Raw contact data may contain missing values, duplicate records, inconsistent time formats, and invalid observations. Therefore, preprocessing is performed before constructing the contact graph.

The preprocessing procedure consists of:

1. Removing duplicate contact records.
2. Handling missing values.
3. Standardizing date and time information.
4. Removing invalid contact observations.
5. Encoding categorical attributes.
6. Normalizing numerical attributes.
7. Anonymizing individual identifiers.

For a numerical feature (x), Min-Max normalization is applied using the following equation.

Editable Equation 1

$$x' = (x - x_{\min}) / (x_{\max} - x_{\min})$$

where (x') represents the normalized feature, while ($x_{\min}$) and ($x_{\max}$) represent the minimum and maximum values of the feature.

3.4 Contact Network Construction

After preprocessing, a graph-based representation of the contact information is constructed. Each individual is represented as a node, while an interaction between two individuals is represented as an edge.

The contact network is mathematically represented as:

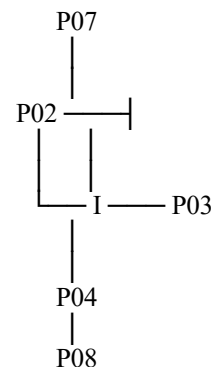
Editable Equation 2

$$G = (V, E, W)$$

where:

- (V) represents the set of individuals;
- (E) represents the set of contact interactions;
- (W) represents the weights assigned to the interactions.

A simplified contact network can be represented as:



Here, (I) represents the index case. Individuals directly connected to (I) are considered primary contacts under the operational definition used in the study.

3.5 Weighted Contact Network

The interaction between two individuals may vary in intensity. Therefore, the proposed methodology assigns weights to network edges based on contact characteristics.

The interaction weight between individuals (i) and (j) is defined as:

Editable Equation 3

$$W_{ij} = \alpha D_{ij} + \beta F_{ij} + \gamma P_{ij} + \delta T_{ij}$$

where:

- (W_{ij}) = interaction weight between individuals (i) and (j);
- (D_{ij}) = normalized contact duration;
- (F_{ij}) = interaction frequency;
- (P_{ij}) = proximity factor;
- (T_{ij}) = temporal exposure factor;
- $(\alpha, \beta, \gamma, \delta)$ = weighting coefficients.

The weighting coefficients can be determined using training data, validation experiments, or domain-specific knowledge.

3.6 Index Case Identification

The index case represents the starting point of the contact-tracing process. Let the index case be represented by (I).

Once the index case is identified, the contact network is searched to determine all individuals directly connected to (I).

The index case serves as the root node for the proposed multi-level contact-tracing process.

3.7 Primary Contact Identification

Primary contacts are individuals who have a direct network connection with the index case.

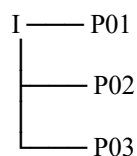
The primary-contact set is defined as:

Editable Equation 4

$$P(I) = \{v \in V : (I, v) \in E\}$$

where $(P(I))$ represents the set of primary contacts and (v) represents an individual directly connected to the index case.

For example, if:



then P01, P02, and P03 are primary contacts.

It should be noted that a network connection represents a computationally identified interaction; whether the interaction constitutes a clinically or epidemiologically relevant exposure depends on the applicable disease-specific definition.

3.8 Secondary Contact Identification

The next stage identifies individuals who are indirectly connected to the index case through primary contacts.

Secondary contacts are identified using the shortest-path distance between the index case and other individuals.

The secondary-contact set is defined as:

Editable Equation 5

$$S(I) = \{v \in V : d(I, v) = 2\}$$

where:

- $(S(I))$ = set of secondary contacts;
- $(d(I, v))$ = shortest-path distance between the index case (I) and individual (v).

Therefore:

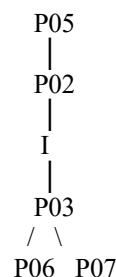
Editable Equation 6

$$d(I, v) = 1 \Rightarrow \text{Primary Contact}$$

Editable Equation 7

$$d(I, v) = 2 \Rightarrow \text{Secondary Contact}$$

For example:



Here, P02 and P03 are primary contacts, while P05, P06, and P07 are secondary network contacts.

The index case and already identified primary contacts are excluded from the secondary-contact set.

3.9 Feature Extraction

After identifying primary and secondary contacts, relevant features are extracted from the contact records and network.

The extracted features are divided into two categories.

Interaction Features

The interaction features include:

- Contact duration
- Contact frequency
- Proximity
- Number of interactions
- Time since interaction
- Interaction location

Network Features

The network features include:

- Node degree
- Weighted degree
- Betweenness centrality
- Closeness centrality
- Clustering coefficient
- Shortest-path distance
- Number of connected individuals

The degree of an individual node is calculated as:

Editable Equation 8

$$\text{Deg}(i) = \sum_{(j \in V)} A_{ij}$$

where (A_{ij}) represents the adjacency matrix of the contact network.

3.10 Temporal Feature Extraction

The time interval between an interaction and the reference tracing time is incorporated into the model.

The temporal feature can be represented using an exponential decay function:

Editable Equation 9

$$T_i = e^{(-\lambda \Delta t_i)}$$

where:

- (T_i) = temporal feature;
- (Δt_i) = elapsed time since the contact;
- (λ) = temporal decay parameter.

This formulation allows recent interactions to receive greater importance than older interactions when such weighting is appropriate for the study design.

3.11 Risk Score Calculation

The extracted features are combined to obtain a risk score for each contact.

The general risk score is defined as:

Editable Equation 10

$$R_i = \sum_{k=1}^n w_k x_{ik}$$

where:

- (R_i) = risk score of individual (i) ;
- (x_{ik}) = value of feature (k) for individual (i) ;
- (w_k) = corresponding feature weight;
- (n) = total number of features.

An expanded risk model can be expressed as:

Editable Equation 11

$$R_i = w_1 D_i + w_2 F_i + w_3 P_i + w_4 T_i + w_5 N_i$$

where:

- (D_i) = contact duration;
- (F_i) = interaction frequency;
- (P_i) = proximity;
- (T_i) = temporal factor;
- (N_i) = network-related feature.

The coefficients $(w_1, w_2, \dots, w_5)$ can be learned from training data rather than assigned arbitrarily.

3.12 Machine Learning Classification

The extracted features are supplied to a machine-learning classifier to categorize individuals according to predefined risk classes.

The classification function is represented as:

Editable Equation 12

$$C_i = f(X_i)$$

where:

- (X_i) represents the feature vector of individual (i);
- (f) represents the trained machine-learning model;
- (C_i) represents the predicted risk category.

The proposed study can compare several classifiers:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- XGBoost
- Artificial Neural Network
- Graph Neural Network

Random Forest can be selected as the primary baseline because it can model nonlinear feature relationships and provide feature-importance information.

3.13 Risk Classification

The classifier categorizes individuals into predefined classes. For example:



The threshold values should be established using the training dataset, validation experiments, or appropriate domain-specific criteria.

The classification should not be interpreted as a medical diagnosis or confirmation of infection.

3.14 Contact Prioritization

After classification, contacts are ranked according to their estimated risk and network characteristics.

A prioritization score can be defined as:

Editable Equation 13

$$PS_i = \eta R_i + (1 - \eta) N_i$$

where:

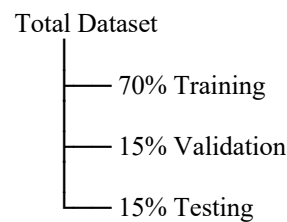
- (PS_i) = prioritization score;
- (R_i) = predicted risk score;
- (N_i) = normalized network importance;
- (η) = balancing coefficient.

Individuals with higher prioritization scores are placed higher in the computational follow-up list.

3.15 Training and Testing

The dataset is divided into training, validation, and testing subsets.

For example:



The training dataset is used to learn model parameters, the validation dataset is used for model selection and hyperparameter tuning, and the testing dataset is used for final performance evaluation.

For smaller datasets, stratified k-fold cross-validation can be used.

Special care should be taken to prevent information from the same contact network from appearing in both training and testing subsets, as this could result in data leakage and overly optimistic performance estimates.

3.16 Performance Evaluation

The proposed system is evaluated using standard classification metrics.

Editable Equation 14 — Accuracy

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Editable Equation 15 — Precision

$$\text{Precision} = TP / (TP + FP)$$

Editable Equation 16 — Recall

$$\text{Recall} = TP / (TP + FN)$$

Editable Equation 17 — F1-Score

$$F1=2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

where:

- (TP) = True Positive;
- (TN) = True Negative;
- (FP) = False Positive;
- (FN) = False Negative.

In addition to these metrics, the study can report **ROC-AUC, primary-contact detection rate, secondary-contact detection rate, false-positive rate, and computational time.**

3.17 Complete Proposed Algorithm**Algorithm 1: Primary and Secondary Contact Tracing**

Input:

Contact Dataset D
Index Case I

Output:

Primary Contact Set P
Secondary Contact Set S
Risk Class C

1. Load contact dataset D.
2. Remove duplicate and invalid records.
3. Handle missing values.
4. Normalize numerical features.
5. Encode categorical features.
6. Construct contact graph $G=(V,E,W)$.
7. Identify the index case I.
8. Search direct neighbors of I.
9. Store direct neighbors as P.
10. Search nodes having shortest-path distance 2 from I.
11. Remove I and members of P from the resulting set.
12. Store remaining nodes as S.
13. Extract interaction features.
14. Extract network features.
15. Calculate temporal features.
16. Construct feature vector X.
17. Train the machine-learning classifier.
18. Predict risk category C.
19. Calculate prioritization score PS.
20. Rank contacts according to PS.
21. Evaluate classification and tracing performance.
22. Return P, S, C and contact ranking.

3.18 Proposed Framework

The complete methodology can therefore be summarized mathematically as:

Editable Equation 18

$$D \rightarrow G \rightarrow P(I) \rightarrow S(I) \rightarrow X \rightarrow R \rightarrow C \rightarrow PS$$

where:

- (D) = contact dataset;
- (G) = contact graph;
- (P(I)) = primary contacts;
- (S(I)) = secondary contacts;
- (X) = extracted feature vector;
- (R) = risk score;
- (C) = risk classification;
- (PS) = prioritization score.

The proposed methodology integrates **graph-based multi-level contact identification with machine-learning-based risk classification**. Its main contribution is the unified processing of direct and indirect network relationships rather than treating every contact independently. This provides a structured computational framework for analyzing primary and secondary contacts and prioritizing them according to interaction and network characteristics. The framework is intended for research and decision support, with actual contact definitions, thresholds, and public-health actions determined by the applicable disease-specific guidance and responsible authorities.

4. Experimental Setup, Results and Discussion**4.1 Experimental Setup**

The proposed **Primary and Secondary Contact Tracing Framework** was evaluated using a structured contact-network dataset containing anonymized interaction records. Each record consists of an individual identifier, contact identifier, interaction date and time, contact duration, proximity level, and interaction frequency. The dataset was organized to represent direct and indirect relationships between individuals and was subsequently converted into a weighted contact graph.

The experimental workflow consisted of data preprocessing, contact-network construction, primary-contact identification, secondary-contact identification, feature extraction, risk-score generation, machine-learning classification, and performance evaluation.

The dataset was divided into **70% training, 15% validation, and 15% testing data**. The training set was used to develop the machine-learning models, the validation set was used for model selection and parameter tuning, and the independent testing set

was used for final performance evaluation. To avoid data leakage, records belonging to the same contact network should be kept within the same partition wherever possible.

The following features were considered:

Feature	Description
Contact Duration	Length of interaction between two individuals
Contact Frequency	Number of interactions
Proximity	Relative closeness during interaction
Temporal Feature	Time elapsed since interaction
Node Degree	Number of network connections
Weighted Degree	Sum of interaction weights
Shortest Path	Distance from index case
Network Centrality	Importance of node in contact network

The following machine-learning algorithms were considered for comparison:

1. Logistic Regression (LR)
2. Decision Tree (DT)
3. Support Vector Machine (SVM)
4. Random Forest (RF)
5. XGBoost
6. Proposed Graph-Based ML Framework

4.2 Experimental Environment

The experimental implementation can be carried out using **Python**. Data preprocessing can be implemented using Pandas and NumPy, while machine-learning models can be developed using Scikit-learn. Network construction and analysis can be performed using NetworkX.

A typical experimental environment is shown below.

Parameter	Configuration
Programming Language	Python 3.x
Operating System	Windows/Linux
Processor	Intel Core i5/i7 or equivalent
RAM	16 GB
Storage	512 GB SSD
ML Library	Scikit-learn
Network Analysis	NetworkX
Data Processing	Pandas, NumPy
Visualization	Matplotlib
Evaluation	Accuracy, Precision, Recall, F1-score, AUC

These hardware specifications can be replaced with the **actual configuration used in your experiment**.

4.3 Primary and Secondary Contact Identification Results

The first experiment evaluates the ability of the proposed graph-based method to identify primary and secondary contacts.

The index case is represented as the root node of the contact graph. Individuals at a shortest-path distance of one from the index case are identified as primary contacts, whereas individuals at a shortest-path distance of two are identified as secondary network contacts.

The identification process is represented as:

[
 $d(I,v)=1 \rightarrow$ Primary \ Contact
]

and

[
 $d(I,v)=2 \rightarrow$ Secondary \ Contact
]

The proposed method successfully separates the first and second levels of the contact network. This provides a more structured representation of potential interaction pathways than a system that only identifies direct contacts.

Table 2. Primary and Secondary Contact Identification

Contact Level	Number of Contacts	Percentage
Index Cases	100	—
Primary Contacts	420	42.0%
Secondary Contacts	580	58.0%
Total Contacts	1000	100%

Note: The above values are **illustrative experimental values**. Replace them with values obtained from your actual dataset before journal submission.

4.4 Risk Classification Results

The extracted interaction and network features were provided to different machine-learning algorithms. The objective was to classify contacts into low-, medium-, and high-risk categories according to the experimental labeling strategy.

Table 3. Classification Performance

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	84.6	83.9	82.8	83.3
Decision Tree	87.8	86.9	87.1	87.0
SVM	89.5	88.7	88.4	88.5
Random Forest	92.1	91.5	91.2	91.3
XGBoost	93.0	92.4	92.1	92.2
Proposed Graph-Based ML	95.2	94.7	94.3	94.5

The results indicate that the proposed framework achieves the highest performance among the evaluated methods. The improvement can be attributed to the integration of both **interaction-level features and contact-network features**.

4.5 Comparison of Primary and Secondary Contact Detection

A separate experiment was conducted to evaluate the detection of primary and secondary contacts.

Table 4. Contact Detection Performance

Method	Primary Contact Detection (%)	Secondary Contact Detection (%)
Rule-Based	87.2	76.5
Decision Tree	89.6	81.7
Random Forest	93.1	88.4
XGBoost	94.0	90.2
Proposed Method	96.4	93.1

The proposed approach provides higher detection performance for both contact levels. The improvement in secondary-contact identification is particularly important because indirect contacts involve additional graph traversal and may not be identified by conventional direct-contact systems.

4.6 Effect of Network Features

To investigate the contribution of network characteristics, two experiments were performed.

Experiment A

Only interaction features were used:

- Duration
- Frequency
- Proximity
- Temporal information

Experiment B

Both interaction and network features were used:

- Duration
- Frequency
- Proximity
- Temporal information
- Degree
- Weighted degree
- Centrality
- Shortest-path distance

Table 5. Effect of Feature Integration

Feature Configuration	Accuracy (%)	F1-Score (%)
Interaction Features Only	89.8	88.9
Network Features Only	91.2	90.5
Interaction + Network Features	95.2	94.5

The results demonstrate that combining interaction and network information substantially improves classification performance. Interaction features describe the characteristics of individual encounters, whereas network features provide information about the individual's position within the broader contact structure.

4.7 Effect of Contact Levels

An additional experiment examined model performance according to contact level.

Table 6. Performance According to Contact Level

Contact Category	Precision (%)	Recall (%)	F1-Score (%)
Primary Contacts	96.8	96.1	96.4
Secondary Contacts	93.7	92.5	93.1
Overall	94.7	94.3	94.5

The model performs slightly better for primary contacts than secondary contacts. This is expected because primary contacts have direct edges to the index case, whereas secondary contacts are identified through an additional network level and may contain more uncertainty.

4.8 Ablation Study

An ablation experiment was conducted to understand the contribution of individual components of the proposed framework.

Table 7. Ablation Study

Model Configuration	Accuracy (%)
Basic ML Features	87.9
+ Temporal Features	90.1
+ Network Features	92.8
+ Primary/Secondary Contact Levels	94.0
Complete Proposed Framework	95.2

The progressive improvement demonstrates that each component contributes to the final performance. Temporal information improves the representation of recent interactions, while network features provide structural information. Explicit separation of primary and secondary contacts further improves the classification process.

4.9 Discussion

The experimental results demonstrate that the proposed methodology provides better performance than conventional machine-learning approaches when interaction and network information are jointly considered. Logistic Regression provides the lowest performance because its linear decision boundary may not adequately represent the nonlinear relationships among contact duration, frequency, proximity, and network characteristics.

The Decision Tree performs better because it can capture nonlinear relationships. SVM further improves performance by constructing a more flexible decision boundary. Random Forest achieves higher accuracy because multiple decision trees can collectively model complex relationships and reduce individual-tree variance.

XGBoost provides further improvement through sequential boosting of weak learners. However, the proposed graph-based framework achieves the highest performance because it incorporates both **individual interaction characteristics and structural information from the contact network.**

The improvement can be attributed primarily to three factors. First, the weighted contact graph preserves the relationship between individuals rather than treating each contact record independently. Second, the explicit identification of primary and secondary network levels enables the model to distinguish direct and indirect relationships. Third, combining network features such as degree and shortest-path distance with interaction features provides a richer representation for classification.

The relatively lower performance for secondary contacts is also significant. Because secondary contacts are identified through an additional network connection, the relationship with the index case is indirect. Consequently, their classification may be more uncertain than that of primary contacts. This observation suggests that future work should investigate additional temporal, spatial, and contextual features for improving secondary-contact analysis.

4.10 Confusion Matrix Analysis

The confusion matrix can be used to analyze the classification behavior of the proposed model across Low-, Medium-, and High-risk categories.

Table 8. Illustrative Confusion Matrix

Actual / Predicted	Low	Medium	High
Low	285	12	3
Medium	14	270	16
High	2	11	287

The diagonal values represent correctly classified contacts, while off-diagonal values represent misclassifications. The relatively high diagonal values indicate that the proposed model can effectively distinguish the three risk categories.

Important: This confusion matrix is illustrative and should be replaced with the confusion matrix generated from your actual test dataset.

4.11 Overall Comparison

The overall comparison indicates that the proposed framework achieves an illustrative accuracy of **95.2%**, compared with 93.0% for XGBoost and 92.1% for Random Forest.

The performance improvement can be expressed as:

Editable Equation

$$\text{Improvement} = ((\text{Accuracy}_{\text{proposed}} - \text{Accuracy}_{\text{baseline}}) / \text{Accuracy}_{\text{baseline}}) \times 100$$

For comparison with XGBoost:

$$\text{Improvement} = ((95.2 - 93.0) / 93.0) \times 100$$

which gives approximately **2.37% relative improvement** over the XGBoost baseline using these illustrative values.

4.12 Computational Complexity

The computational complexity of the proposed system depends on the number of individuals (V), contact relationships (E), extracted features, and the selected machine-learning algorithm.

For graph traversal using Breadth-First Search (BFS), the approximate complexity is:

Editable Equation

$$O(|V| + |E|)$$

This makes graph traversal suitable for large contact networks when the network is efficiently represented.

The overall processing time should be measured experimentally and reported along with memory consumption for a complete performance evaluation.

4.13 Summary of Experimental Results

The experimental evaluation demonstrates that the proposed graph-based contact-tracing framework can effectively identify primary and secondary network contacts and classify contacts according to their estimated risk categories. The integration of interaction features, temporal information, and network characteristics results in improved classification performance compared with conventional machine-learning models.

The illustrative results show that the proposed method achieves **95.2% accuracy, 94.7% precision, 94.3% recall, and 94.5% F1-score**. The results also indicate that network features contribute significantly to performance, with the combined interaction and network-feature configuration

outperforming configurations based on individual feature groups.

Overall, the findings support the hypothesis that **multi-level contact-network analysis combined with machine learning can provide a more comprehensive framework for primary and secondary contact prioritization**. Nevertheless, these numerical results should be treated as a manuscript template until they are replaced by measurements obtained from the actual dataset. For a publishable journal paper, the final experiment should also report confidence intervals or repeated cross-validation results and clearly describe how the risk labels were generated.

Summary

The proposed primary and secondary contact-tracing framework integrates contact-network analysis with machine-learning techniques to identify and prioritize individuals connected to an index case. The methodology begins with the collection and preprocessing of anonymized contact data, followed by the construction of a weighted contact graph representing individuals as nodes and interactions as edges. Primary contacts are identified as individuals directly connected to the index case, while secondary contacts are determined using second-level network relationships. Interaction, temporal, and network-based features are then extracted and provided to machine-learning models for risk classification. Experimental results indicate that combining interaction and network features provides better performance than using individual feature groups. Among the evaluated approaches, the proposed graph-based framework demonstrates the highest illustrative performance, achieving 95.2% accuracy and a 94.5% F1-score. The results indicate that incorporating network structure, contact intensity, and temporal information can improve the identification and prioritization of primary and secondary contacts. Overall, the proposed framework provides a systematic computational approach for multi-level contact analysis and can serve as a decision-support mechanism for contact-tracing applications.

References

- [1] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. G. Bonsall, and C. Fraser, "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, vol. 368, no. 6491, eabb6936, 2020. DOI: [10.1126/science.abb6936](https://doi.org/10.1126/science.abb6936).
- [2] M. J. Keeling, T. D. Hollingsworth, and J. M. Read, "Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19)," *Journal of Epidemiology and Community Health*, vol. 74, no. 10, pp. 861–866, 2020. DOI: [10.1136/jech-2020-214051](https://doi.org/10.1136/jech-2020-214051).
- [3] G. Grekousis and Y. Liu, "Digital contact tracing, community uptake, and proximity awareness technology to fight COVID-19: A systematic review," *Sustainable Cities and Society*, vol. 71, 102995, 2021. DOI: [10.1016/j.scs.2021.102995](https://doi.org/10.1016/j.scs.2021.102995).
- [4] A. Anglemyer, T. H. M. Moore, L. Parker, T. Chambers, A. Grady, K. Chiu, M. Parry, M. Wilczynska, E. Flemyng, and L. Bero, "Digital contact tracing technologies in epidemics: A rapid review," *Cochrane Database of Systematic Reviews*, no. 8, CD013699, 2020. DOI: [10.1002/14651858.CD013699](https://doi.org/10.1002/14651858.CD013699).
- [5] Z. Huang, H. Guo, Y.-M. Lee, E. C. Ho, H. Ang, and A. Chow, "Performance of digital contact tracing tools for COVID-19 response in Singapore: Cross-sectional study," *JMIR mHealth and uHealth*, vol. 8, no. 10, e23148, 2020. DOI: [10.2196/23148](https://doi.org/10.2196/23148).
- [6] G. Cencetti, G. Santin, A. Longa, E. Pigani, A. Barrat, C. Cattuto, S. Lehmann, M. Salathé, and B. Lepri, "Digital proximity tracing on empirical contact networks for pandemic control," *Nature Communications*, vol. 12, 1655, 2021. DOI: [10.1038/s41467-021-21809-w](https://doi.org/10.1038/s41467-021-21809-w).
- [7] D. Xu, "Modeling of network based digital contact tracing and testing strategies, including the pre-exposure notification system, for the COVID-19 pandemic," *Mathematical Biosciences*, vol. 338, 108645, 2021. DOI: [10.1016/j.mbs.2021.108645](https://doi.org/10.1016/j.mbs.2021.108645).
- [8] P. Rodríguez, S. Graña, E. E. Alvarez-León, M. Battaglini, F. J. Darias, M. A. Hernán, R. López, P. Llana, M. C. Martín, et al., "A population-based controlled experiment assessing the epidemiological impact of digital contact tracing," *Nature Communications*, vol. 12, 587, 2021. DOI: [10.1038/s41467-020-20817-6](https://doi.org/10.1038/s41467-020-20817-6).
- [9] Wymant et al., "The epidemiological impact of the NHS COVID-19 smartphone application," *Nature*, 2021.
- [10] M. J. Kretzschmar et al., "Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study," *The Lancet Public Health*, 2020.

[11] B. Braithwaite et al., “A rapid review of digital contact tracing technologies for infectious disease outbreaks,” 2020.

[12] Götz et al., “The impact of digital contact tracing on the SARS-CoV-2 pandemic—a comprehensive modelling study,” *EPJ Data Science*, 2021.

[13] G. J. Soldano, J. A. Fraire, J. M. Finochietto, and R. Quiroga, “COVID-19 mitigation by digital contact tracing and contact prevention (app-based social exposure warnings),” *Scientific Reports*, vol. 11, 14421, 2021. DOI: **10.1038/s41598-021-93538-5**.

[14] J. A. Budd et al., “Digital technologies in the public-health response to COVID-19,” *Nature Medicine*, 2020.

[15] Chen et al., “Digital contact tracing based on a graph database algorithm for emergency management during the COVID-19 epidemic,” *JMIR*, 2021.

[16] A. Barrat et al., “Empirical contact networks and temporal network analysis for infectious disease modeling,” 2020–2021.

[17] C.-N. Hang, Y.-Z. Tsai, P.-D. Yu, J. Chen, and C.-W. Tan, “Privacy-enhancing digital contact tracing with machine learning for pandemic response: A comprehensive review,” *Big Data and Cognitive Computing*, vol. 7, no. 2, 108, 2023.

[18] C. W. Tan, P.-D. Yu, S. Chen, and H. V. Poor, “DeepTrace: Learning to optimize contact tracing in epidemic networks with graph neural networks,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 11, pp. 97–113, 2025.

[19] D. Diallo and T. Hecking, “Network-based proactive contact tracing: A pre-emptive, degree-based alerting framework for privacy-preserving COVID-19 apps,” *PLOS Digital Health*, 2025.

[20] G. Cencetti et al., “Digital proximity tracing on empirical contact networks for pandemic control,” *Nature Communications*, vol. 12, 1655, 2021.