UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

TOWARDS AI-ENABLED DIAGNOSTICS FROM NOISY AND INCOMPLETE DATA

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY IN ELECTRICAL AND COMPUTER ENGINEERING

BY

MUHAMMAD SAJID RIAZ
Norman, Oklahoma
2024

TOWARDS AI-ENABLED DIAGNOSTICS FROM NOISY AND INCOMPLETE DATA

A DISSERTATION APPROVED FOR THE
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

BY THE COMMITTEE CONSISTING OF

Dr. Ali Imran, Chair

Dr. Samuel Cheng

Dr. Timothy Ford

Dr. Choon Yik Tang

Acknowledgements

I would like to express my deepest gratitude to my advisor Dr. Ali Imran, for his invaluable support and guidance throughout my PhD. Given the interdisciplinary nature of my research, Dr. Imran's continuous feedback, cross-domain expertise, and practical insights have been instrumental in shaping my research directions that helped me do applied research which was my major goal as a PhD student. His unwavering dedication and passion for research have been truly inspiring and have helped me gain a deeper understanding of applied and impactful research. Thank you, Dr. Imran, for being an exceptional mentor, kind leader, and for helping me at every stage of my academic and professional development.

My heartfelt appreciation also goes to Dr. Samuel Cheng, Dr. Choon Yik Tang, and Dr. Timothy Ford for taking the time to review my work through this journey. Their valuable feedback and suggestions were immensely helpful in improving my work.

I express my sincere appreciation to Renee Wagenblatt, Josh Davis, Lee Lawson, Lauren Condry, Monica Goodwin, Denise Davis, and Krista Pettersen for their exceptional support and assistance in handling all administrative tasks during my time as a PhD student at OU-Tulsa. Their promptness to my requests mostly closer to the deadlines and tireless work behind the scenes always made me feel safe and well taken care of.

I extend my sincere thanks to my friends and colleagues at AI4Networks Research Center, to Dr. Farooq, Dr. Manalastas, Waseem, Umar, Charles, Dr. Masood, Dr. Zaidi, Dr. Kasi, Aneeqa, Tabish, and especially Dr. Quereshi for continuous support and assistance during my academic journey. I feel proud and privileged to be a part of this Center, which provided me with a nurturing and stimulating environment to learn and grow.

I am grateful to extend my appreciation to my friends and fellow teammates at Green Country Cricket Club (GCCC) for granting me the chance to participate in a sport that holds immense significance for me. Throughout my Ph.D. journey, cricket, facilitated by

GCCC, has played a pivotal role in maintaining my well-being and serving as a means of emotional release.

Finally, I want to dedicate this work to my incredible parents, whose unwavering dedication has always propelled me towards success, as well as to my admirable brothers, Wajid and Arshad, and my loving sister. Additionally, I am profoundly grateful to my caring, supportive, and beautiful wife, Tahira Saleem for being my support system and taking care of our brilliant kids Anaya Sajid and Ali Sher, to help me focus on my academic and professional goals. Their unending support, boundless love, and constant encouragement have served as my steadfast pillars of strength and inspiration throughout my academic journey.

Table of Contents

List of Figures

List of Tables

Abstract

The advent of artificial intelligence (AI) has revolutionized diagnostic procedures, promising enhanced accuracy and efficiency across various domains. This dissertation leverages AI capabilities for diagnostic purposes, particularly focusing on the challenges posed by noisy and incomplete data, with a specific emphasis on image and audio data modalities.

The modern diagnostic landscape often encounters datasets fraught with noise and incompleteness, stemming from diverse sources such as sensor limitations, environmental factors, or inherent complexities in the data collection process. Addressing these challenges necessitates the development of robust AI methodologies capable of extracting meaningful insights despite the presence of imperfections.

This research endeavors to bridge this gap by proposing innovative AI techniques tailored to handle noisy and incomplete image and audio data. One of the major contributions of this dissertation is a hybrid deep learning-based framework (HYDRA) for root cause analysis of coverage anomalies in emerging cellular networks using minimization of drive tests (MDT) reports. A complete schematic, detailed design, and comparative analysis of HYDRA are given in chapter II of this dissertation. The main objective of this research was to devise novel algorithms that can effectively mitigate the adverse effects of noise and incompleteness, thereby enhancing the reliability and accuracy of diagnostic outcomes.

Furthermore, this study adopts a multi-faceted approach, encompassing both theoretical advancements and practical implementations of a concept framework for proactive future pandemic prediction leveraging multi-modal biosensing data. This approach is presented in the dissertation as a framework called iPREDICT, which leverages the power of deep learning architectures, a plethora of biomarkers acquired using free-life biosensors available in daily life usage smart devices like smartwatches and smartphones.

Moreover, the practical significance of this research extends beyond academic realms, with potential applications spanning diverse sectors such as healthcare diagnostics, emerging cellular networks, and biomedical signal processing. By empowering AI systems to navigate through the complexities of noisy and incomplete data, this dissertation endeavors to pave the way for more robust and dependable diagnostic tools, ultimately contributing to improved decision-making processes and enhanced societal well-being.

In summary, this dissertation represents a significant step towards realizing the full potential of AI-enabled diagnostics in real-world scenarios characterized by noisy and incomplete data. Through a synergy of cutting-edge AI methodologies and practical insights, this research aspires to catalyze advancements that will shape the future of diagnostic technologies across multiple domains.

CHAPTER 1

Introduction

1.1 Background and Motivation

The integration of Artificial Intelligence (AI) into diagnostic work has undeniably revolutionized several industries [1, 2] including telecom and healthcare, enabling faster and more accurate diagnoses. AI-based solutions leverage vast amounts of data to identify patterns, predict outcomes, and aid in decision-making processes. However, the full potential of AI in diagnostics remains hindered by the persistent challenge of dealing with noisy and incomplete data [3, 4]. The accuracy and effectiveness of AI algorithms heavily depend on the quality and quantity of data available for training. In real-world scenarios, obtaining comprehensive and clean data can be complex and resource-intensive, leading to limitations in the scalability and efficiency of AI-based diagnostic solutions. Addressing these issues is crucial to unlocking the true transformative power of AI in diagnostic solutions.

1.2 Problem Statements and Contributions

In this section, I present the above-mentioned issues as Problem Statements (PS) and present the outline of proposed solutions in the form of contributions (C) of this dissertation.

PS1: The first problem that is addressed in this dissertation is the assessment of the health of cellular networks. There are no practical solutions in the literature to investigate the health of cellular networks using MDT-based coverage maps, because current solutions lack: (1) Capability to operate with sparse/incomplete network coverage data.

(2) Capability to diagnose multiple faults in multi-base station deployment scenarios.

To address these challenges, we present a tri-pronged approach as follows:

C1-A: We designed a realistic cellular network based on Brussels city map using an RF-engineering simulator widely used in the cellular industry. The designed network simulates the multi-fault in multiple BS network, that helps acquire practical network data (MDT reports) that is incomplete as well as noisy.

C1-B: This dissertation propose the usage of raw MDT reports as coverage maps to investigate the health of the network without compromising on the practicality as well as the efficiency of the solution. To achieve this we developed custom software to leverage our indigenous over-the-air 5G and beyond testbed (TurboRAN) for data collection and network health investigation.

C1-C: To maximize the efficiency of this solution I proposed and analyzed a hybrid deep learning-based solution to mitigate the challenge of data sparsity for the AI-enabled diagnostics in cellular networks domain.

PS2: Inspired by the seminal work led by Dr. Imran on COVID-19 and other respiratory illnesses detection through cough sounds, a significant amount of research has been conducted in recent years, with some studies reporting accuracies of up to 98%. Nevertheless, the majority of these studies have not taken into account the practical challenges posed by background noise. Overcoming this hurdle is crucial for the practical application of cough or acoustic-based diagnostics, including the iPredict framework and beyond.

C2: We propose a robust screening solution based on several noise removal, robustness, and mitigation techniques. While the current solutions show higher accuracy on certain datasets, our solution presents a more robust performance on noisy acoustic data.

PS3: Another challenge in the scalability of acoustic diagnostic solutions e.g., in the smartphone-based screening is the hardware and software diversity of such acoustic data collection devices. This major limitation of the existing solutions is also not fully understood in the literature.

C3: We present an analysis of several scalability issues that compromise the AI-based screening of respiratory health conditions. We carry out case studies using various devices to identify the variations in the acoustic data collection devices that lead to compromised efficacy of the current solutions. This analysis highlights many future research directions that will lead to more robust and scalable screening solutions.

PS4: Based on the current literature there is no practical solutions available for future pandemic prediction. This is primarily due to the complex nature of the problem and also the novelty of each pandemic based on characteristics of virus that cause the epidemic that later turns into a pandemic.

C4: Leveraging the previous work from of our research center and building on the work that I did as a solution to PS2 and PS3, we present a hotspot predictions framework iPREDICT. This is a concept framework that presents the design of a proactive pandemic prediction solution enabled by omni-present cellular networks, a plethora of biomarkers data collected using biosensing devices like smartwatches, and advances in AI and data analysis capabilities. iPREDICT is expected to inspire more research that will lead towards practical solutions for the prediction of future pandemics.

1.3 Current and Planned Dissemination and Publications

Awards

A1: Awarded Gallogly College of Engineering Dissertation Excellence Award by the University of Oklahoma, Spring 2024.

A2: Awarded 2nd position at 3-Minute Thesis Competition by University of Oklahoma, Tulsa, Spring 2022.

Peer-Reviewed Journal Articles

J1: M. S. Riaz, Shaukat, M., Saeed, T., Ijaz, A., Qureshi, H.N., Posokhova, I., Sadiq, I., Rizwan, A. and Imran, A., 2024. iPREDICT: AI enabled proactive pandemic prediction using biosensing wearable devices. Informatics in Medicine Unlocked, p.101478.

J2: M. S. Riaz, H. N. Qureshi, U. Masood, A. Rizwan, A. Abu-Dayya and A. Imran, ”A Hybrid Deep Learning-Based (HYDRA) Framework for Multifault Diagnosis Using Sparse MDT Reports,” IEEE Access, IEEE Access, vol. 10, pp. 67140-67151, IEEE, June 2022.

J3: A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, **M. S. Riaz**, K. Ali, C. N. John, I. Hussain, and M. Nabeel, ”AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App,” Elsevier Informatics in Medicine Unlocked, Elsevier Informatics in Medicine Unlocked, vol. 20, pp. 100378, Elsevier, June 2020.

J4: M. S. Riaz, H. N. Qureshi, and A. Imran, ”RAISED: Towards More Reliable AI-enabled Screening of COVID-19 from Cough” **[Under Internal Review]**

Peer-Reviewed Conference Papers

C1: M. S. Riaz, H. N. Qureshi, U. Masood, A. Rizwan, A. Abu-Dayya, and A. Imran, ”Deep Learning-based Framework for Multi-Fault Diagnosis in Self-Healing Cellular Networks,” IEEE Wireless Communications and Networking Conference (WCNC 2022), pp. 746-751, April 2022.

1.4 Organization

The rest of the dissertation is organized as follows: Chapter 2 presents the AI-enabled health assessment i.e. fault diagnosis in cellular networks. It also addresses the challenges of incomplete network data and how to enrich the data and its respective challenges. Chapter 3 presents the statistical analysis of acoustic features for screening various respiratory conditions from noisy acoustic data. Chapter 4 presents iPREDICT framework, an outbreak or hotspot prediction framework by leveraging multi-modal data from omni-present commodity biosensing devices, Chapter 5 addresses in detail the scalability challenges in leveraging smartphone-based preliminary screening of respiratory illness. Building on scalable and robust-to-noise diagnostic capabilities achievable via low-cost omni present devices such as smartphones. Finally, Chapter 6 presents the conclusion and outlines potential future work in image and acoustic analysis for anomaly diagnosis.

CHAPTER 2

Assessing the Health of Network from Incomplete Noisy Image Data

2.1 Introduction

In this chapter, we will dive deep into the diagnostics from noisy and incomplete data for the root cause analysis of coverage anomalies in emerging cellular networks. Management tasks in emerging cellular networks are becoming more complex due to evolving network architecture, rapidly increasing and diversifying network traffic, and the growing number of network parameters [4, 5]. Among several management challenges in emerging cellular networks, one major challenge is the timely detection and diagnosis of faults. The increasing complexity of emerging cellular networks and the ultra-reliability requirements of numerous emerging applications are intensifying the challenges of the detection and diagnosis of faults.

Faults, that can lead to hard outages (complete coverage degradation) or soft outages (partial service degradation) in cellular networks can occur due to several reasons. These include poor network design, including improperly configured parameters such as the number, types, and locations of the base stations (BS), antenna heights, sector orientation, tilt, power, frequency reuse patterns, or the number of carriers, among others. Other types of faults can occur due to hardware, software, or functionality failures (e.g., power supply or radio board and network connectivity failures) [6].

Traditionally, outages resulting from faults are detected by human-based monitoring of either alarms, performance counters, or complaints filed by network subscribers [6, 7]. This can take hours and at times days to resolve outage issues. Therefore, for better Quality of Experience (QoE) and Quality of Service (QoS), network providers spend a lot of money, time, and resources to do coverage testing via drive tests. This helps them resolve problems caused by poor parameter configuration and environment change at the cost of heavy capital and operational expenditures. Outages caused by parameter misconfiguration or hardware or software failure that did not raise an alarm are even more challenging to detect and diagnose. These require network experts to manually analyze network logs which can, in turn, further slow down the outage compensation process. Moreover, this challenge of fault detection and root cause analysis is especially aggravated in emerging ultra-dense networks, where the same advances in network design that bring advantages such as higher data rates and capacity as compared to legacy networks, e.g., densification, also lead to the growing complexity of the network, making it difficult to manually detect and diagnose faults. The additional burden of growing operational and capital expenditures is making matters worse. Therefore, outage detection and fault diagnosis through the conventional human monitoring of logs and counters or manual collection of data through the drive-test is neither a practical nor viable option, particularly in emerging complex and dynamic network environments [8, 9].

Network automation solutions, i.e., self-healing solutions are needed to automate the process of fault detection and diagnosis. Only when the outages and their root cause are detected promptly without drive tests and humans in the loop, will the network be able to take actions to compensate for these outages autonomously. The automatic root cause analysis of outage problems can save billions of dollars to network providers annually, by replacing manual resolution of coverage-related anomalies [7]. To wake of this need, the 3GPP has introduced minimization of drive test (MDT) reports feature [10], where the user equipment (UE) periodically sends network coverage related key performance indicators (such as Reference Signal Received Power and Quality, RSRP, and RSRQ respectively) along with their geographical location to their serving base stations, thus eliminating the need for drive tests. Following the standardization of MDT reports, the problem of outage detection and automated fault diagnosis using MDT reports has gained significant attention in the literature.

2.2 Related Work and Open Challenges

Although outage detection has been studied extensively in the literature [11, 12, 13, 14, 15, 16, 17], relatively a small number of studies have focused on outage diagnosis [18, 19, 20].

Among the several studies that focus on outage detection, are works that use convolutional neural networks (CNN) [11], deep neural networks (DNN) [13], support vector machine (SVM) [15], and several other machine learning (ML)-based methods [21, 13, 15, 22, 23, 24]. For a thorough review of outage detection, the reader is referred to a recent survey presented in [5]. A key insight from the extensive review presented in [5] is that almost all existing studies on outage detection and diagnosis overlook a major practical challenge while using MDT reports i.e., the spatial sparsity of MDT reports in the real network. That is most studies assume that MDT reports are available from each point in the area under concern, an assumption that does not hold in a real deployment. This is because MDT reports can be received only from bins where users are present. This usually is a small fraction of the area of interest. In ultra-dense deployments, small cells contain even fewer users compared to macro cells. This makes the number of MDT reports per cell even smaller. This poses a major practical problem for automation solutions that leverage MDT data. However, this problem is often overlooked in the literature by assuming that ample MDT reports are available to represent network KPIs in the whole coverage area.

In comparison to outage detection, outage diagnosis that is the focus of this research, remains relatively under-investigated in the literature. The study in [18] is among those few works that focus on outage diagnosis using self-organizing maps (SOM). Another such study in [25] also present a fault diagnosis framework using SOM. However, the solutions presented in both [18] and [25] are semi-supervised and require input from experts for accurate labeling of the clusters (formed based on different fault classes). Apart from that the SOMs are not robust to varying distributions of data, which makes them nongeneralizable to use with real network MDT reports, due to the sparse nature of MDT data.

In [26] researchers propose an ensemble model (combining two or more classification techniques) for fault diagnosis, which use multiple classifiers that diagnose the current state of the network (normal or anomalous) based on a majority vote, from given network key performance indicators (KPIs) e.g., signal-to-interference-plus-noise ratio (SINR), received signal received power (RSRP), etc. This solution adds cost sensitivity based on misclassification of faults. The cost function assigns different costs for different faults depending upon the severity of the fault. However, the MDT training data sparsity and multi-fault in multiple BSs are not addressed in this work either.

The most relevant to this work are the studies presented in [27, 28, 29]. The researchers in [27] and [28] present a fault diagnosis solution using neuromorphic AI and classical ML methods, respectively. They use MDT reports to generate radio environment maps (REMs). Their analysis shows that random forest (RF) outperforms CNN when MDT data is available from the entire coverage area. However, as explained earlier real network MDT data is expected to be both sparse and noisy. Also, both these studies consider faults in a single BS, i.e., they assume a single fault at a time. Hence, while offering a promising first set of results on fault diagnosis using only MDT data in a multi-BS network, the aforementioned assumptions in [27, 28] render these solutions unsuitable for real deployment. In [29] authors present a deep learning (DL)-based solution for the multi-fault diagnosis in one BS. But the proposed solution requires the drive-test to get the data of a key feature used in the model (throughput). The time and cost for the drive-tests make this solution less scalable for practical deployment.

Based on the literature review summarized above there does not exist a practical fault diagnosis solution in the literature, which has the following capabilities.

1. **Capability to diagnose multiple faults in multi-base station deployment scenario.**
2. **Capability to operate with sparse/incomplete MDT reports.**

In order to address the above-mentioned gaps in the literature, this dissertation proposes

Hybrid Deep Learning-based Root Cause Analysis (HYDRA), which is the first solution that includes the capabilities identified above. HYDRA can detect multiple faults in multi-BS deployment scenarios with realistic sparse MDT data in a fully automated fashion. To achieve these capabilities HYDRA has two key innovative components as illustrated in Fig. 3.1. The first component solves the MDT report sparsity problem by leveraging data enrichment techniques (explained in section 2.5). The second component consists of a novel hybrid of CNN and XGBoost-based models to achieve reliable diagnosis despite noise and sparsity in the enriched or raw REMs.

2.3 Major Contributions of Proposed Framework

The primary contributions of this work are summarized as follows:

1. This dissertation presents first of its kind fault diagnosis solution that can reliably diagnose multiple faults in multiple BSs in the network, caused by both hard outages(network failures leading to no coverage) or soft outages (occurring due to inefficient configuration of network parameters) while using sparse MDT reports. In a real network, faults can occur in different BSs, and they can be of different types. HYDRA is robust to not only different kinds of faults and BS locations but also to variable user densities in the network. This makes HYDRA more feasible to implement in a real cellular network where user density and distribution, never remain static.

2. HYDRA is designed to work with realistic raw sparse MDT reports from the network. These reports are converted into REMs. The REMs are incomplete due to the spatial sparsity of MDT reports resulting from varying user density. I present a practical solution to complete REMs. This is done by investigating and comparing state-of-the-art data enrichment techniques suitable for the problem. I perform a multi-KPI comparative analysis of frequency selective reconstruction (FSR), Bi-

harmonic Equations, and TELEA. Results show that FSR outperforms others and therefore is best suited for REM completion tasks in HYDRA.

3. The inherent noise in the REMs from the sparsity of MDT reports and/or reconstruction makes the task of fault diagnosis using REMs even more challenging. This challenge is ignored in the literature by often assuming the availability of complete and noise-free REMs. Therefore, classic models such as SVM[15], RF [27, 28] and CNN[11] are often used and observed to yield adequate performance in the literature. Our analysis shows that these classic models do not attain adequate performance when realistically sparse MDT data is used to generate REMs. I also address this issue of data sparsity/scarcity by proposing and evaluating a hybrid deep learning model where a CNN is first used to extract the features and the features are then fed into an XGboost model to diagnose network coverage anomalies using raw MDT reports. It is observed that the extensive performance evaluation (using several suitable performance metrics) with varying degrees of MDT data sparsity shows that the proposed hybrid model performs better than both the models when used standalone in terms of robustness to noise and variable UE density.

2.4 Network Topology and Data Acquisition

Figure 3.1 provides a holistic view of the HYDRA framework. The framework has three major blocks: The first block is the acquisition of MDT reports from the network explained in this section. The second block is the conversion of raw MDT reports into REMs and data enrichment using image inpainting, thoroughly explained in Section 3.3.1. The third block is hybrid deep learning-based root cause analysis using sparse REMs data, rigorously described with rationale and implementation details in Section 2.6.

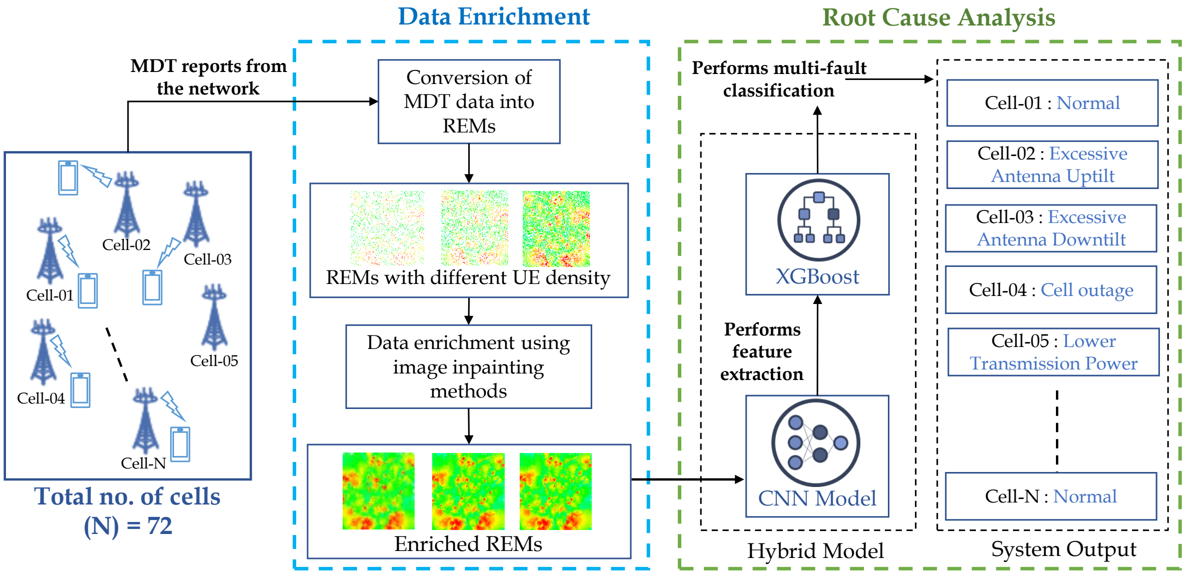

Data Enrichment
Root Cause Analysis
MDT reports from the network
Conversion of MDT data into REMs
REMs with different UE density
Data enrichment using image inpainting methods
Enriched REMs
Cell-01
Cell-02
Cell-03
Cell-04
Cell-05
Cell-N
Total no. of cells (N) = 72
Performs multi-fault classification
Performs feature extraction
XGBoost
CNN Model
Hybrid Model
Cell-01 : Normal
Cell-02 : Excessive Antenna Uptilt
Cell-03 : Excessive Antenna Downtilt
Cell-04 : Cell outage
Cell-05 : Lower Transmission Power
Cell-N : Normal
System Output

Fig. 2.1: Proposed HYDRA Framework for Root-Cause Analysis of Multi-Fault in Multiple BSs, based on Image Enrichment and Hybrid Deep Learning.

2.4.1 Network Topology

The root-cause analysis framework I consider is designed for a real network but due to the unavailability of real data, a realistic commercial RF planning and optimization tool, **Forsk Atoll**[30] is used to generate and collect MDT reports. The simulated network topology considers an area from Brussels City, Belgium as shown in Fig. 2.2. I consider 15 different clutter types based on environmental conditions and terrain profiles. Aster propagation (advanced ray-tracing) is used as a propagation model because of its ability to better capture the idiosyncrasies in the environment as compared to empirical propagation models. I use the same locations and configuration parameters of BS used by a real network provider for its deployment in Belgium. Table 2.1 reports these settings. Therefore, the obtained coverage data can be assumed to be a very close representation of the ground truth of the MDT reports in the area used in the simulation. The area of simulation is 13.292 km^2 with 24 macrocell BSs (72 cells) to generate data with multiple fault classes in multiple BSs simultaneously.

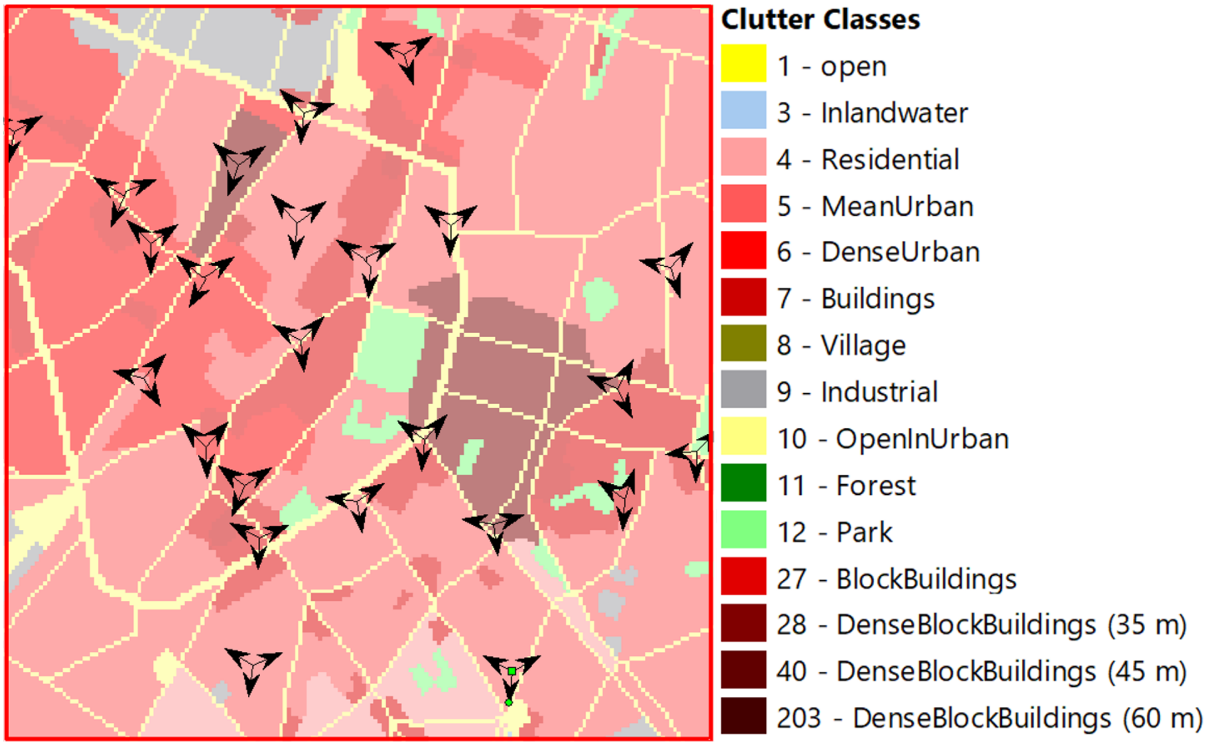


Fig. 2.2: Network topology and geographical clutter information used in the simulator for generation of synthetic MDT data.

2.4.2 Data Acquisition

I acquired MDT reports with 4 highly used fault classes in the literature for root cause analysis and self-healing frameworks: cell outage, low transmission power, excessive antenna uptilt, and excessive antenna downtilt [18, 31]. Figure 2.3 presents a visualization using SINR maps of different fault classes when induced on a selected cell in the designed network in the simulator. Fig. 2.3a represents a normal coverage scenario and the impact of other fault classes on cell coverage is illustrated in Fig. 2.3 (b-e). The parameter configuration of the 4 fault classes is described as follows:

1. **Cell Outage (CO):** To simulate cell outage, I deactivate the transmitter on a selected site in the simulator. This simulates a no-coverage fault scenario around that cell. Figure 2.3b presents the CO scenario for the highlighted cell.

2. **Low Transmission Power (LTP):** The maximum transmission power is 43 dBm for a normal BS in our designed network based on the recommended value by [10]. I

Table 2.1: Network Scenario Settings

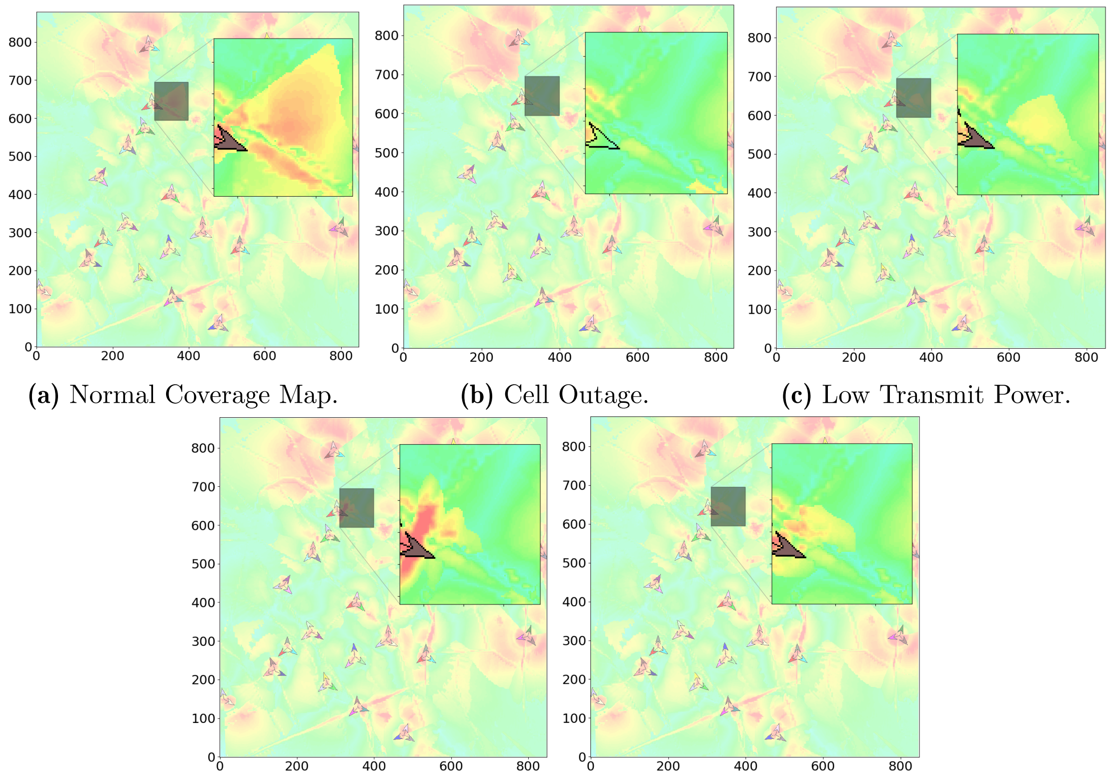
Network parameters	**Values**
Network layout	24 macro BSs (eNodeBs)
Sectors per BS	3 sectors/cells per BS
Carrier frequency	2100 MHz
Simulation area	13.292 km^2
Bin size	30m × 30m
Antenna height	Actual site heights
Propagation model	Aster propagation model (ray-tracing)
Clutter types	15 classes
Maximum transmission power	43 dBm
Cell individual offset (CIO)	0 dB
Antenna tilt	0^o
Antenna gain	18.3 dBi
Geographical information	Digital Terrain Model (ground heights) + Digital Land Use Map (clutter classes)

simulate the LTP fault scenario by reducing the maximum transmit power of a cell to 25 dBm. Figure 2.3c shows an LTP scenario.

3. **Excessive Antenna Downtilt (EAD):** To induce excessive antenna downtilt I change the tilt value from 0^o to 20^o. Figure 2.3d presents an EAD scenario.

4. **Excessive Antenna Uptilt (EAU):** Normal antenna tilt is 0^o. I change the tilt value from 0^o to -20^o. The impact of EAU can be seen in 2.3e for a selected cell.

To ensure the practicality of the HYDRA, while generating simulated MDT reports, I randomly select four cells out of 72 cells in total and induce a random fault in them through a different independent random process. In this way, not only I can have different cells (based on location in the network) in each MDT report but also different types of fault (CO, LTP, EAD, or EAU). I have 19933 different MDT reports of the network, each



(a) Normal Coverage Map. **(b)** Cell Outage. **(c)** Low Transmit Power. **(d)** Excessive UpTilt. **(e)** Excessive DownTilt.

Fig. 2.3: REMs presenting different network conditions. (a) Normal (b) Cell outage (c) Low transmission power, this image is showing when transmission power drops to 25dBm (d) Excessive antenna uptilt, this is +20^o tilt (e) Excessive antenna downtilt, -20^o tilt.

having 4 anomalous and 68 normal cells and each anomalous cell with a different fault.

I then convert raw MDT reports into SINR REMs. The REMs built from MDT reports are expected to be sparse by varying degrees depending on the user density, time interval used to aggregate MDT reports in a bin, and size of the bin [32]. I model this practical constraint by creating REMs with varying degrees of sparsity as shown in Fig. 2.4. To overcome the errors in fault detection and diagnosis caused by the sparsity in REMs in this report I leverage image inpainting to enrich the sparse REMs. To the best of our knowledge image inpainting is not used in the cellular domain because there are just a few published works, where MDT reports are used in the form of REMs. The available literature which considers REM-based outage diagnosis [27, 28], does not consider sparsity. This study is the first to investigate the impact of data sparsity and provide a detailed performance comparison of state-of-the-art inpainting techniques to address the practical problem of sparsity in MDT data and REMs in Section 3.3.1.

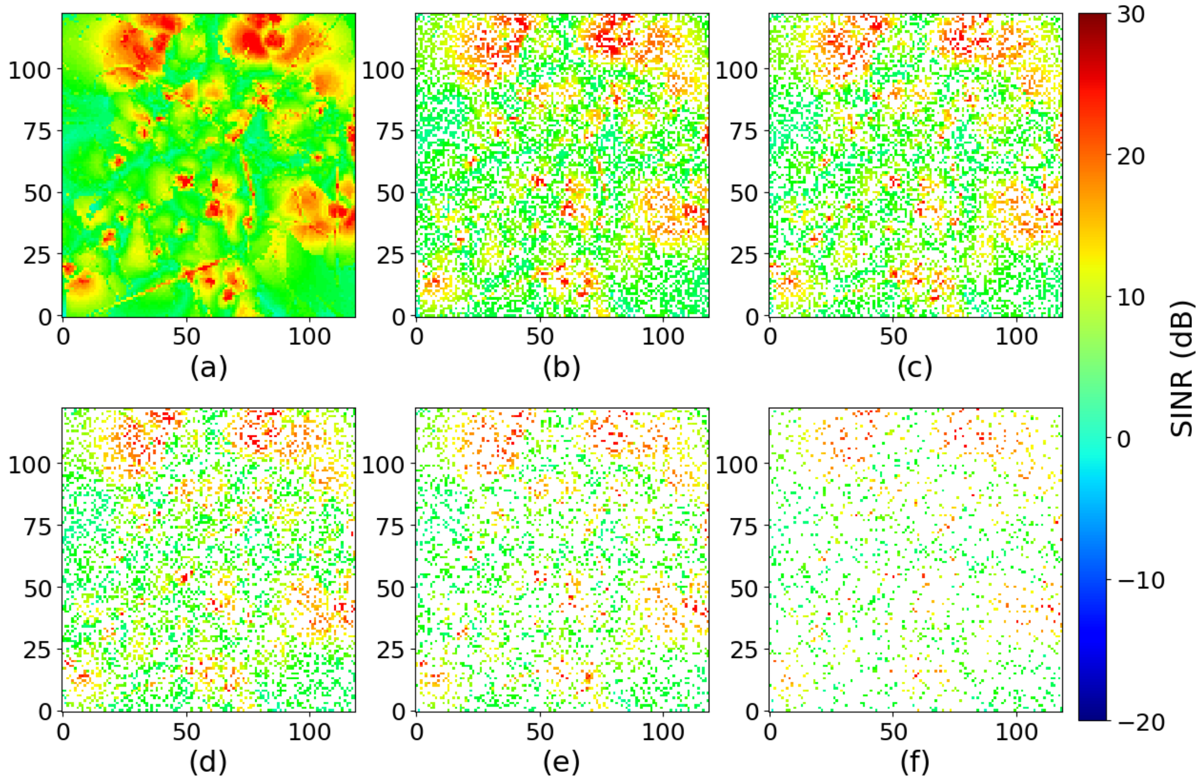


Fig. 2.4: REMs with various MDT report densities (a) Complete REM (203 UEs/cell (1101 UEs per km^2)). (b) 100 UEs/cell (550 UEs per km^2). (c) 80 UEs/cell (440 UEs per km^2). (d) 60 UEs/cell (330 UEs per km^2). (e) 40 UEs/cell (220 UEs per km^2). (f) 20 UEs/cell (110 UEs per km^2)

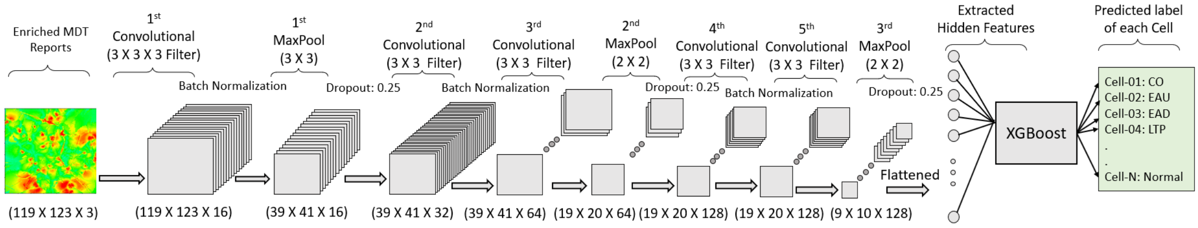


Fig. 2.5: Proposed hybrid deep learning model. CNN extracts hidden features from REMs and XGBoost performs the classification operation.

2.5 Data Enrichment using Image Inpainting

In this section, I present intuition and implementation details of block II (data enrichment) and block III (root cause analysis) of the HYDRA framework explained in Figure

3.1. MDT reports in a real network are expected to be sparse due to reasons such as low UE density [32]. There are various methods available to enrich sparse MDT data such as interpolation [33], regression clustering [34], and kriging [35]. While these methods work well for the enrichment of numerical data, our goal is to ultimately build REMs i.e., images. Therefore, instead of classical interpolation techniques image inpainting methods are more suited to our purpose here. Image inpainting methods based on fast marching methods [36], frequency selective reconstruction (FSR) [37], and Biharmonic equation provides a good balance between accuracy and time required to reconstruct a complete REM from a given sparse REM. Building on insights from these works, to address REM sparsity, I use image inpainting methods to enrich data before passing it on to root cause analysis block of Fig. 2.5. Through an extensive survey of the state-of-the-art image inpainting techniques, I select (based on accuracy and efficiency) the following methods to recover missing SINR values in the REMs. A comparative analysis of these methods on a sparse dataset (100 MDT reports/call) in the cellular networks domain is given in Table 2.2.

2.5.1 Frequency Selective Reconstruction (FSR)

FSR reconstructs missing SINR values using Fourier basis functions from available neighboring SINR values in the REM. This is a computationally expensive method but is highly parallelizable and with the use of GPUs can achieve significantly accurate results in considerably less time [37].

2.5.2 Biharmonic Equations

This method estimates the missing pixels using fourth-order partial differential equations. This is a computationally very expensive process due to the computation of multiple derivatives to estimate missing SINR values. It is quite accurate on small datasets but requires a lot of time to reconstruct bigger datasets.

2.5.3 TELEA

TELEA uses the principles of the fast marching method [36] to reconstruct missing SINR values in the REM using a normalized weighted sum computed from known neighborhood SINR values.

2.5.4 Navier-Stokes

This method reconstructs the missing SINR values in the REM using the principle of heuristic based on fluid dynamic equation (Navier-Stokes) [38]. The reconstruction process starts on the edges and keeps on filling the missing SINR data towards the center of the REM.

To evaluate the performance of data enrichment methods I use root mean square error (RMSE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR) as performance metrics. It is evident from Table 2.2 that FSR outperforms the rest of the inpainting methods. In this work, RMSE is the most important metric because that represents the difference between the estimated (inpainted) value of SINR against the ground truth.

Table 2.2: Performance Evaluation of Data Enrichment Methods

Method	**RMSE (dB)**	**SSIM (%)**	**PSNR (dB)**
FSR	**8.6**	**90**	**22.81**
Biharmonic Equation	9.3	87	21.23
Navier-Stokes (NS)	9.45	86	20.8
TELEA	9.7	85	20.84

2.6 Root Cause Analysis based on Hybrid Deep Learning

In this section, I elaborate on the **root cause analysis** block in the HYDRA framework (right most block in Fig. 3.1). I also explain the intuition behind using a hybrid deep learning model, implementation, and performance metrics to compare the proposed model against the widely used techniques for root cause analysis.

2.6.1 Why a hybrid deep learning model?

The rationale behind using a hybrid model stems mainly from the fact that in cellular networks, the availability of training data is still a challenge even in the age of big data [32]. For this reason, I cannot use CNN alone as it requires large training data and is computationally more expensive as compared to classical ML models. On the other hand, classical ML methods when used alone are less robust to noisy data (noise is induced due to the application of image inpainting to enrich sparse MDT data) as compared to CNN. To overcome this challenge, I use a hybrid approach where I use CNN for hidden feature extraction from REMs to take advantage of the robustness it offers towards noisy images [39, 40]. Then I pass the extracted features to XGBoost which is computationally efficient as compared to the classification layer of CNN [41] and provides better accuracy when used as a hybrid [42, 43]. XGBoost takes the extracted features and performs fault diagnosis. Figure 2.5 provides a detailed elaboration of the proposed hybrid model.

2.6.2 Minimization of drive test reports:

MDT reports are introduced by 3GPP in release 10, these reports have several features e.g. user location and network quality of service based on certain KPIs like RSRP, RSRQ, and SINR to name a few [10]. In this research, I use SINR one of the available KPIs in MDT reports. The major advantages that MDT reports offer are a reduction of human intervention, a reduction in operational expenditure as well as the reduction in time-inefficiency arising from offline configurations required for coverage-related faults detection and diagnosis. These features make MDT reports a key enabler for ML-based self-organization envisioned for emerging cellular networks.

2.6.3 Convolutional neural network:

A CNN is a class of neural networks that specializes in processing data that has a grid-like topology, such as an image. CNN uses a dynamic kernel, and convolutional layers instead of fully connected layers, which reduces the number of weights in each layer and hence requires less computation time, which makes CNN computationally more efficient. I used CNN for feature extraction from REMs that are image representations of network coverage maps. The CNN architecture in Figure 2.5 is our top-performing model. The loss vs epoch graph of this model is present in Figure 2.6, this graph is based on mean loss values of 5-fold cross-validation for 50 epochs. This graph shows the lack of overfitting as well as underfitting in the model training process as the gap between training and validation loss converges after 50 epochs. A detailed explanation of each feature extraction function is given as follows:

1. Convolution: convolution extracts important hidden features e.g. boundary edges from the input coverage map. In this study, boundaries are very important to distinguish between coverage regions of different sites. The dimensions of the output matrix of convolution operation are defined as Equation 2.1 and 2.2.

   $$O_r = \frac{(I_r - F + 2P)}{S} + 1 \tag{2.1}$$

   $$O_c = \frac{(I_c - F + 2P)}{S} + 1 \tag{2.2}$$

   where O_r , O_c , I_r and I_c represent the number of rows and columns of the output and input matrix respectively, while F, P and S represents the size of kernel, padding, and length of the stride.

2. Batch Normalization: To accelerate the learning of CNN and to address internal covariate shift, batch normalization is used [44]. This transformation normalizes the input to a layer by maintaining the mean and standard deviation close to 1 and

0 respectively.

3. Pooling: Pooling is used for down-sampling of the feature matrix which reduces its sensitivity and makes the feature extraction process robust to changes. I use max pooling to ensure the presence of the most activated features.

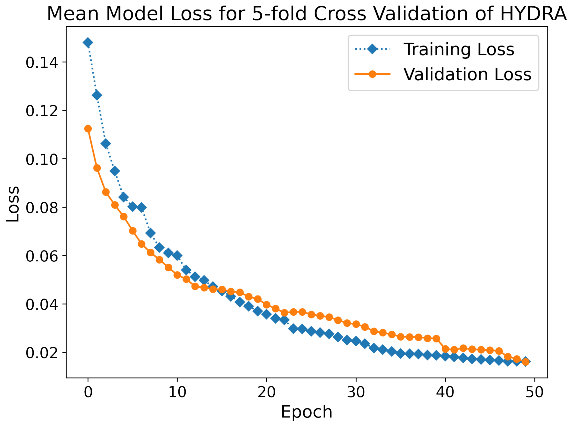


Fig. 2.6: Mean model loss for 5-fold cross-validated HYDRA

2.6.4 Extreme gradient boosting:

XGBoost is a decision-tree-based ensemble ML algorithm that uses a gradient boosting framework [45]. The tree-based nature and gradient boosting make XGBoost yield superior results using fewer computing resources in the shortest amount of time. Time and computation efficiency is the reason I use XGBoost as a classification model of HYDRA instead of other ML models like artificial neural networks (ANN) [41], RNN, MLP, or extreme learning machines (ELM) [42, 43]. Furthermore, it gives more accurate results as compared to SVM and random forest. A detailed performance analysis of XGBoost against SVM and random forest is present in Section 3.4.

2.7 Results and Comparative Analysis of Proposed Framework

2.7.1 Performance metrics:

As I propose a solution for a multi-label multi-class problem, unlike a simple classification problem, it requires special performance measuring metrics [46, 47], due to the biased nature of data towards normal class. Hence, I choose the following performance metrics to evaluate the HYDRA.

1. **F1-Score (F1):** F1-score combines both precision and recall in one metric by taking their harmonic mean. This provides a more realistic performance analysis because it minimizes the chance of bias towards the majority class in the data. F1-score of a class is defined by 2.3

$$F_1 = \frac{Tp}{Tp + \frac{1}{2}(Fp + Fn)} \tag{2.3}$$

where *Tp* is true positive (%), *Fp* is false positive (%), and *Fn* is false negative (%) of the respective class.

2. **Exact Match Ratio/Subset Accuracy (EMR):** According to EMR, a diagnosis made by the model will be correct only if the network condition of all the cells in the network are diagnosed correctly. In this study I have a network designed with 72 cells, even if the network condition of 1 out of 72 cells for a given REM is predicted incorrectly, that REM will be considered as an incorrect prediction. This is considered a very strict performance metric, but to present a critical performance analysis I include it in our results. EMR is defined by Equation 2.4.

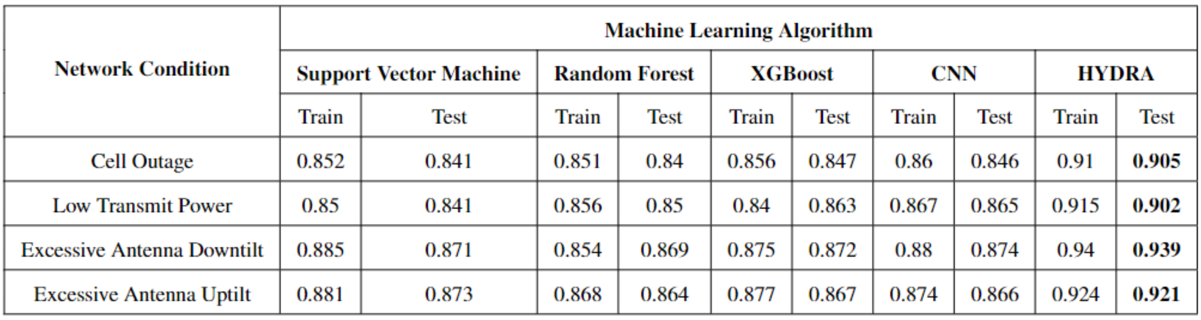
$$EMR = \frac{1}{N}\sum_{i=1}^{N} I(P_i = T_i) \tag{2.4}$$

where I is a proposition function that returns 1 if all 72 cells are correctly diagnosed else returns 0.

3. **Proportionally Correct Diagnosis (PCD):** I present a worst-case performance analysis of HYDRA using PCD, because even if it diagnoses 2 out of 4 faults correctly that can help compensate half of the anomalous cells in the network. So, besides EMR I present proportionally correct results. PCD presents the percentage of cases when 4, 3, 2, or 1 faults (out of 4) are correctly diagnosed from a given REM.

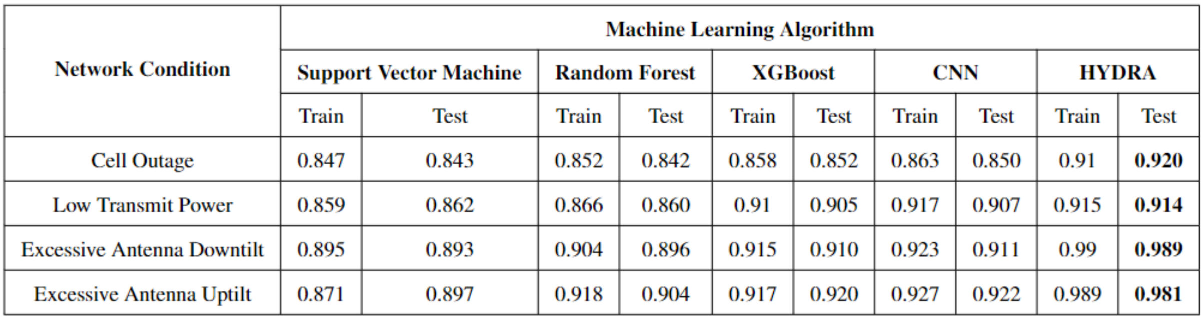
2.7.2 Results

I used a 5-fold cross-validation method [48] for the performance analysis presented in this section, because cross-validation ensures that each sample from the original dataset has an equal chance of appearing in the training and validation set. Which is a well-suited approach when we have limited input data. I compare the performance of HYDRA against the state-of-the-art ML methods used for the detection and diagnosis of outages in the literature. These include SVM [15, 49, 50], RF [27, 28], standalone XGBoost and standalone CNN [27]. I evaluate HYDRA with different UE densities to analyze its efficacy in realistic settings (i.e., robustness to sparsity of MDT reports in a cell/area). Figure 2.10 presents the performance evaluation of state-of-the-art ML/DL models for sparse data (considering various UE densities) and enriched data (enhanced using FSR image inpainting as explained in Section 3.3.1). Figures 2.10b - 2.10f provide a comparative analysis of sparse and enriched data using EMR as a metric.



Network Condition	**Machine Learning Algorithm**									
	Support Vector Machine		**Random Forest**		**XGBoost**		**CNN**		**HYDRA**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Cell Outage	0.852	0.841	0.851	0.84	0.856	0.847	0.86	0.846	0.91	**0.905**
Low Transmit Power	0.85	0.841	0.856	0.85	0.84	0.863	0.867	0.865	0.915	**0.902**
Excessive Antenna Downtilt	0.885	0.871	0.854	0.869	0.875	0.872	0.88	0.874	0.94	**0.939**
Excessive Antenna Uptilt	0.881	0.873	0.868	0.864	0.877	0.867	0.874	0.866	0.924	**0.921**

Fig. 2.7: F-score comparison of ML algorithms for different network conditions on sparse training and testing data (100 MDT reports/cell)



Network Condition	Machine Learning Algorithm									
	Support Vector Machine		**Random Forest**		**XGBoost**		**CNN**		**HYDRA**	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Cell Outage	0.847	0.843	0.852	0.842	0.858	0.852	0.863	0.850	0.91	**0.920**
Low Transmit Power	0.859	0.862	0.866	0.860	0.91	0.905	0.917	0.907	0.915	**0.914**
Excessive Antenna Downtilt	0.895	0.893	0.904	0.896	0.915	0.910	0.923	0.911	0.99	**0.989**
Excessive Antenna Uptilt	0.871	0.897	0.918	0.904	0.917	0.920	0.927	0.922	0.989	**0.981**

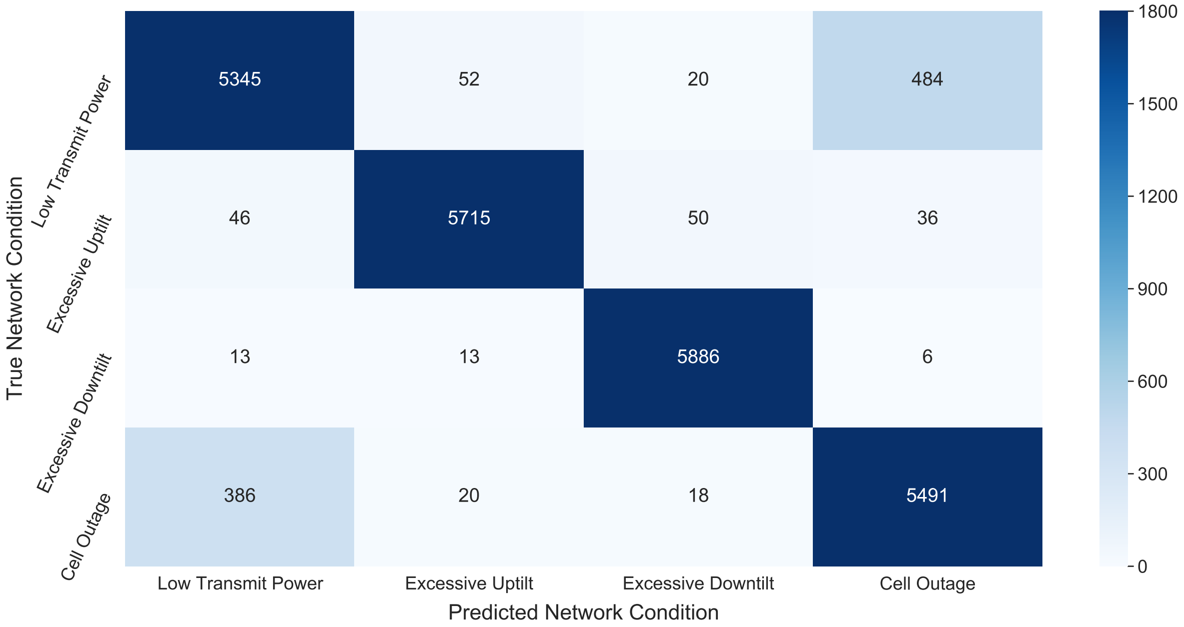
Fig. 2.8: F-score comparison of ML algorithms for different network conditions on enriched (using FSR image inpainting) training and testing data

2.7.3 Performance Analysis

The insights coming from the performance analysis of selected models on sparse and enriched data can be summarized as follows:

1. **Performance for each fault class based on F-1 score:** Figure 2.7 presents a comparison of HYDRA based on F-1 score for each fault class against SVM, random forest, XGBoost, and standalone CNN on sparse data. This can be seen that HYDRA outperforms other methods based on F-1 score with a consistent F-1 score of above 0.90 for each network fault. Figure 2.8 provides fault diagnosis results on enriched data. A significant improvement in F-1 score can be seen for both antenna uptilt and antenna downtilt but there is relatively less improvement for cell outage and low transmit power, on enriched data. The reason for no improvement in cell outage and low transmit power is because, a cell going through a complete or partial outage will not have much coverage even after the data enrichment process, so the diagnosis accuracy remains almost the same.

2. **Performance analysis based on confusion matrix:** Figure 2.9 presents thorough insights about the performance of HYDRA for each network fault. Two network conditions that are not well distinguished are lower transmit power and cell outage, both faults are confused by HYDRA as one another. That makes sense if we look at Fig. 2.3b and Fig. 2.3c cell outage and low transmit power affect the coverage of

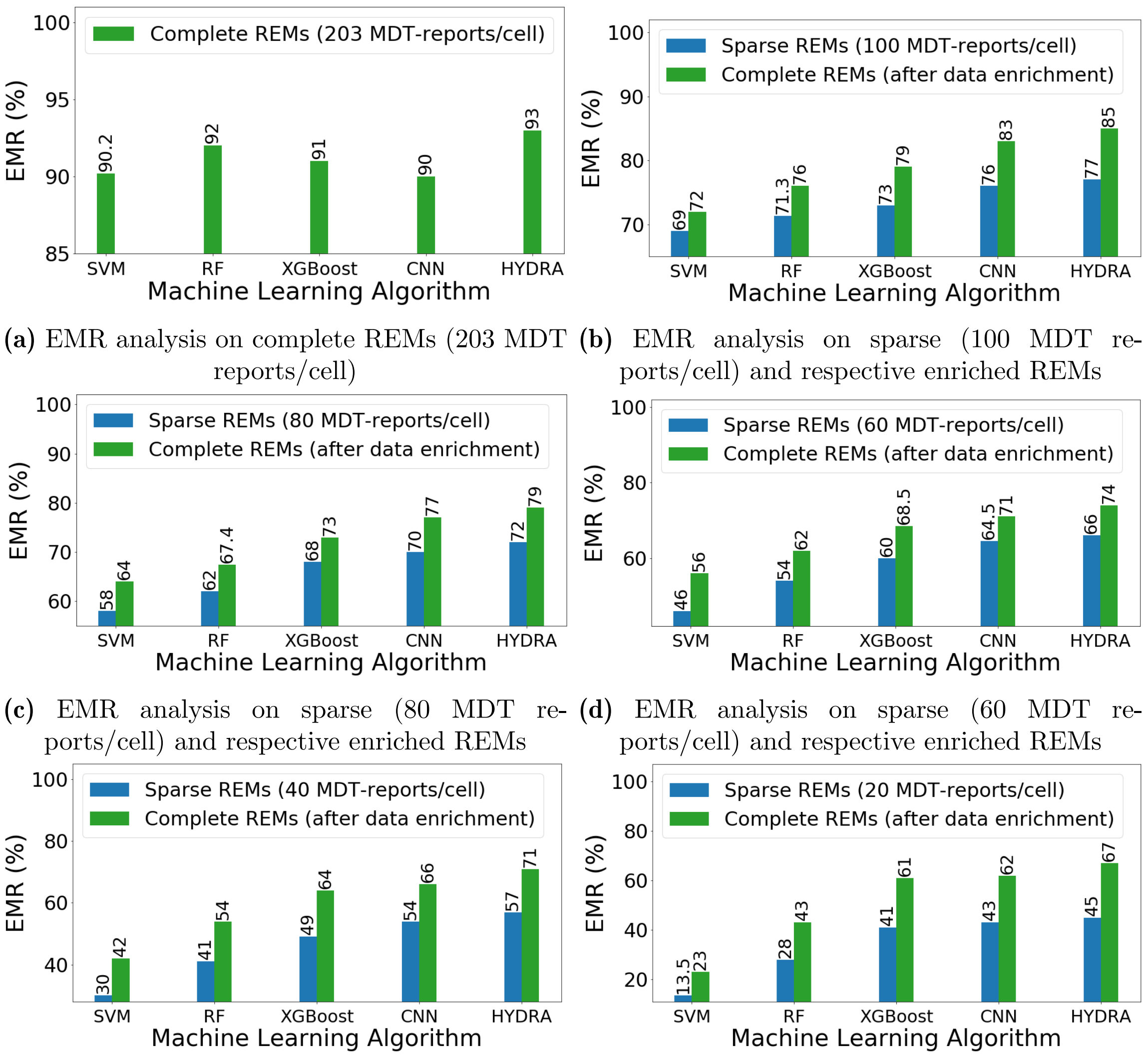


True Network Condition \ Predicted Network Condition	Low Transmit Power	Excessive Uptilt	Excessive Downtilt	Cell Outage
Low Transmit Power	5345	52	20	484
Excessive Uptilt	46	5715	50	36
Excessive Downtilt	13	13	5886	6
Cell Outage	386	20	18	5491

Fig. 2.9: Mean confusion matrix for multi-fault diagnosis for HYDRA using 5-fold cross-validation.

a cell in a similar fashion. Therefore, it is relatively more challenging to distinguish between cell outage and low transmit power as compared to excessive antenna uptilt and downtilt. Resolving the high rate of confusion between these two specific faults i.e. low transmission power and cell outage for other reasons, can be an interesting topic for a dedicated future study as it may have to leverage Bayesian analysis and historical fault logs to establish priors.

3. **Performance on sparse REMs:** From Fig. 2.10a, we can observe that SVM and RF perform slightly better than CNN and XGBoost on complete REMs (i.e., when the number of users is large enough to send MDT report from each bin of the area under consideration). This justifies the popularity of SVM for self-healing in the literature [15]. However, a drastic drop in diagnosis performance can be seen for SVM and RF on sparse data. i.e., EMR drops from 90.2% to 69% and from 92% to 71.3% respectively, as the density of MDT reports drop from 203 in Fig. 2.10a to 100 per cell in Fig. 2.10b. The downward trend in performance continues as the number of reports decreases per cell. EMR of SVM drops to 13.5% when MDT reports per cell



(a) EMR analysis on complete REMs (203 MDT reports/cell)

(b) EMR analysis on sparse (100 MDT reports/cell) and respective enriched REMs

(c) EMR analysis on sparse (80 MDT reports/cell) and respective enriched REMs

(d) EMR analysis on sparse (60 MDT reports/cell) and respective enriched REMs

(e) EMR analysis on sparse (40 MDT reports/cell) and respective enriched REMs

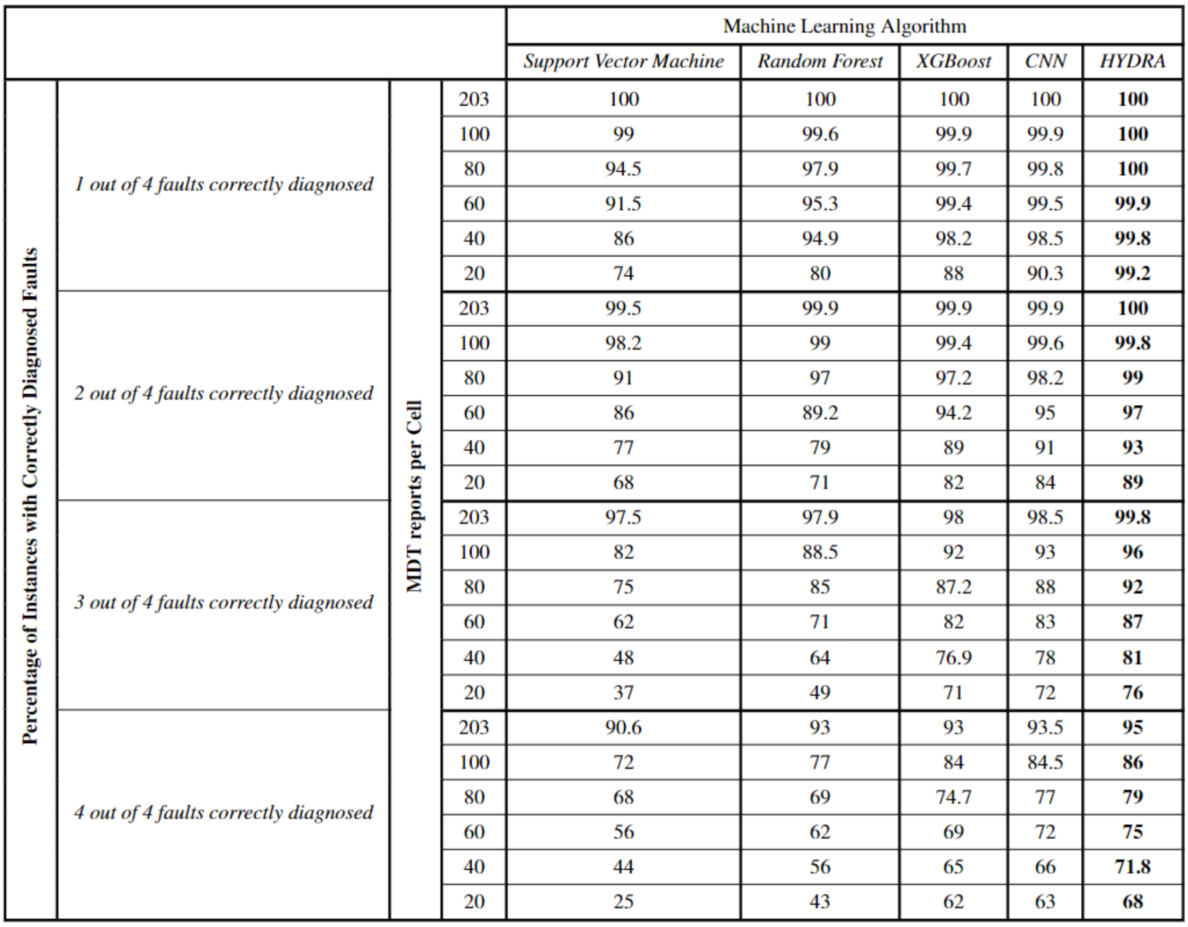
(f) EMR analysis on sparse (20 MDT reports/cell) and respective enriched REMs

Fig. 2.10: EMR based performance analysis on sparse and respective enriched REMs. Note: (REMs with 203 MDT reports per cell are full coverage maps, so do not require enrichment).

decrease to 20 in Fig. 2.10f. In contrast, the results show that HYDRA is relatively robust to the REMs sparsity and can diagnose faults with an EMR of 45%, even when REMs are extremely sparse with just 20 MDT reports/cell.

4. **Performance on enriched REMs:** From Fig. 2.10f HYDRA shows a promising improvement in enriched data as compared to sparse data. Figure 2.10f, shows that HYDRA can diagnose faults with an EMR of 67% on data enriched from highly sparse (20 MDT reports/cell) REMs. This is a 22% improvement in EMR

achieved using data enrichment and hybrid deep learning-based diagnosis proposed in HYDRA.



			Machine Learning Algorithm				
			Support Vector Machine	*Random Forest*	*XGBoost*	*CNN*	*HYDRA*
Percentage of Instances with Correctly Diagnosed Faults	*1 out of 4 faults correctly diagnosed*	203	100	100	100	100	**100**
		100	99	99.6	99.9	99.9	**100**
		80	94.5	97.9	99.7	99.8	**100**
		60	91.5	95.3	99.4	99.5	**99.9**
		40	86	94.9	98.2	98.5	**99.8**
		20	74	80	88	90.3	**99.2**
	2 out of 4 faults correctly diagnosed	203	99.5	99.9	99.9	99.9	**100**
		100	98.2	99	99.4	99.6	**99.8**
		80	91	97	97.2	98.2	**99**
		60	86	89.2	94.2	95	**97**
		40	77	79	89	91	**93**
		20	68	71	82	84	**89**
	3 out of 4 faults correctly diagnosed	203	97.5	97.9	98	98.5	**99.8**
		100	82	88.5	92	93	**96**
		80	75	85	87.2	88	**92**
		60	62	71	82	83	**87**
		40	48	64	76.9	78	**81**
		20	37	49	71	72	**76**
	4 out of 4 faults correctly diagnosed	203	90.6	93	93	93.5	**95**
		100	72	77	84	84.5	**86**
		80	68	69	74.7	77	**79**
		60	56	62	69	72	**75**
		40	44	56	65	66	**71.8**
		20	25	43	62	63	**68**

Fig. 2.11: Performance analysis showing performance of ML algorithms for correctly diagnosing a proportion of faults

5. **Proportionally correct diagnosed faults on enriched data:** Furthermore, a worst-case performance analysis is presented in Fig 2.11, showing the proportion of correctly diagnosed faults using different MDT report densities. It can be seen from Fig 2.11 that HYDRA can correctly diagnose at least 2 faults in the network with 100% accuracy for complete REMs and with 89% accuracy from highly sparse (20 MDT reports/cell) MDT data. Furthermore, the proposed model can reliably diagnose 3 of the 4 faults 76% of the time from highly sparse MDT data. It can identify at least one fault in around 99% of the maps enriched from just 20 MDT reports/cell sparse data.

CHAPTER 3

Screening Various Respiratory Conditions from Noisy Acoustic Data

3.1 Introduction

In the previous chapter, HYDRA is presented, a framework for coverage anomaly detection from noisy incomplete image data. In this chapter, I present diagnostics using acoustic data. The modality of data is changed but the focus of the research remains the same, which is diagnostics from noisy and incomplete data. I leverage the potential benefits of cough-based AI-enabled COVID-19 screening. Its benefits are multi-fold due to its non-invasive nature, rapidness, and cost-effectiveness. However, there are also many challenges highlighted by recent research that must be addressed [51, 3]. One major challenge highlighted in [51, 3] results due to the lack of standardization of the cough datasets that are being used for the training of AI models. This challenge is a superset of many underlying challenges that result in heterogeneity of audio data at different fronts. To name a few, 1) noise in the data due to the collection in different environmental settings (hospital, home, outdoor, etc.), 2) different sampling rates (48kHz, 44.1kHz, 22.05kHz, etc.) at which the data is collected that contribute to the quality of audio, and 3) different data collection devices (smartphone, laptop, desktop, etc.). Overall, while cough-based AI-enabled COVID-19 screening shows promise as a potential tool in the fight against the pandemic, careful consideration must be given to these challenges in order to ensure that it can be implemented effectively and safely.

To address these challenges, the respective research community needs to focus on two major research and development directions, 1) highlight the potential challenges that compromise the robustness of published solutions, and 2) quantify the impact of those challenges on the performance of AI-based solutions. Only through quantification of the effects of specific challenges, reliable AI-based solutions be developed and effectively implemented at scale.

3.1.1 Related Work

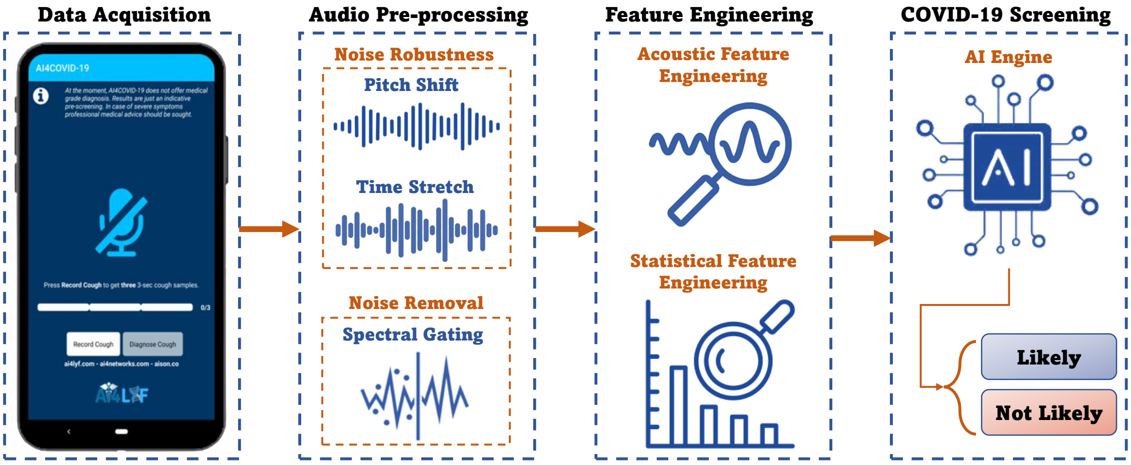
In the recent past multiple studies have been published highlighting the challenges and future directions for reliable AI-enabled COVID-19 diagnosis [51, 3, 52]. However, the current literature lacks to provide a comprehensive quantification of these challenges on the performance of COVID-19 screening. This is primarily due to the unavailability of standardized as well as adequate clinical data that can be used to quantify the impact of these challenges.

However, researchers in [53] present an analysis of their framework on cough data collected at different sampling rates. They considered 4 kHz, 8 kHz, and 48 kHz sampling rates, and based on their results higher sampling rates help yield higher diagnostic accuracy for the ML models. This is a promising insight that leads toward a more robust ML-based solution for cough-based disease diagnostics.

3.1.2 Gaps in the Current Literature

Based on the available literature summarized above there does not exist a robust AI-enabled solution that quantifies the following prominent challenges in cough-based COVID-19 screening.

1. **Analysis of different acoustic and statistical features used in the ML models training.**
2. **Analysis of different cough sampling rates, because different cough collecting devices have different hardware and software capabilities.**
3. **Analysis of device heterogeneity based on the built-in hardware of the devices from different make, model, and brand.**



At the moment, AI4COVID-19 does not offer medical grade diagnosis. Results are just an indicative pre-screening. In case of severe symptoms professional medical advice should be sought.

Press **Record Cough** to get **three** 3-sec cough samples.

Fig. 3.1: RAISED: A framework for Reliable AI-enabled COVID-19 Screening.

To address the above-mentioned prominent challenges and limitations in the literature, I propose RAISED Reliable AI-enabled Screening of COVID-19, which quantifies the impact of challenges mentioned above and also highlights several new research directions for more reliable AI-enabled cough-based disease diagnostics.

3.1.3 Contributions and organization

The following are the key components and contributions of RAISED:

1. This dissertation presents RAISED, a first-of-its-kind framework that quantifies several of the prominent challenges in AI-enabled COVID-19 screening using cough. This includes the quantification of more than 100 acoustic and 10 statistical features published in our previous work [54], that impact the reliability of such solutions.

2. RAISED demonstrate the impact of several commonly used cough sampling rates (4 kHz, 8 kHz, 22.05 kHz, 44.1 kHz, and 48 kHz) on the screening performance of RAISED to highlight the need for reliable and robust screening solutions. The results show this is a challenge that needs to be considered while designing AI-enabled solutions..

3. This research provides a detailed analysis of pre-processing methods used for noise removal and robustness. The results demonstrate that removing the environmental noise using the traditional speech recognition methods, impacts the performance of RAISED negatively which opens new research directions to explore, discussed in detail in Section 3.4.

4. RAISED highlight and quantify the impact of heterogeneity in the cough data collection devices based on the hardware components like microphone variance. The insights from the results show some promising research questions that need to be answered for reliable AI-enabled disease screening.

3.2 Data and Feature Acquisition Methods

3.2.1 Data Acquisition

I acquired the cough samples using our indigenous mobile and web application that I used for our previous work [2]. Each sample is a 3-second long, uncompressed PCM 16-bit cough stream, captured at 48 kHz. Healthy and COVID-19 samples are collected using the same app hence maintaining the same audio features and characteristics which is essential for our analysis in this research, cough features are given in Table 3.1.

3.2.2 Cough Features for AI-enabled Screening

We published a comprehensive list of audio features that can be leveraged to build a robust AI-enabled solution for cough-based diagnostics in our previous work [54]. Out of the published list, I selected 14 acoustic and 10 statistical features to quantify the impact of these features on the AI-based COVID-19 screening. I present our analysis in detail in Section 3.4 by highlighting the impact of each feature individually as well as combined for the sake of robustness.

Table 3.1: Cough Data Summary

Cough Parameters	**Values**
Total samples	792 (400 Normal, 392 COVID-19)
Sample duration	3 seconds
Sampling rate	48 kHz
Bit rate	PCM 16 bits
Channels	Stereo
Gender distribution	Male: 622, Female: 170
Age distribution	Min Age: 19 Years, Max Age: 74 Years
User consent	Electronic consent

3.2.3 Cough Collection Device Heterogeneity

A wide variety of devices are used to collect the cough samples. The diversity of data collection devices is essential for the development of RAISED because each device has different hardware which leads to self-noise in the collected cough data. I present the device diversity analysis in Section 3.4 considering several different smartphone and laptop devices.

3.3 System Design of RAISED

RAISED shown in Fig. 3.1 is designed with robustness and reliability as the major motivations. To achieve robustness RAISED has the following components explained below:

3.3.1 Audio Pre-processing

Recording cough using smart devices in public settings introduces environmental noise and reverberation, which can contaminate the recordings and compromise the accuracy of ML models for disease diagnosis. While reverberation can be characterized by sound propagation models in indoor and outdoor environments [55], environmental noise can vary significantly based on the surroundings. The amplitude, frequency distribution, and signal-to-noise ratio of the noise in each recording can be unique. Furthermore, even in the same ambient setup, variations in microphone-to-mouth positioning, such as distance and angle, can result in noise variations in the recordings. Although filtering and smoothing algorithms can be employed to reduce noise, this often leads to the elimination of high-frequency components in the recordings. However, these high-frequency components are not always noise-induced and can be crucial for accurate cough-based diagnosis, thus they cannot be entirely removed. Additionally, ML models trained on noiseless data or data with limited noise scenarios tend to overfit and lack generalization to real-life settings where a multitude of environmental noises exist. Therefore, the challenge lies in developing noise-aware machine learning models that exhibit robustness to environmental distortions during both the training and inference stages [56].

RAISED leverages several methods to overcome the challenge of noise in the cough data. Firstly using the traditional approaches in the speech recognition domain to remove static environmental noise. Secondly, provided the cough sound is itself more like a noise than speech. Therefore I use noise robustness methods that can overcome the noise as well as the biases like age and gender in the data.

3.3.2 Feature Engineering

Building on our survey [54], I focus on acoustic and statistical feature engineering for cough-based screening. This is a relatively unexplored direction in the literature for AI-enabled COVID-19 diagnosis. Therefore I present a comprehensive analysis for both time and frequency domain features. Furthermore, I demonstrate the quantitative impact of each feature class individually as well as when combined with other features for the model training.

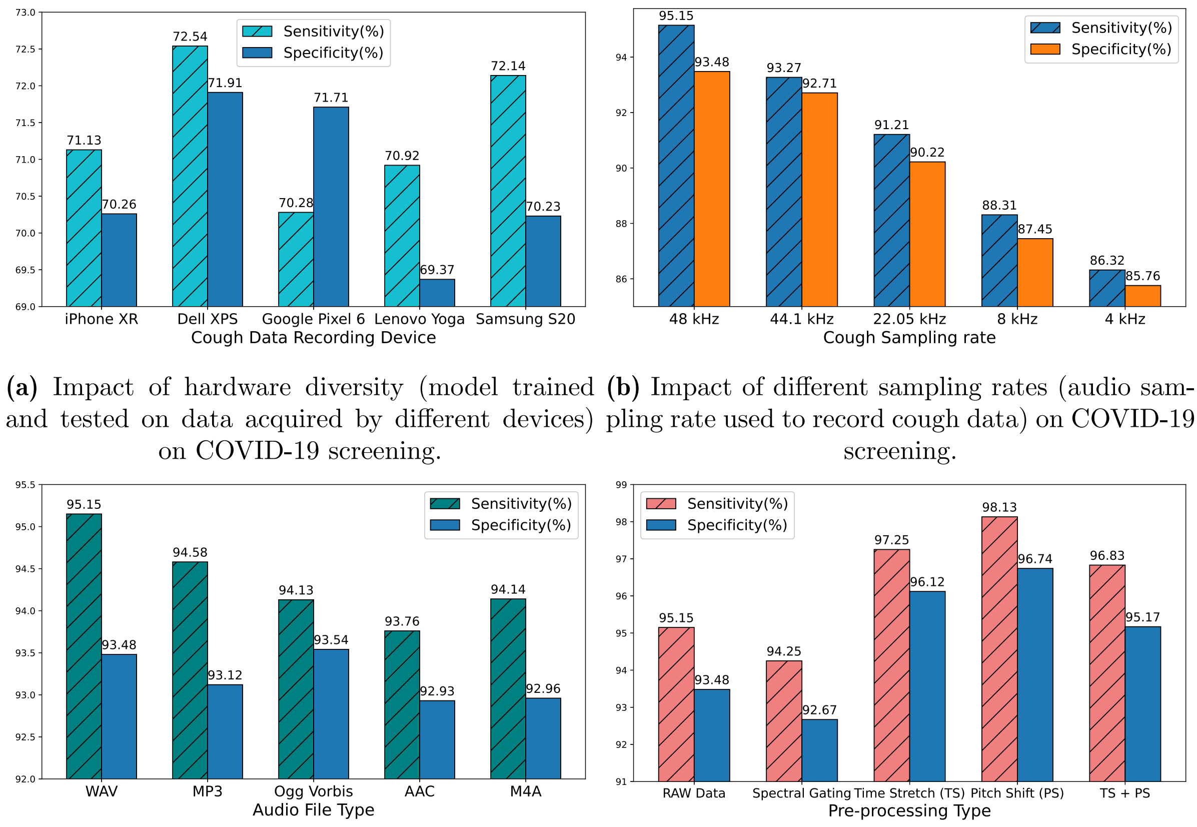
3.3.3 AI-engine for COVID-19 Screening

We use extreme gradient boosting (XGBoost) as our AI engine for RAISED, because of its efficiency and robustness to noisy data. XGBoost is inherently reliable for the robust classification of noisy data due to its ensemble learning approach, gradient boosting mechanism, and regularization techniques. These features enable XGBoost to effectively reduce the impact of noise, and improve the overall classification performance even in challenging sparse and noisy environments [57]. Moreover, it outperforms our previously published models for multi-class classification [2], demonstrating superior performance as a binary classifier.

3.4 Results and Comparative Analysis

3.4.1 Evaluation of Device Diversity

The ambulatory sound-based cough diagnostic can be influenced by the hardware diversity of audio recording devices across three levels. Firstly, there is variability in device types, including cellphones, laptops, microphones, and smartwatches. Secondly, even devices of the same type may differ due to various manufacturing brands such as Apple, Google, Samsung, and variations in specifications such as frequency response, phase response, sensitivity, noise level, sound pressure level, and signal-to-noise ratio. Thirdly, even devices with identical specifications, belonging to the same brand and model, can exhibit electro-acoustic variations stemming from inherent manufacturing process uncertainties of microphone chips [58]. In Fig. 3.2a I present the performance analysis of RAISED on test data collected using different devices. The results show a rapid decrease



(a) Impact of hardware diversity (model trained and tested on data acquired by different devices) on COVID-19 screening.

(b) Impact of different sampling rates (audio sampling rate used to record cough data) on COVID-19 screening.

(c) Impact of different file types (cough stored in different file formats for efficient transmission and manipulation) on COVID-19 screening.

(d) Impact of different pre-processing methods (used for sound-based data robustness and environmental noise reduction) on COVID-19 screening.

Fig. 3.2: Performance analysis of RAISED based on various software and hardware related data diversities, different (a) sampling rates (b) file types (compression rates) (c) pre-processing methods (d) device diversity, used to acquire training data.

in accuracy which is also seen in evaluation by other researchers [58]. This brings forth opportunities for further research, such as the development of ML models capable of generalizing across the electro-acoustic disparities among microphones. Another avenue involves training ML algorithms to accurately identify the diverse models and brands of recording microphones. By achieving this identification, it becomes feasible to mitigate the impact of microphone profiles on the recorded sounds.

3.4.2 Evaluation of Audio Sampling Rates

In addition to the hardware disparities among microphones, software characteristics, including the sampling rate, contribute to producing sound recordings with varying sizes and qualities. As a result, a trade-off between ML model accuracy and speed arises, involving the management of data quality versus size. I emphasize the challenge posed by the diversity of audio sampling rates by providing a quantified assessment of the impact of five distinct sampling rates on the performance of RAISED in Fig. 3.2b. Our findings based on results in Fig. 3.2b show though not significant but higher sampling rates yield higher accuracy. The results show a 10% decrease in the sensitivity and specificity of RAISED for a 12 times lower sampling rate. This is a promising step toward reliable screening and can be strengthened with further research in this direction.

3.4.3 Evaluation of Audio File types

In tandem with the sampling rate, file type is another software-based characteristic of the audio data, that contributes to the size and quality of the cough samples. In Fig. 3.2c I present the quantified impact of five different commonly used audio file types on the performance of RAISED. A dip in performance is evident as we move from uncompressed WAV file format to compressed file formats. This is primarily due to the quantization error as well as the data encoding used for the compressed file formats. Both quantization and encoding errors end up losing some latent features (MFCCs, band power, energy, etc.) in the audio which leads to a decrease in performance. Consequently, this opens up new avenues for research into audio compression techniques tailored specifically for cough data.

3.4.4 Evaluation of Noise Removal and Robustness Methods

To evaluate RAISED against noisy data I present a two-pronged approach, 1) remove the environmental noise from the cough before feeding it to the AI engine, 2) add more noise to the training data to make the AI engine robust to learn from the noisy data. Fig. 3.2d presents results from both approaches. I use spectral gating (SG) as a method to remove the environmental noise from cough but the performance goes down on the noise-removed data. This is counter-intuitive because clean data should intuitively yield better performance. Based on our analysis this must be due to two reasons, 1) SG is a common method to remove noise from speech data and cough is more like a noise than speech, so SG end up removing some cough in addition to static environmental noise 2) SG removes high-frequency components from the speech considering them noise and cough also has high-frequency components, which end up being removed and hence losing some of the important features that help the AI engine learn to distinguish between healthy and COVID-19 coughs.

In addition to the noise removal method, I also present the performance analysis of RAISED for noise robustness techniques in Fig. 3.2d. I use time stretch (TS) and pitch shift (PS) to add more noise to the cough samples. I train the AI engine on the synthetic noisy data to make it learn to classify COVID-19 and healthy coughs in the presence of noise. I select TS and PS keeping in mind this will also help mitigate the age and gender bias in the data which will lead to the overall robustness of RAISED. The results in Fig. 3.2d show both TS and PS improves the performance of RAISED. PS demonstrates greater improvement compared to TS, primarily because our dataset consists of approximately 80% male participants. Therefore, PS effectively addresses the bias arising from gender variations. Additionally, I provide results for a cascade of TS and PS as a pre-processing technique. While these results indicate an enhancement in performance compared to using the raw data, the improvement is lower compared to the individual performance achieved by TS and PS separately.

3.4.5 Evaluation of Acoustic and Statistical Features

In conjunction with several software and hardware challenges I also present a comprehensive analysis of cough features that are used for COVID-19 screening. The acoustic features extracted from cough samples play a crucial role in the analysis and classification of cough sounds. Various acoustic parameters are utilized to capture the distinctive

Table 3.2: Performance Analysis of Acoustic Features Used Individually for COVID-19 Screening

Acoustic Features Used	**Performance Metrics**	
	Sensitivity(%)	**Specificity (%)**
MFCC	87.59	86.91
MFCC Delta	76.93	76.13
MFCC Delta2	79.38	77.34
Melspectrogram	77.69	74.71
Zero Crossing Rate	61.14	62.18
Spectral Rolloff	62.02	64.12
Spectral Centroid	66.29	65.23
Spectral Bandwidth	64.7	63.27
RMS Energy	73.21	72.19
Chroma STFT	67.16	68.03
Chroma CQT	64.13	63.75
Chroma CENS	65.67	64.53
Spectral Contrast	65.13	62.34
Tonnetz	64.23	62.78

characteristics of cough events. Our recent survey work [54] provides a comprehensive compilation of more than 300 features relevant to reliable AI-enabled cough-based diagnostics. For the development of RAISED, I specifically selected 102 features from this extensive list, encompassing 14 distinct classes based on time and frequency domain audio features. Table 3.2 presents an in-depth analysis of the individual performance of various acoustic features in the context of RAISED. The findings highlight the significant role of Mel-frequency cepstral coefficients (MFCC) in enabling an AI engine to effectively differentiate between diseases based on cough sounds. Notably, two novel features, namely

Table 3.3: Performance Analysis of Acoustic Features Combined for COVID-19 Screening (* ZC Rate: Zero Crossings Rate, Melspec: Melspectrograms)

Acoustic Features Used	**Performance Metrics**	
	Sensitivity	**Specificity**
MFCC + MFCC Delta	90.41	89.17
MFCC + MFCC Delta + MFCC Delta2	92.13	90.23
MFCC + MFCC Delta + MFCC Delta2 + Melspec	92.38	91.03
MFCC + MFCC Delta + MFCC Delta2 + Melspec + ZC Rate	91.76	90.34
MFCC + MFCC Delta + Melspec + ZC Rate + RMS Energy	92.57	90.27
MFCC + Melspec + RMS Energy + Spectral Rolloff	92.83	91.12
MFCC + Melspec + RMS Energy + Spectral Centroid	93.89	92.28
MFCC + Melspec + RMS Energy + Tonnetz	94.17	92.79
MFCC + MFCC Delta + MFCC Delta2 + RMS Energy	**95.15**	**93.48**

MFCC Delta and MFCC Delta2, which represent the first and second-order derivatives of MFCC respectively, demonstrate substantial importance despite their limited usage in existing COVID-19 screening literature. This underscores the potential of exploring these related MFCC-derived features. Additionally, widely recognized features such as Melspectrograms and RMS energy exhibit strong performance, aligning with their extensive utilization in previous studies. Conversely, certain features including Spectral rolloff, Chroma STFT, and Tonnentz, while commonly employed in the literature, demonstrate relatively lower contributions according to our analysis.

In Table 3.3, I present the performance evaluation of a collective set of acoustic features used for training the AI engine in RAISED. To construct this combined feature set, I merged the top-performing individual features identified in Table 3.2. The results indicate that the combination of MFCC, MFCC Delta, MFCC Delta 2, Melspectrograms, and RMS Energy yields the most favorable performance. However, it is worth noting that

Table 3.4: Performance Analysis of Statistical Features Used Individually for COVID-19 Screening

Statistical Features Used	Performance Metrics	
	Sensitivity	Specificity
Mean	95.15	93.48
Median	94.25	93.67
Max	86.12	84.35
Min	86.37	85.19
1st Quartile	89.24	90.37
3rd Quartile	90.27	91.78
Inter Quartile range	88.17	87.67
Standard Deviation	82.46	81.76
Skewness	77.35	74.29
Kurtosis	76.21	81.3

features such as Zero Crossing Rate, Spectral Rolloff, and Tonnetz exhibit a negative contribution when included in the feature set.

Table 3.4 presents an analysis of different statistical features used for acoustic features aggregation. Based on the analysis and results presented in Table 3.4 Mean and Median yield the best results for COVID-19 diagnosis, these results are aligned with our previous paper [2] on the same subject.

3.5 Conclusion

While recent research has shown promising potential in AI-enabled COVID-19 screening. It is important to acknowledge the numerous challenges that need to be addressed for the successful implementation of cough-based diagnosis systems. One significant set of challenges lies in the software and hardware aspects of cough-based diagnostics. The diverse range of audio recording devices and their associated hardware disparities pose difficulties in achieving accurate and robust classification. Furthermore, variations in software characteristics, such as sampling rates and file types, introduce complexities in data management and machine learning model performance.

To address these challenges I present a framework named RAISED Reliable AI-enabled Screening of COVID-19. RAISED quantify the impact of several challenges that, if addressed can pave the way towards cough-based disease diagnostics. Furthermore, it highlights the need for the development of noise-aware machine learning models that can handle the unique characteristics of cough data and the environmental factors that influence it. Additionally, there is a need for research into audio compression techniques specifically tailored for cough data, as well as effective pre-processing methods to mitigate the impact of environmental noise and reverberation.

While it may be challenging, the advancements made in the AI domain hold immense potential for enhancing the reliability, speed, accuracy, and accessibility of COVID-19 screening.

CHAPTER 4

Hotspot Prediction by Leveraging Multi-modal Noisy Data from Biosensing Devices

4.1 Introduction

The previous chapter presents RAISED, a robust framework for COVID-19 diagnosis from noisy acoustic data. In this chapter, I present a scalable version of RAISED that is a concept framework for proactive pandemic prediction using biosensing wearable devices.

4.1.1 Motivation and Background

While COVID-19 is not the first pandemic in the 21st century, it is one of the most devastating ones taking millions of lives and annihilating trillions of dollars from the global economy [59, 60, 61]. In the US alone, the social and economic damage of the COVID-19 pandemic has surpassed that of all the natural disasters in the last century combined [62]. The absence of proactive and scalable outbreak detection or pandemic prediction mechanisms is one of the core reasons why pandemics spread and thus cause greater socioeconomic damage than natural disasters like hurricanes and tsunamis. Predictive or early detection systems for calamities have helped humanity minimize human fatalities and economic losses in the past. The prevention of the catastrophic repercussions of potential pandemics requires establishing a similar early detection system. Such a system needs to be scalable to allow global surveillance and detection of infectious disease outbreaks, and thus predict pandemics at the pre-emergence or local outbreak stage.

4.1.2 Related Work

Previous epidemiological studies of the pathogenic diseases [63, 64, 65, 66] and extensive insights from different coping strategies for the COVID-19 pandemic [67, 68, 2], provide evidence that the early stage detection of the pandemic when it is just a local outbreak can be a game changer in containing the infection and preventing it from becoming a full-blown epidemic and then pandemic. The current laboratory-based diagnostic tests, that are often conducted after the infection has spread at the local level, do not offer the continual screening, agility, safety, scalability, and ubiquity to serve as a fast and proactive outbreak detection and thus a pandemic prediction and prevention system. Without such a system, COVID-19 cannot be expected to be the last pandemic of its scale and resultant catastrophic impact on the global health and economic system.

Based on the literature at our disposal, it is evident that the majority of infectious diseases present symptoms that can be detected and monitored through commodity wearable or ambient sensors. The biomarkers that can be measured to screen for these symptoms and thus detect an infection are also identified in this table. With advances in biosensing, nanotechnology, and wireless communications most of these biomarkers can be measured nonintrusive at population level and analyzed centrally. This observation combined with promising results and the impact of our seminal work [2] on screening for COVID-19, anytime anywhere just from the cough sounds by using an app installable on any commodity phone or watch, motivates us to propose iPREDICT (see fig. 4.1 for the schematic of the iPREDICT) an innovative framework that can enable in-situ and continuous screening at the population level to detect a new outbreak of an existing or a new disease at an early stage thus serving as potential pandemic prediction and prevention system that world direly needs.

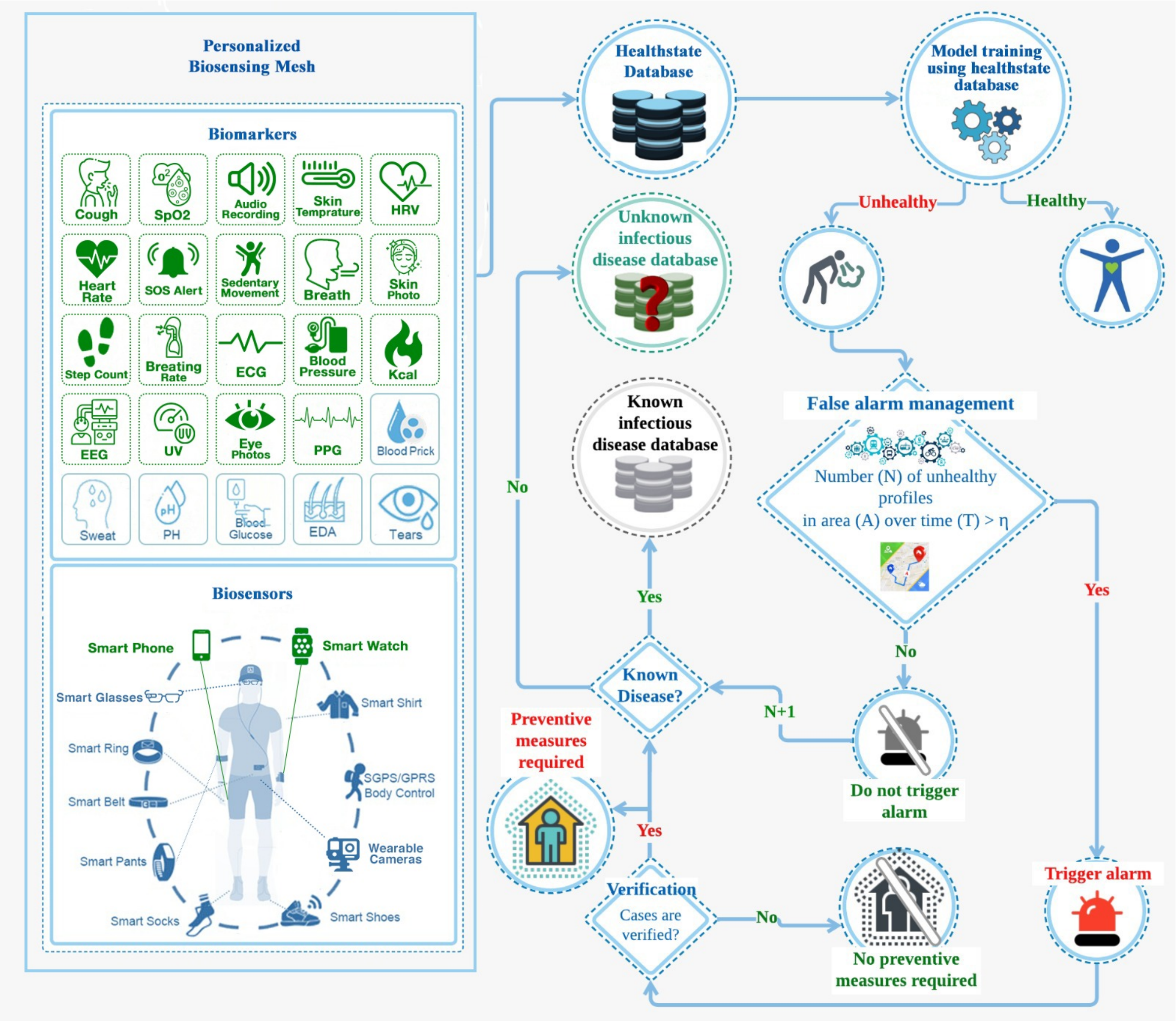


Fig. 4.1: iPREDICT: AI-enabled proactive pandemic prediction framework using wearable biosensing devices. *Biomarkers: SpO2(oxygen saturation), HRV (heart rate variability), ECG (electrocardiogram), Kcal (kilocalories burnt), EEG (electroencephalogram), UV (ultraviolet exposure), PPG (photoplethysmography), pH (saliva pH), EDA (electrodermal activity)

4.2 iPREDICT: AI-enabled Proactive Pandemic Prediction Framework

4.2.1 Overview of iPREDICT

iPREDICT presents a comprehensive framework given in fig. 4.1 for future pandemic prediction comprising four integral components. First, a personalized biosensing mesh forms the foundation, enabling real-time data collection. Second, a curated array of biomarkers, efficiently gathered through the biosensing mesh, facilitates intricate health assessments. Third, the synergy of AI models leverages individual biosensor data streams to facilitate personalized training, enhancing predictive accuracy. Fourth, by diligently analyzing these streams, the framework adeptly identifies burgeoning anomalies through the AI models' predictions, thereby enabling timely outbreak alarms, exemplifying iPREDICT's potential in proactive pandemic prediction. A detailed description of the individual components of iPREDICT is provided in the following subsections.

4.2.2 Components of iPREDICT

Key components and contributions of iPREDICT can be summarized as follows:

1. iPREDICT is a novel, AI-powered proactive pandemic prediction framework that uses wearable biosensors. The framework integrates expertise from various fields such as AI, epidemiology, and distributed system software development, to provide a comprehensive solution for accurate prediction of future pandemics. iPREDICT acquires essential biomarkers from free-life biosensors, analyzes the transmission pattern of infectious diseases based on location, and employs AI algorithms to raise alerts and prevent the rapid spread of the disease and potential pandemic outbreaks.

2. iPREDICT proposes a method that involves using graph neural networks (GNNs) to determine the pandemic prediction threshold, taking into account various environmental, geographical, and biological parameters. As the problem is complex, with no existing mathematical model that includes all these parameters, the proposed approach uses historical pandemic data to build a GNN-based framework capable of predicting epidemic thresholds at different scales and resolutions of the population. The expertise of the authors in applying AI in different domains informs the development of this method.

3. In the next chapter I present several crucial challenges that must be addressed in the widespread deployment of iPREDICT. With a strong focus on the engineering challenges within our research domain, which include AI, signal processing, and

cellular networks. The challenges I address include the collection of audio data (cough sounds) using various smartphone devices, at different audio sampling rates (for the efficient storage of audio data, which is critical for large-scale systems), and the transfer of audio data in different file sizes and formats, over cellular networks for analysis and diagnosis.

4. I demonstrate the feasibility of iPREDICT by leveraging our previous work AI4COVID-19 [2] as a case study and provide an analysis of the quantitative impact of four different engineering challenges on the performance of AI4COVID-19 by considering one biosensor (microphone) and one biomarker (cough sound) out of a massive list of available biosensors and biomarkers in a variety of biosensing devices, see fig. 4.1.

4.3 System Design of iPREDICT

4.3.1 Personalized Biosensing Mesh

Within the iPREDICT framework, I propose the ”personalized biosensing mesh” as a pivotal component, which capitalizes on the capabilities of diverse wearable devices such as smartwatches and smartphones. By ingeniously integrating these commodity wearables, a comprehensive biosensing ecosystem is forged, capable of capturing an array of vital biomarkers highlighted (in green color and dotted border) in fig. 4.1 such as skin temperature, SpO2, audio recording, heart rate variability, and cough sounds (proposed an even detailed list of biomarkers in fig. 4.1. Moreover, a description and how these biomarkers can be used as a symptom for the detection of various diseases is presented in Table 4.1). These biomarkers can be acquired using readily available wearable devices through their built-in biosensors [69, 70].

iPREDICT proposes the use of smartwatches and smartphones as wearable devices due to their usage convenience, acceptance, and availability to the masses. These wearable devices encompass an assortment of sensors, each equipped to measure specific biomarkers. For instance, heart rate sensors embedded in smartwatches meticulously track pulse rate variability and resting heart rates [71]. Accelerometers, commonly featured in smartphones can monitor movement patterns and quantify activity levels. Also, both the devices have microphones that can collect cough and audios that is used for respiratory disease diagnosis with impressive results [2]. Moreover, cutting-edge wearables incorporate photoplethysmography (PPG) sensors that ascertain blood oxygen saturation, while electrodermal activity sensors gauge stress levels. Temperature sensors integrated into devices like smartwatches serve as sentinels, detecting fluctuations indicative of fever or irregularities [72].

By harnessing this confluence of wearable devices and their inherent biosensors, a rich and diverse multimodal data stream is cultivated. The cough sound data is analyzed for the preliminary diagnosis of several respiratory diseases like Bronchitis, Pertusis, and COVID-19 [2]. Likewise, heart rate data, culminate in a comprehensive portrayal of an individual's activity levels and overall health. I highlighted a few of the biomarkers (cough, audio i.e. counting, heart rate) that showed promising results in COVID-19 screening [73, 2]. This cumulative biomarker dataset serves as the foundation for constructing a multidimensional individual health profile, emblematic of the iPREDICT framework's prowess.

4.3.2 Creation of Biomarker Profiles and Respective Challenges

The next component of iPREDICT is the creation of a comprehensive database of historical biomarkers of the population, I call it Healthstate database in the iPREDICT framework presented in fig. 4.1. Table 4.1 presents a list of biomarkers and the respective description of what latent information these biomarkers provide which can be exploited for the disease diagnosis. Healthstate database consists of biomarker measurements of individuals either labeled as 'healthy/normal' or as 'not normal' i.e., profiles of

Table 4.1: Description of Different Biomarkers for Disease Diagnosis

Biomarker	**Description**
Breath	Provides latent information via sound, smell, and intensity about a person's health.[74]
Audio Recording	Carries latent features that can be used in the acoustic analysis of respiratory diseases.[2]
Step Count	Provides insights about the lifestyle of a person that can relate to overall health. [75]
Burnt Kilocalories	Can be associated with cachexia which can be caused due to cancer or other chronic diseases.[76]
SOS Alert	Can be used in emergency cases when a person's vital signs fall outside a normal range.[77]
Heart Rate	Is a major biomarker for respiratory disease diagnosis.[74]
Sweat	Can be used for cystic fibrosis that causes damage to lungs and digestive system.[78]
Sedentary Movement	Can be used for cardiovascular disease diagnosis.[79]
Skin Temperature	Is a major symptom of the diseases that cause fever.[80]
Skin Photos	Provides information about allergic viruses.
Heart Rate Variability	Is a major biomarker for cardiovascular disease diagnosis.[74]
Oxygen Saturation (SpO_2)	Can be used as a biomarker for respiratory disease diagnosis such as COPD.[81]
Electrocardiogram	Is widely used biomarker for coronary heart disease diagnosis.[74]
Blood Pressure	Is a basic biomarker used by physicians for cardiovascular disease diagnosis.[74]
Saliva pH	Is used as a biomarker for stress examination and monitoring.[82]
Blood Glucose	Is a commonly used biomarker for the diagnosis of diabetes.[74]
Cough	Contains the signature of several respiratory diseases e.g. asthma, pertussis, bronchitis etc.[2]
Breathing Rate	Can be used as a biomarker for several diseases and conditions such as asthma, COPD, and pneumonia.[74]
Retinal Images	Is used as a biomarker for the diagnosis of diseases like chronic kidney problems and anemia.[74]
Electroencephalogram	Is a widely used biomarker for neurodegeneration problems like epilepsy, sleep disorders, and brain injuries.[74]
Photoplethysmograph	Is a biomarker used for cardiovascular disease detection.[83]
Ultraviolet Exposure	Can be used as a biomarker for systemic oxidative stress.[84]
Electrodermal Activity	Is used as a biomarker for the diagnosis of anxiety disorder and Parkinson's disease.[85]
Tears	Contains useful information in the fluid that can be used for the diagnosis of ocular and breast cancer.[86]

individuals that have been pre-identified to have some medical condition. These profiles will enable the detection of anomalous data points that lie within the health data stream of the individuals. However, biomarker profiling comes with several challenges brought by the variability and complexity of the biomarker data. These challenges can be broadly categorized into two categories explained below:

1. **Challenge 1: Inter-person Variability**: The creation of a personalized biosensing mesh comes with a complexity challenge. One way to address this is to create the

models on edge devices, and only send triggers along with select when anomaly is noted for central examination by AI and or medical and public health professionals. To cope with the low computational power of the edge devices, instead of advanced deep learning (DL) models, simpler template matching methods can be used.

2. **Challenge 2: Intra-person variability**: Non-infectious diseases, seasons, lifestyle changes, and stress can cause variations. Addressing this challenge requires not only the fusion of multiple biomarkers that reflect a multi-system state of the human body but also deep medical expertise. As an example, HRV is a biomarker that drops usually with the onset of most types of sickness. However, these sicknesses may not be the cause of concern for the iPREDICT system as they may not be infectious. Therefore, a reliable method to detect the spread of an infection is to have a higher-level model, that looks for patterns of anomalies among people who have been in close proximity. For example, if HRV of multiple people who have been in close contact starts dropping within a time window, then it can be considered as a case for further analysis of iPREDICT system. This further analysis is carried out in iPREDICT components described in the next sections.

4.3.3 AI-based Anomaly Detection Using Biomarker Profiles

In tandem with the challenges highlighted in the previous section, the high dimensionality (multiple biosensors capturing multiple biomarkers) and the large volume of data in an individual's biomarker profile add to the complexity of the anomaly detection component of iPREDICT. Due to such challenges, I propose a potential disease outbreak to be modeled as a time series anomaly detection problem. To achieve this, I propose a novel mechanism for identifying anomalous readings at an individual's biomarker level for detection of viral infections, at their onset. The biosensor time series data will be used to train an AI model for identifying individuals with biomarker levels deviating from their normal trend. Thus, the trained AI models will be patient-specific, mitigating the effects of intra as well as inter-personal variability and promoting precision medicine. Time series anomaly detection can be achieved using several machine learning (ML) models such as ARIMA, SARIMAX, etc. [87], and DL models (e.g., recurrent neural networks (RNNs), long short-term memory (LSTM) [88], and autoencoders [89]). The predictive results from these models will identify a potential disease outbreak and suspected individuals will be further tested to verify if a cluster of such anomalous data is present in close spatio-temporal proximity.

4.4 NAT: An Adaptive Thresholding for Disease Prevalence

iPREDICT identifies the trends of similar irregularities in the biomarker values of multiple individuals residing in close proximity over a brief time duration. Once the cluster of infected people is identified, iPREDICT triggers an alarm based on a disease-specific threshold to alert the authorities about a potential outbreak. I propose a pandemic threshold η based on several magnitude/number (N), area(A), and time(T) of infection parameters given in fig. 4.2. The threshold η is modeled based on **NAT** in eq.4.1.

$$\eta(N_d, A, T) > \eta_d \tag{4.1}$$

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

Finding the quantitative value of η_d is a challenging task for the epidemiology research community. An even bigger challenge lies in adaptively setting this threshold to minimize the intervention time for the authorities to take necessary measures. While we know from the epidemiology literature [90, 91, 92] that **NAT** depends on a wide range of factors as listed in fig. 4.2, we do not have a quantitative representation of **NAT** that includes all the factors of fig. 4.2, and developing a quantitative understanding will take decades of research by the epidemiology research community or we may never know its closed-form mathematical equation. The challenge is due to the diversity of the nature of infectious diseases (reproduction rate, nature of spread (airborne or touch), association with other infectious diseases, etc.) and general human behaviors (mobility pattern, response to intervention policies, etc.). Therefore, leveraging the capability of AI to learn such complex relationships between a large variety of parameters and the availability of data on the recent pandemics, I propose a GNN-based framework for alarm management as the next component of iPREDICT.

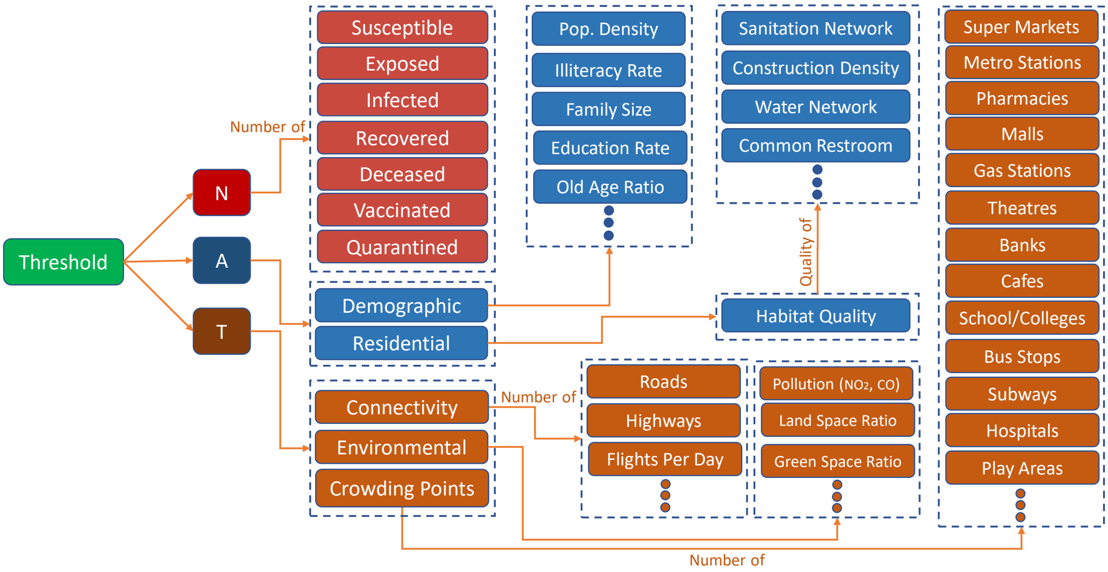


Fig. 4.2: Qualitative representation of pandemic threshold based on NAT using associated parameters from epidemiology.

4.4.1 Population Resolution and Scale-Agnostic Graph Neural Network System for Alarm Management

To overcome the challenges highlighted and discussed in the previous section, I propose an AI-enabled data-driven approach to learn pathogen and population dynamics-specific values for η_d by learning from historical data of epidemics. I aim to take benefit from the recent advances in DL on graphs, i.e., Graph Neural Networks [93] which learn to predict the η_d by performing convolutions on a graphical representation of the population and its features listed in fig. 4.2.

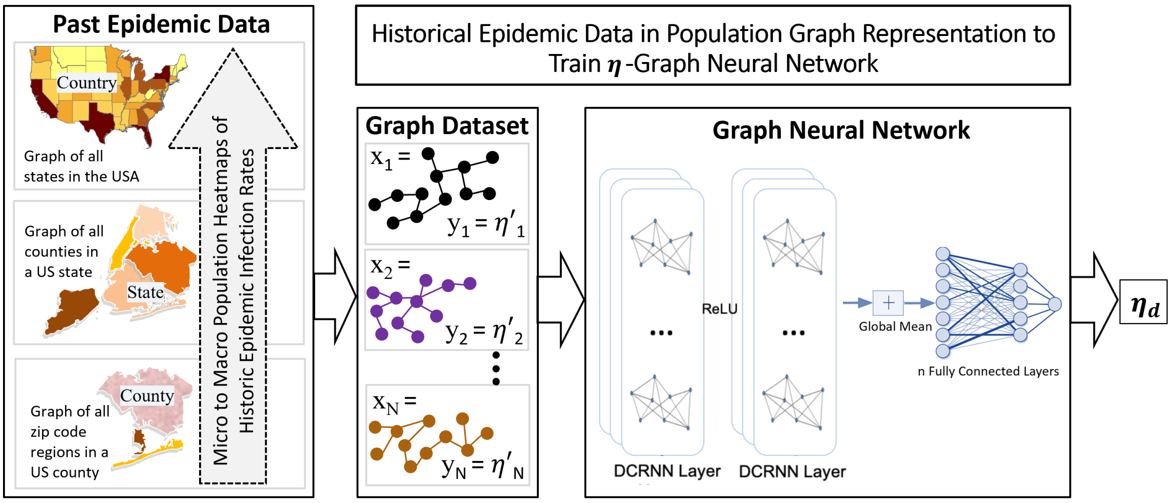

Past Epidemic Data
Country
Graph of all states in the USA
Graph of all counties in a US state
State
Graph of all zip code regions in a US county
County
Micro to Macro Population Heatmaps of Historic Epidemic Infection Rates
Historical Epidemic Data in Population Graph Representation to Train η-Graph Neural Network
Graph Dataset
$x_1 =$
$y_1 = \eta'_1$
$x_2 =$
$y_2 = \eta'_2$
$x_N =$
$y_N = \eta'_N$
Graph Neural Network
ReLU
Global Mean
n Fully Connected Layers
DCRNN Layer
DCRNN Layer
η_d

Fig. 4.3: Graphical modeling of historical epidemic data and training of Graph Neural Network to predict η_d

Extensive literature in epidemiology exists where historical epidemic data is used for forecasting the future state of an epidemic. Firstly, compartmental models such as SIR [94], SIERD [95], and SIRV [96], etc., comprise systems of ordinary differential equations which predict epidemic parameters and spread. Secondly, ML models like SARIMA [97] predict future infection rates through time series forecasting on past infection rates. Thirdly, time series forecasting is also performed via Deep Neural Networks (DNNs) such as LSTMs [98, 99, 100]. However, these approaches rely solely on the temporal aspect of the epidemic, i.e., historical infection rates, while not accounting for the spatial dynamics of the population such as density, distribution, inter-mobility, and population characteristics like hygiene, humidity, etc., which can be vital in driving an epidemic. This is evident from [101], where integrating rainfall data with infection rates significantly improved the forecast of Dengue, since humidity and stagnant water caused by rain breed Dengue carrier mosquitoes. Similarly, [102] shows that meteorological factors like atmospheric pressure positively influenced the forecast of Influenza B, and [103] examined the influence of mobility data on Influenza spread modeling. Despite such studies advocating for the efficacy of spatial features in epidemic prediction, an all-encompassing predictive model with spatial as well as temporal features is yet to be established. Recently, several studies [104, 105, 106, 107, 108, 109] emerged where a population is modeled as

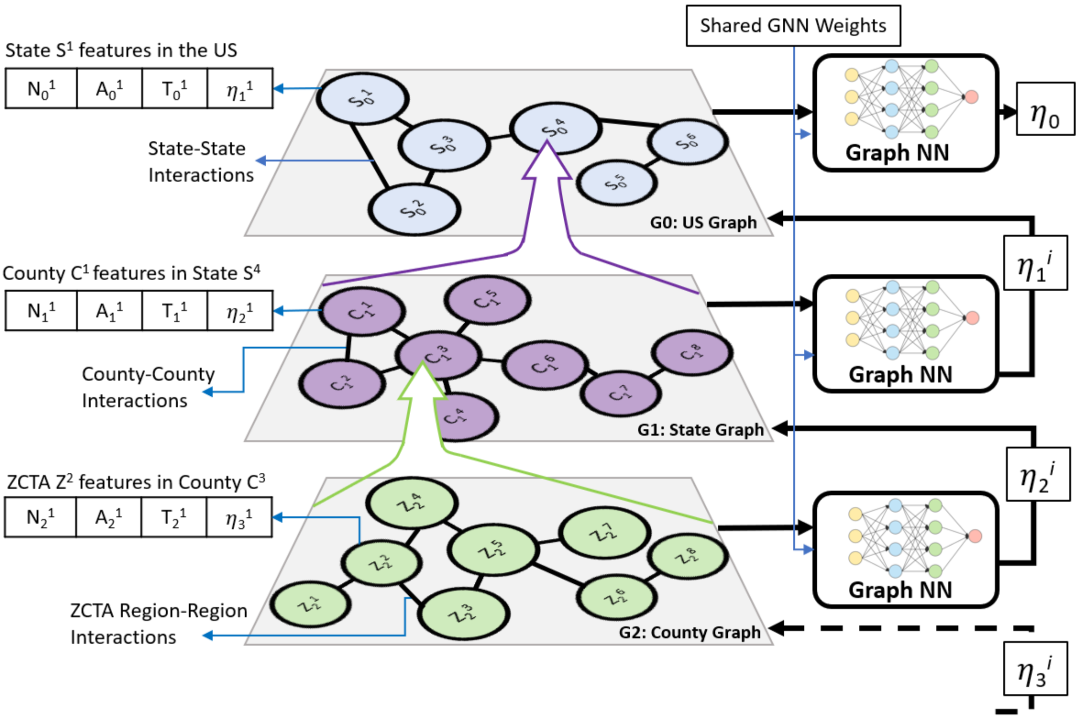


Fig. 4.4: Multi-hierarchy and Multi-scale modeling of Populations into graphs where η_d is predicted from lower hierarchy graphs inform the prediction of η_d at higher hierarchy.

a knowledge graph such that it captures the temporal characteristics of the population as graph node features and spatial dynamics as well as mobility as graph structure i.e., adjacency matrix. Such graphical modeling aligns with the widespread use of graphs in epidemiology where spot maps, heat maps, and area (Patch or Choropleth) graphs are employed to illustrate the geographical spread of outbreaks on a 2D plane [110]. However, the representation of such maps as knowledge graphs which can train DL models, and the DL on graphs with Graph Neural Networks are emergent research directions in AI which have demonstrated improved prognostic capability over classical ML and DL for epidemic forecasting [104, 105, 106, 107, 108, 109]. Therefore, based on 1) the limitation of compartmental, ML, and DNN models to capture population features and mobility, 2) the conventional capability of graphs in epidemiology to encapsulate these factors effectively, and 3) the recent advancement in DL on graphs to build predictive models from spatio-temporal graph datasets, I suggest a GNN model that learns from the dynamic graphical representation of populations during past epidemics to predict η_d future epidemics.

In [109, 108, 106, 104] Graph neural networks embed graphs of US counties as low-dimensional latent embeddings which capture the population characteristics. As graph features (e.g, infection rates) vary against time, the embeddings of population graphs from past time $\{t-d, t-d+1, t-d+2, ..., t\}$ are treated as a time series with either Transformer or LSTM for forecasting of future $t+1$ infection rates. However, these approaches do not take into account the resolution and scale of the population graphs as variables. As the resolution increases from country to state to county and further, the properties of the population graphs change and hence a GNN model trained with data exclusively at low resolution (such as at the state level) cannot be employed at high resolution such as zip code and vice versa. In [109] and [108], the authors acknowledge resolution as a significant variable, but the multi-resolution nature of their model comes from the clustering of graph nodes and making condensed graphs from the pooled features of the clustered nodes. However, the clustering of graph nodes is data-driven therefore it disregards the natural clustering of regions due to standard geographic divisions e.g., all counties in one state can exhibit similar characteristics due to proximity, inter-mobility, cultural and environmental similarities, so they should be clustered together. The clustering of regions based on geography takes advantage of Tobler's first law of geography [111], which states that spatially closer regions have higher similarity than spatially distant regions. Such geography-aware clustering, therefore, eliminates the need for training data and hyper-parameter search required in data-driven clustering. In addition to resolution, the scale of the graphs is a bottleneck. For instance, [108] makes a graph with all the counties in the US as nodes. Given that there are 3,142 counties and equivalent regions in the US, one graph will contain as many nodes and up to 4.9 million edges. Further increasing the resolution will result in a graph of 40,000 nodes = no. of five-digit zip codes in the US [112]. On the other hand, the lower the resolution, the smaller the graph but the crucial early-stage infection data is lost. For instance, if all the US states are modeled as a graph of 50 nodes (high scale, low resolution), then the pathogen breakout can be detected when it is already epidemic across states while the goal should be early detection during spreads over one county. Therefore, the amount of area covered in a graph, i.e., the scale of the graph should be inversely proportional to the resolution of the graph. Building on this understanding, I propose a multi-resolution multi-scale hierarchical approach for modeling populations as graphs and training an end-end resolution and scale-agnostic GNN which learns to predict pandemic alarm threshold η_d from population features listed in fig. 4.2.

We divide the population in a nested manner based on the Standard Hierarchy of Census Geographic Entities [113], from micro to macro-region i.e., ZIP Code Tabulation Areas (ZCTA), county, and state. The set of all ZCTAs in one county forms a county-level graph as shown in fig. 4.4. Similarly, all counties in a state form one state-level graph, and so on. Therefore, the total number of spatial graphs is, $S_{total} * C_{avg} * Z_{avg}$ where $S_{total} = 50$ for the states in the US, C_{avg} is the average number of counties in US states, and Z_{avg} is the average number of ZCTAs in US counties. For each geographic level, there can be multiple temporal graphs that have the same spatial structure but vary in node/edge features as each temporal graph consists of historical data from time t to $t+T$ where T is the time span of a week.

To summarize, the historical epidemic data is to be modeled into a set of graphs $\{G_h^t | h \in H, t \in T\}$ where H is the spatial granularity/hierarchy such as {ZCTA, County, State} and T is time granularity of data such as $\{week_0$, $week_1$,...,$week_T\}$. Each graph G in G_h^t is represented as $G = (V, E)$ where V is the set of nodes representing regions at hierarchy h and E is the set of edges between nodes. Each node has a set of features $X = \{x_0, x_1, ..., x_k\}$ which represent **NAT** features listed in fig. 4.2. Hence, in node features, I combine a plethora of data sources which together affect the risk of pathogen breakouts. Historical data of past epidemics and pandemics consisting of the number of infected, susceptible, recovered individuals, etc., over time in a region form the dynamic node features. Properties of the population, such as census data, population density, death/birth rate, poverty, literacy rate, age and gender demographics, etc. along with environmental factors that highlight the population's limitations and resources such as weather, pollution, and terrain form the static features. One node thus consists of all the relevant features to qualify the breakout within that node i.e., a population section (such as Zip code no. 11005 or the Queens county, depending on whether the population resolution is zipcode-level or county-level). To join nodes with edges, I determine the geographical connectivity between regions by using spatial distance as well as road network density and borders between the regions which are modeled as nodes. For each edge, the weight e is a function of human mobility pattern from Facebook Data for Good [114] which uses the location history from mobile devices to track air, road, or train travel between two regions and also specifies normal mobility ranges of communities, cohabitation and co-movement of groups. So, in summary, the graphical modeling of historical epidemic and pandemic data is such that the node features represent all the variables that can quantify pathogen spread within a region while edges and edge weights represent the variables that account for the spread of a pathogen from one region to another.

The most significant aspect of this geographically nested graph dataset is that η_d from the lower hierarchy serves as a node feature in the higher hierarchy graph as depicted in fig. 4.4. In the training dataset, the η_d at each hierarchy level comes from the averaging of η_d of nodes at the lower level, e.g,

$$\eta_{NYState} = \frac{\sum_{i=1}^{C_{NYState}} \eta_i}{|C_{NYState}|} \tag{4.2}$$

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

This dataset of spatio-temporal graphs described above is ingested by a Graph Neural Network (GNN) as shown in fig. 4.3, specifically Diffusion Convolution Recurrent Neural Network (DCRNN). DCRNN, introduced by Y. Li et al. [115] is a type of GNN designed for addressing spatio-temporal forecasting tasks. It combines diffusion convolutional layers and recurrent layers to capture spatial and temporal dependencies, respectively, and integrates them into a unified framework. The diffusion step captures spatial dependencies by propagating information through the graph structure of the data. It allows each node to aggregate information from its neighboring nodes. The recurrent step captures temporal dependencies by incorporating historical information from previous time steps using a recurrent architecture, such as a Gated Recurrent Unit (GRU). The matrix multiplication in GRU is replaced with the diffusion convolution described above, thus integrating the diffusion, convolutional, and recurrent steps in DCRNN, i.e., effectively modeling both the spatial dependencies among different locations in the graph and the temporal dependencies over time. DCRNN transforms the input node features into lower dimensional embedding in latent space. The embeddings are optimized at every training step to best capture the information from node features and node neighbors. The latent embeddings from all the nodes are then combined by either concatenation, mean pooling or trainable pooling layers such as hierarchical pooling [116] and Self-attention graph pooling [117]. Then, the pooled embedding is passed through fully connected neural network layers to finally output η_d, a real number that embodies the threshold for **NAT** parameters such that when $\eta(N_d, A, T) > \eta_d$, the alarm is triggered in the system as shown by fig. 4.1. As GNNs are independent of the graph structure, a GNN such as DCRNN trained on graphs of multiple resolutions and scales, can learn features that are resolution and scale-agnostic. Hence, the resultant trained GNN can be deployed at ZCTA, county, or state level with the shared weights as shown in Fig 4.4.

4.4.2 Verification and Preventive Measures for iPREDICT

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

CHAPTER 5

Addressing the Engineering Challenges in Population-Level Biosensed Data

5.1 Introduction

In the previous chapter, I present the concept and system architecture of iPREDICT. iPREDICT leverages population-level biomarker data that comes with several challenges including engineering challenges. This chapter focuses on mitigation techniques for these challenges and numerical results to verify their viability.

5.2 Engineering Challenges in Implementation of iPREDICT

A variety of challenges in the medical and engineering domains will be faced when implementing iPREDICT. The challenges in engineering arise from devices used to record a variety of biomarkers using biosensing technology, see fig. 4.1. The recording devices involved vary in terms of their biomarker-capturing mechanisms, hardware components, and software capabilities (operating system, middleware, etc.) that introduce variability and randomness in the data collection process. Moreover, environmental factors such as ambient noise can also impact the performance of iPREDICT. Therefore, in this study, I identified and quantitatively assessed several key engineering challenges encountered in diagnosing COVID-19 based on cough sound biomarkers, recorded using a smartphone microphone. The following engineering challenges included questions related to audio signal processing, such as the effects of contamination of cough sound with the environmental noise, variability of self-induced noise (by active circuitry, and Brownian movement of air particles) by a microphone [118], variation in sampling frequencies (in order to capture

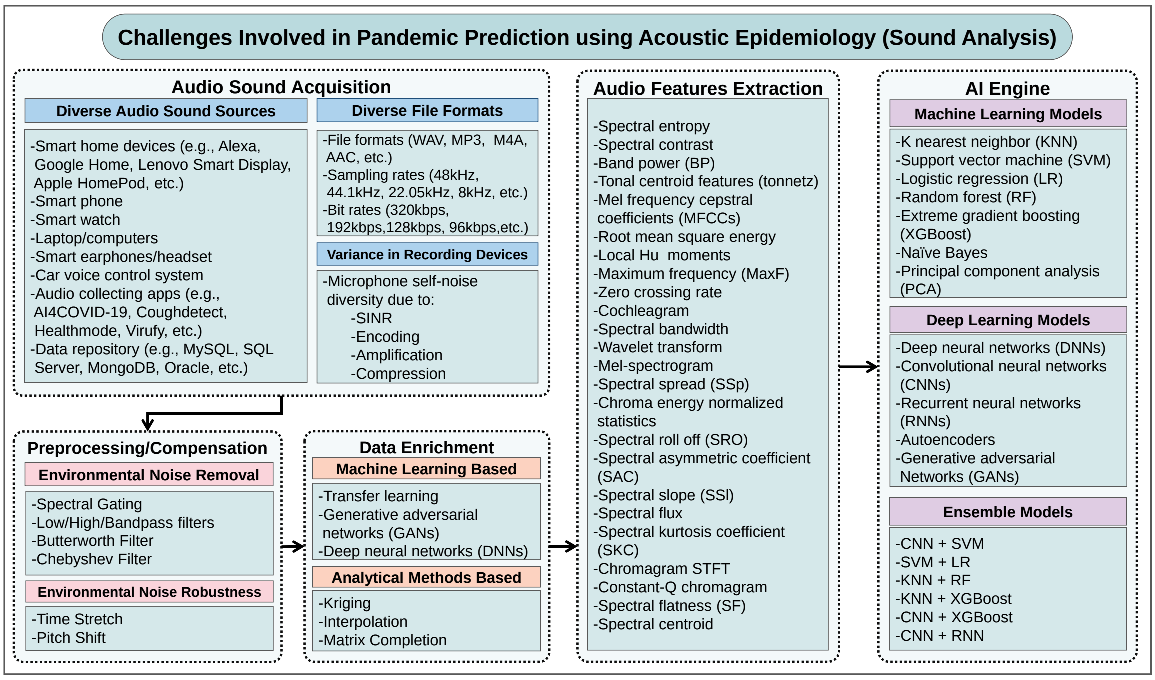

Challenges Involved in Pandemic Prediction using Acoustic Epidemiology (Sound Analysis)
Audio Sound Acquisition
Diverse Audio Sound Sources
-Smart home devices (e.g., Alexa, Google Home, Lenovo Smart Display, Apple HomePod, etc.)
-Smart phone
-Smart watch
-Laptop/computers
-Smart earphones/headset
-Car voice control system
-Audio collecting apps (e.g., AI4COVID-19, Coughdetect, Healthmode, Virufy, etc.)
-Data repository (e.g., MySQL, SQL Server, MongoDB, Oracle, etc.)
Diverse File Formats
-File formats (WAV, MP3, M4A, AAC, etc.)
-Sampling rates (48kHz, 44.1kHz, 22.05kHz, 8kHz, etc.)
-Bit rates (320kbps, 192kbps,128kbps, 96kbps,etc.)
Variance in Recording Devices
-Microphone self-noise diversity due to:
-SINR
-Encoding
-Amplification
-Compression
Preprocessing/Compensation
Environmental Noise Removal
-Spectral Gating
-Low/High/Bandpass filters
-Butterworth Filter
-Chebyshev Filter
Environmental Noise Robustness
-Time Stretch
-Pitch Shift
Data Enrichment
Machine Learning Based
-Transfer learning
-Generative adversarial networks (GANs)
-Deep neural networks (DNNs)
Analytical Methods Based
-Kriging
-Interpolation
-Matrix Completion
Audio Features Extraction
-Spectral entropy
-Spectral contrast
-Band power (BP)
-Tonal centroid features (tonnetz)
-Mel frequency cepstral coefficients (MFCCs)
-Root mean square energy
-Local Hu moments
-Maximum frequency (MaxF)
-Zero crossing rate
-Cochleagram
-Spectral bandwidth
-Wavelet transform
-Mel-spectrogram
-Spectral spread (SSp)
-Chroma energy normalized statistics
-Spectral roll off (SRO)
-Spectral asymmetric coefficient (SAC)
-Spectral slope (SSl)
-Spectral flux
-Spectral kurtosis coefficient (SKC)
-Chromagram STFT
-Constant-Q chromagram
-Spectral flatness (SF)
-Spectral centroid
AI Engine
Machine Learning Models
-K nearest neighbor (KNN)
-Support vector machine (SVM)
-Logistic regression (LR)
-Random forest (RF)
-Extreme gradient boosting (XGBoost)
-Naïve Bayes
-Principal component analysis (PCA)
Deep Learning Models
-Deep neural networks (DNNs)
-Convolutional neural networks (CNNs)
-Recurrent neural networks (RNNs)
-Autoencoders
-Generative adversarial Networks (GANs)
Ensemble Models
-CNN + SVM
-SVM + LR
-KNN + RF
-KNN + XGBoost
-CNN + XGBoost
-CNN + RNN

Fig. 5.1: Engineering challenges associated with the implementation of the pandemic prediction framework.

high-quality cough samples needed for AI-based pandemic prediction), and compression rates for efficient storage and transmission of cough samples.

5.2.1 Environmental Noise and Noise Variations

Recording cough sounds via smartphones in a public setting induces environmental noise and reverberation which contaminate the recording and consequently compromise the accuracy of the ML models for disease diagnosis. While reverberation can be characterized by sound propagation models in indoor and outdoor settings, [55] environmental noise can vary greatly based on the surroundings. The noise amplitude and frequency distribution along with the signal-to-noise ratio in each sound recording can be unique. Moreover, even in the same ambient setup, the microphone-to-mouth positioning in terms of distance and angle causes noise variations in recordings. Although noise can be reduced by leveraging filtering and smoothing algorithms, this results in the elimination of high-frequency components in the recording. However, these high-frequency components are not always noise-induced and can be crucial for accurate cough-based diagnosis and hence, can not be completely removed. Moreover, ML models trained on noiseless data or data with limited noise scenarios tend to overfit and cannot be generalized to real-life settings where infinite types of environmental noises exist. Therefore, the challenge is to build noise-aware ML models that are robust to environmental distortions at training and inference [56].

5.2.2 Heterogeneity of Microphones

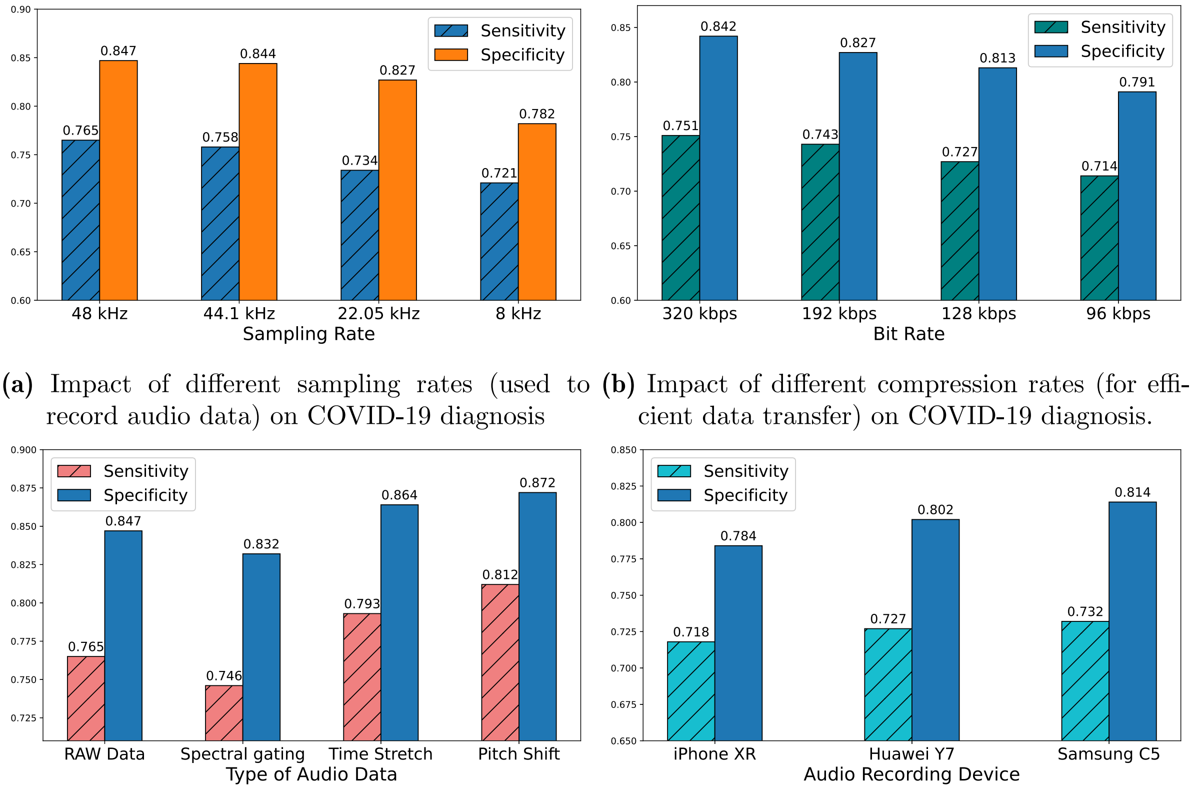
Another factor that complicates ambulatory sound-based cough diagnostics is the heterogeneity of recording devices at three levels: 1) varying device types (e.g., cellphones, laptops, lapel microphones, and smartwatches); 2) devices of the same type from different manufacturers (Apple, Google, and Samsung, etc.) with varying specifications (frequency response, phase response, sensitivity, noise level, sound pressure level, signal to noise ratio, and; 3) devices with the same specifications (same brand and same model) that exhibit electro-acoustic variations due to inherent manufacturing process uncertainties of microphone chips [118], [58]. Our focus is on the differences in recording microphones from the same or different manufacturers. Sound recorded via different microphones is not identical and hence affects the performance of iPREDICT. Therefore, the challenge is to generalize the ML models beyond the electro-acoustic differences in microphones.

5.2.3 Diversity in Audio Sampling Rate

In conjunction with the hardware dissimilarities of the microphones presented in the previous section, software characteristics such as the sampling rate (at which the audio is recorded) generate sound recordings with variable size and quality. Thus, poses the caveat of the trade-off between ML model accuracy and speed (audio with a higher sampling rate will take more time to process, which will compromise on efficiency of the ML model but yield more accurate results). I highlight the audio sampling rate diversity challenge by presenting the quantified impact of 4 different sampling rates in Section. 5.3. The results present a comparative analysis of different sampling rates on the diagnosis of COVID-19 using cough audio data. The cough sounds can have frequency components up to 20kHz [119]. Therefore I highlight the impact of 8kHz, 22.05kHz, 44.1kHz, and 48kHz sampling rates in the case study observing the Shannon-Nyquist sampling theorem (i.e., the sampling frequency must be more than double the highest frequency component) [120].

5.2.4 Diversity in Audio File Format

In addition to the sampling rate, the data is lost from the sound recordings through compression as they are stored on the recording devices using different file formats. Although lossless audio formats such as WAV, AIFF, ALAC, and FLAC exist, their larger storage size renders them inefficient for mobile transmission over the network, storage in recording devices and cloud, and consumption by the ML models at scale. Therefore, compression formats such as 3GP, WMA, AAC, M4A, and MP3, with MP3 being the most common [121] reduce the size of the audio file. Furthermore, the reduced file size is efficient for storing as well as transmitting over the network, to process the audio file on the cloud for biomarker profile signature creation and matching for the potential pandemic prediction. However, file compression does not only lose data but also, does so in favor of keeping human audibility while discarding human indiscernible components. Moreover, to discriminate the nuances of cough originating through closely related respiratory disorders, the frequency components beyond human audible interest can be significant. In addition to the data loss, compression formats also engender the challenge of portability over many codecs used by different recording software. A generalized cough-based diagnosis system must therefore be able to process different codecs or re-encode variable formats into one while being robust to the compression rates and mechanisms of varying formats. This challenge is further highlighted and quantified in Section 5.3.



(a) Impact of different sampling rates (used to record audio data) on COVID-19 diagnosis

(b) Impact of different compression rates (for efficient data transfer) on COVID-19 diagnosis.

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

(d) Impact of device diversity (model trained on data acquired by one device and tested on data acquired by another device) on COVID-19 diagnosis.

Fig. 5.2: Performance analysis of AI4COVID-19 based on various software and hardware related data diversities, different (a) sampling rates (b) bit rates (compression rates) (c) pre-processing methods (d) device diversity, used to acquire training data.

5.3 Case Study: Showing Feasibility of Smartphone-Based Biosensing

We leverage our seminal work AI4COVID-19 [2] as a case study to show the feasibility of iPREDICT. By exploiting one biomarker (cough sound) and one biosensing device (smartphone microphone) I analyze and quantify the engineering challenges discussed in Sec. 5.2. These challenges come from: 1) audio data acquisition, 2) data transfer over the wireless network, 3) data pre-processing for noise removal and noise robustness, and, 4) the diversity of data acquisition devices. Interested readers can refer to [2] for the details regarding the multi-pronged data-driven AI model, cough sound features, and dataset used in AI4COVID-19.

The first challenge is in the audio data acquisition phase due to the variable sampling rates discussed in Sec. 5.2 and highlighted in fig. 5.1. I present an analysis of AI-based COVID-19 diagnosis using cough data acquired at 4 different sampling rates 48kHz, 44.1kHz, 20.05kHz, and 8kHz. Fig. 5.2a presents a comparison of COVID-19 diagnosis performance using a variation in true positive and true negative rates (sensitivity and specificity). The results in fig. 5.2a show that the sensitivity of COVID-19 diagnosis decreases from 0.765 to 0.721 when the sampling rate of cough sounds used for investigation is decreased from 48kHz to 8kHz. Moreover, the specificity is also reduced from 0.847 to 0.844, 0.827, and 0.782 for the respective sampling rates. The drop in performance is a function of 3 factors that are involved in the process of up and downsampling of cough data. These factors include interpolation, anti-aliasing, and decimation [122]. This challenge can be further investigated as future work, as an optimization problem between diagnosis performance (diagnosis accuracy of the ML model) and efficiency (time taken by ML model for the inference) of the proposed framework.

Once the audio data is acquired by a biosensor, it needs to be transmitted over a wireless network that has different transmission capabilities. This requires audio data to be compressed, which has several compression formats and bit rates highlighted in Sec. 5.2 and fig. 5.1. To further highlight this challenge I present MP3 file format compressed at 320kbps, 192kbps, 128kbps, and 96kbps bit rates, to see the effect of cough data compression on the COVID-19 diagnosis. Fig. 5.2b shows a performance deterioration in sensitivity from 0.751 to 0.714 when the compression bitrate of MP3 is changed from 320kbps to 96kbps. Also, the specificity is reduced from 0.842 to 0.79 for the respective bitrates. The drop in performance is a function of quantization error, as well as data encoding[123]. Both, the quantization error and data encoding compromise the quality of audio, which leads to losing some latent features such as MFCCs, band power, and energy, (a comprehensive list is given in fig. 5.1) that are important for COVID-19 diagnosis. The detrimental impact of file compression can be further investigated using more sophisticated ensemble methods (e.g. CNN+XGBoost, CNN+LSTM, and CNN+SVM) that are robust to noise caused by data compression[57].

The third challenge arises before transferring the audio data to the ML model, due to the use of pre-processing methods to mitigate the impact of environmental noise in sound recording. There can be two approaches to reduce the effect of environmental noise on the efficiency of the ML method: 1) remove the environmental noise, 2) add more noise to the training data to make the ML model robust to learning from the noisy data. I present the results of spectral gating (SG) as a noise removal method and time stretch (TS) and pitch shift (PS) as noise robustness methods for this case study in fig. 5.2c, while more methods can be found in fig. 5.1. Fig. 5.2c shows that TS and PS improve the sensitivity from 0.765 to 0.793 and 0.812 respectively, and also achieve an enhanced specificity for both TS and PS. In contrast, the noise reduction technique SG brings the sensitivity down to 0.746 from 0.765, the specificity is also decreased slightly. A major reason for the drop in performance can be the nature of cough data which is more like noise itself, and when a static noise removal technique like SG is applied it removes some of the latent frequency features that contribute towards the COVID-19 diagnosis, which leads to slightly poor performance. The slight change in the performance of the ML models can be attributed to the lack of variation in the environmental noise due to the controlled nature of the environment setting (hospital setting) used for the data collection for this feasibility of the case. With more diverse data gathering settings, it is expected to have more noise variations, and hence it remains crucial to further investigate the impact of noise on the performance of the ML models.

The fourth challenge results from the variance in audio data recording devices. The devices can have different hardware (microphone, speakers, etc.) and software (operating system, middle-ware, etc.) based on brand, make, and model. In this case study, I focus on diversity based on the microphone because that is used to record the audio data. I present an analysis of a diverse set of devices consisting of an iPhone XR, Huawei Y7, and Samsung C5. I trained the AI4COVID-19 framework on data acquired using an Android device and tested on data from iPhone XR, Huawei Y7, and Samsung C5. The results in fig. 5.2d, show a decrease in sensitivity from 0.765 to 0.718 and specificity from

0.847 to 0.784, for the cough data that has an added noise signature of iPhone XR. In contrast, for the Android devices (Huawei Y7 and Samsung C5) the dip in performance is relatively lesser. The potential reasons can be 1) the software of Android devices differs from an iPhone device, and 2) the microphone chips differ for different mobile phone brands. These diversities contribute to the self-noise profiles of each device which leads to variation in the diagnosis performance.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

We are living in a transformative era where the convergence of artificial intelligence (AI), wireless networks, and biosensing technologies is reshaping the landscape in various domains starting from fault diagnosis in cellular networks to disease screening in humans, with far reaching impact all the way to management of future pandemics. In this thesis, I have embarked on a journey to harness the potential of AI-enabled diagnostics, particularly amidst the challenges posed by noisy and incomplete data in image and audio modalities.

The introduction of HYDRA in this thesis—a novel framework for root cause analysis of coverage-related anomalies—demonstrates a paradigm shift in fault diagnosis methodologies. By enriching sparse minimization of drive tests (MDT) data through image inpainting methods and leveraging a hybrid deep learning model, HYDRA exhibits unparalleled robustness in fault diagnosis, even amidst noisy data environments. Through rigorous evaluation against varying levels of data sparsity and comparison with state-of-the-art approaches, HYDRA emerges as a beacon of reliability and efficiency in diagnosing faults within complex cellular networks.

Moreover, our exploration extends beyond the realm of telecom network health assessment to address critical challenges in scalable anytime, anywhere respiratory disease screening using ordinary smartphones through cough sounds and other biomarkers. Our investigation has unveiled promising avenues toward the realization of AI-enabled diagnostics for scalable respiratory disease screening and pandemic management. By leveraging crowdsourced biomarkers from biosensing wearable devices and employing real-time anomaly detection techniques within a spatio-temporal framework, our proposed concept framework, iPREDICT, offers a beacon of hope in the battle against emerging epidemics. Through the integration of graph neural networks (GNNs) for threshold prediction and the lessons learned from our previous endeavor, AI4COVID-19, I present a promising strategy to mitigate the devastating impact of future pandemics. The practical feasibility and engineering challenges associated with pandemic prediction based on sound analysis further underscore the urgency and significance of our proposed framework.

In essence, this thesis presents the transformative potential of AI and emerging technologies in revolutionizing diagnostic procedures across diverse domains. By addressing the inherent challenges of noisy and incomplete data through innovative methodologies and practical implementations, it aspires to usher in a future where accurate and timely diagnostics pave the way for improved decision-making and enhanced societal well-being. Through iPREDICT and HYDRA, it paves the path toward a future where the specter of pandemics and network anomalies is met with proactive intelligence and resilient solutions.

6.2 Future Work

In the realm of AI-enabled diagnostics and fault diagnosis, future research must navigate critical frontiers to realize the full potential of these technologies. One pressing concern is the intricate challenge of data privacy in population-level frameworks, demanding novel methodologies to ensure ethical data handling while preserving individual privacy rights. Simultaneously, the quest for enhanced data quality and standardization looms large, necessitating the refinement of assessment techniques and the establishment of robust protocols to counter the adverse effects of low-quality data, especially evident in our exploration of audio data degradation. Fostering engagement and participation across diverse stakeholders is equally vital, prompting the development of strategies to promote collaboration and data sharing among healthcare professionals, researchers, policymakers, and the public.

On the technical front, research endeavors must venture into innovative fault delineation techniques, particularly to disentangle confounding faults like low transmitter power and cell outages. Leveraging Bayesian or other conditional classifiers informed by historical fault data holds promise for more accurate fault diagnosis. Expanding the diagnostic scope to encompass additional coverage anomalies such as cell individual offset (CIO) and exploring alternative Key Performance Indicators (KPIs) like Reference Signal Received Power (RSRP) offer avenues for improved fault detection. Moreover, tailoring machine learning (ML) algorithms and data enrichment methods to problem-specific contexts is paramount. Future research should focus on optimizing ML models and enrichment techniques for the unique challenges posed by diagnostic tasks, ensuring their efficacy in domains ranging from pandemic prediction to cellular network fault diagnosis.

Bibliography

[1] F. A. Afsar, M. Riaz, and M. Arif, “A comparison of baseline removal algorithms for electrocardiogram (ecg) based automated diagnosis of coronory heart disease,” in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering.* IEEE, 2009, pp. 1–4.

[2] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, “AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app,” *Informatics in medicine unlocked*, vol. 20, p. 100378, 2020.

[3] S. Ghrabli, M. Elgendi, and C. Menon, “Challenges and opportunities of deep learning for cough-based COVID-19 diagnosis: A scoping review,” *Diagnostics*, vol. 12, no. 9, p. 2142, 2022.

[4] A. Imran, A. Zoha, and A. Abu-Dayya, “Challenges in 5g: how to empower SON with big data for enabling 5G,” *IEEE network*, vol. 28, no. 6, pp. 27–33, 2014.

[5] A. Asghar, H. Farooq, and A. Imran, “Self-healing in emerging cellular networks: review, challenges, and research directions,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1682–1709, 2018.

[6] M. Amirijoo, L. Jorguseski, T. Kurner, R. Litjens, M. Neuland, L. Schmelz, and U. Turke, “Cell outage management in LTE networks,” in *2009 6th International Symposium on Wireless Communication Systems.* IEEE, 2009, pp. 600–604.

[7] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, “Planning wireless cellular networks of future: Outlook, challenges and opportunities,” *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[8] R. K. Sahoo, M. S. Squillante, A. Sivasubramaniam, and Y. Zhang, “Failure data analysis of a large-scale heterogeneous server environment,” in *International Conference on Dependable Systems and Networks, 2004.* IEEE, 2004, pp. 772–781.

[9] F. Xing and W. Wang, “On the survivability of wireless ad hoc networks with node misbehaviors and failures,” *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 3, pp. 284–299, 2008.

[10] 3rd Generation Partnership Project. Universal terrestrial radio access (UTRA) and evolved universal terrestrial radio access (E-UTRA); radio measurement collection for minimization of drive tests (MDT); overall description; stage 2 (release 10), 3GPP standard ts 37.320, version 10.2.0, tech. rep., june 2011.

[11] B. Hussain, Q. Du, A. Imran, and M. A. Imran, "Artificial intelligence-powered mobile edge computing-based anomaly detection in cellular networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 4986–4996, 2019.

[12] T. Zhang, K. Zhu, and D. Niyato, "Detection of sleeping cells in self-organizing cellular networks: An adversarial auto-encoder method," *IEEE Transactions on Cognitive Communications and Networking*, 2021.

[13] U. Masood, A. Asghar, A. Imran, and A. N. Mian, "Deep learning based detection of sleeping cells in next generation cellular networks," in *2018 IEEE Global Communications Conference (GLOBECOM).* IEEE, 2018, pp. 206–212.

[14] I. de-la Bandera, R. Barco, P. Munoz, and I. Serrano, "Cell outage detection based on handover statistics," *IEEE Communications Letters*, vol. 19, no. 7, pp. 1189–1192, 2015.

[15] A. Zoha, A. Saeed, A. Imran, M. A. Imran, and A. Abu-Dayya, "A learning-based approach for autonomous outage detection and coverage optimization," *Transactions on Emerging Telecommunications Technologies*, vol. 27, no. 3, pp. 439–450, 2016.

[16] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, "A cell outage detection algorithm using neighbor cell list reports," in *International Workshop on Self-Organizing Systems.* Springer, 2008, pp. 218–229.

[17] M. Alias, N. Saxena, and A. Roy, "Efficient cell outage detection in 5G hetnets using hidden markov model," *IEEE Communications Letters*, vol. 20, no. 3, pp. 562–565, 2016.

[18] A. Gómez-Andrades, P. Muñoz, I. Serrano, and R. Barco, "Automatic root cause analysis for LTE networks based on unsupervised techniques," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2369–2386, 2016.

[19] R. Barco, V. Wille, and L. Díez, "System for automated diagnosis in cellular networks based on performance indicators," *European Transactions on Telecommuni-*

cations, vol. 16, no. 5, pp. 399–409, 2005.

[20] R. Barco, P. Lazaro, and P. Munoz, “A unified framework for self-healing in wireless networks,” *IEEE Communications Magazine*, vol. 50, no. 12, pp. 134–142, 2012.

[21] D. Mulvey, C. H. Foh, M. A. Imran, and R. Tafazolli, “Cell fault management using machine learning techniques,” *IEEE Access*, vol. 7, pp. 124 514–124 539, 2019.

[22] T. Zhang, K. Zhu, and D. Niyato, “A generative adversarial learning-based approach for cell outage detection in self-organizing cellular networks,” *IEEE Wireless Communications Letters*, vol. 9, no. 2, pp. 171–174, 2019.

[23] P.-C. Lin, “Large-scale and high-dimensional cell outage detection in 5G self-organizing networks,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).* IEEE, 2019, pp. 8–12.

[24] O. Onireti, A. Zoha, J. Moysen, A. Imran, L. Giupponi, M. A. Imran, and A. Abu-Dayya, “A cell outage management framework for dense heterogeneous networks,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, 2015.

[25] U. S. Hashmi, A. Darbandi, and A. Imran, “Enabling proactive self-healing by data mining network failure logs,” in *2017 International Conference on Computing, Networking and Communications (ICNC)*, 2017, pp. 511–517.

[26] Y. Wang, K. Zhu, M. Sun, and Y. Deng, “An ensemble learning approach for fault diagnosis in self-organizing heterogeneous networks,” *IEEE Access*, vol. 7, pp. 125 662–125 675, 2019.

[27] S. Bothe, U. Masood, H. Farooq, and A. Imran, “Neuromorphic AI empowered root cause analysis of faults in emerging networks,” in *2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom).* IEEE, 2020, pp. 1–6.

[28] J. B. Porch, C. H. Foh, H. Farooq, and A. Imran, “Machine learning approach for automatic fault detection and diagnosis in cellular networks,” in *2020 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom).* IEEE, 2020, pp. 1–5.

[29] K.-F. Chen, C.-H. Lin, M.-C. Lee, and T.-S. Lee, “Deep learning-based multi-fault

diagnosis for self-organizing networks," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[30] Forsk. Atoll overview, [online], available: http://www.forsk.com/atoll-overview. [accessed: 12-oct-2018]. [Online]. Available: https://www.forsk.com/atoll-overview

[31] P. Muñoz, I. de la Bandera, E. J. Khatib, A. Gómez-Andrades, I. Serrano, and R. Barco, "Root cause analysis based on temporal analysis of metrics toward self-organizing 5G networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2811–2824, 2017.

[32] H. N. Qureshi, A. Imran, and A. Abu-Dayya, "Enhanced MDT-based performance estimation for AI driven optimization in future cellular networks," *IEEE Access*, vol. 8, pp. 161 406–161 426, 2020.

[33] J. D. Naranjo, A. Ravanshid, I. Viering, R. Halfmann, and G. Bauch, "Interference map estimation using spatial interpolation of MDT reports in cognitive radio networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2014, pp. 1496–1501.

[34] F. Sohrabi and E. Kuehn, "Construction of the RSRP map using sparse MDT measurements by regression clustering," in *2017 IEEE international conference on communications (ICC)*. IEEE, 2017, pp. 1–6.

[35] H. Braham, S. B. Jemaa, G. Fort, E. Moulines, and B. Sayrac, "Fixed rank kriging for cellular coverage analysis," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4212–4222, 2016.

[36] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[37] A. Regensky, S. Grosche, J. Seiler, and A. Kaup, "Real-time frequency selective reconstruction through register-based argmax calculation," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[38] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[39] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.

[40] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2691–2699.

[41] N. Memon, S. B. Patel, and D. P. Patel, "Comparative analysis of artificial neural network and xgboost algorithm for polsar image classification," in *International Conference on Pattern Recognition and Machine Intelligence.* Springer, 2019, pp. 452–460.

[42] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "Convxgb: A new deep learning model for classification problems based on cnn and xgboost," *Nuclear Engineering and Technology*, vol. 53, no. 2, pp. 522–531, 2021.

[43] W. Jiao, X. Hao, and C. Qin, "The image classification method with cnn-xgboost model based on adaptive particle swarm optimization," *Information*, vol. 12, no. 4, p. 156, 2021.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning.* PMLR, 2015, pp. 448–456.

[45] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[46] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *International Conference on Machine Learning.* PMLR, 2017, pp. 3780–3788.

[47] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier, "Regret analysis for performance metrics in multi-label classification," in *Proceedings of the 21st European Conference on Machine Learning*, pp. 280–295.

[48] SKlearn. https://scikit-learn.org, [online], available: [accessed: 26-march-2022]. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html

[49] G. Ciocarlie, U. Lindqvist, K. Nitz, S. Nováczki, and H. Sanneck, "On the feasibility

of deploying cell anomaly detection in operational cellular networks," in *2014 IEEE Network Operations and Management Symposium (NOMS)*. IEEE, 2014, pp. 1–6.

[50] G. F. Ciocarlie, U. Lindqvist, S. Nováczki, and H. Sanneck, "Detecting anomalies in cellular networks using an ensemble method," in *Proceedings of the 9th International Conference on Network and Service Management (CNSM 2013)*, 2013, pp. 171–174.

[51] J. Han, T. Xia, D. Spathis, E. Bondareva, C. Brown, J. Chauhan, T. Dang, A. Grammenos, A. Hasthanasombat, A. Floto *et al.*, "Sounds of COVID-19: exploring realistic performance of audio-based digital testing," *NPJ digital medicine*, vol. 5, no. 1, p. 16, 2022.

[52] W. Z. Khan, F. Azam, M. K. Khan *et al.*, "Deep Learning Based COVID-19 Detection: Challenges and Future Directions," *IEEE Transactions on Artificial Intelligence*, 2022.

[53] L. H. Nguyen, N. T. Pham, L. T. Nguyen, T. T. Nguyen, H. Nguyen, N. D. Nguyen, T. T. Nguyen, S. D. Nguyen, A. Bhatti, C. P. Lim *et al.*, "Fruit-cov: An efficient vision-based framework for speedy detection and diagnosis of sars-cov-2 infections through recorded cough sounds," *Expert Systems with Applications*, vol. 213, p. 119212, 2023.

[54] A. Ijaz, M. Nabeel, U. Masood, T. Mahmood, M. S. Hashmi, I. Posokhova, A. Rizwan, and A. Imran, "Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey," *Informatics in Medicine Unlocked*, p. 100832, 2022.

[55] L. Wijayasingha and J. A. Stankovic, "Robustness to noise for speech emotion classification using cnns and attention mechanisms," *Smart Health*, vol. 19, p. 100165, 2021.

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

[57] M. S. Riaz, H. N. Qureshi, U. Masood, A. Rizwan, A. Abu-Dayya, and A. Imran, "A hybrid deep learning-based (HYDRA) framework for multifault diagnosis using sparse MDT reports," *IEEE Access*, vol. 10, pp. 67 140–67 151, 2022.

[58] A. Mathur, T. Zhang, S. Bhattacharya, P. Velickovic, L. Joffe, N. D. Lane, F. Kawsar, and P. Lió, "Using deep data augmentation training to address soft-

ware and hardware heterogeneities in wearable and smartphone sensing devices," in *2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN).* IEEE, 2018, pp. 200–211.

[59] S. Žižek, *Pandemic!: COVID-19 shakes the world.* John Wiley & Sons, 2020.

[60] R. J. Barro, J. F. Ursúa, and J. Weng, "The coronavirus and the great influenza pandemic: Lessons from the "spanish flu" for the coronavirus's potential effects on mortality and economic activity," National Bureau of Economic Research, Tech. Rep., 2020.

[61] UNCTAD, "Global economy could lose over $4 trillion due to COVID-19 impact on tourism | UNCTAD," Accessed on: Aug. 10, 2021 [Online] Available: https://unctad.org/news/global-economy-could-lose-over-4-trillion-due-covid-19-impact-tourism.

[62] W.-K. Wu, J.-M. Liou, C.-C. Hsu, Y.-H. Lin, and M.-S. Wu, "Pandemic preparedness in Taiwan," *Nature biotechnology*, vol. 38, no. 8, pp. 932–933, 2020.

[63] P. Van den Driessche, "Reproduction numbers of infectious disease models," *Infectious Disease Modelling*, vol. 2, no. 3, pp. 288–303, 2017.

[64] M. d. P. M. H. A. Biswas, L. T. Paiva, "A SEIR model for control of infectious diseases with constraints," *Mathematical Biosciences Engineering*, vol. 11, no. 4, pp. 761–784, 2014.

[65] H. Berestycki, J.-M. Roquejoffre, and L. Rossi, "Propagation of epidemics along lines with fast diffusion," *Bulletin of Mathematical Biology*, vol. 83, no. 1, pp. 1–34, 2021.

[66] P. Delamater, E. Street, T. Leslie, Y. T. Yang, and K. Jacobsen, "Complexity of the Basic Reproduction Number (R0)," *Emerging Infectious Disease journal*, vol. 25, no. 1, p. 1, 2019. [Online]. Available: https://wwwnc.cdc.gov/eid/article/25/1/17-1901_article

[67] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system in four weeks," *MedRxiv*, 2020.

[68] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis," *arXiv preprint arXiv:2003.05037*, 2020.

[69] A. Abdel-Ghani, Z. Abughazzah, M. Akhund, K. Abualsaud, and E. Yaacoub, "Efficient pandemic infection detection using wearable sensors and machine learning," in *2023 International Wireless Communications and Mobile Computing (IWCMC).* IEEE, 2023, pp. 1562–1567.

[70] M. Iosa, P. Picerno, S. Paolucci, and G. Morone, "Wearable inertial sensors for human movement analysis," *Expert review of medical devices*, vol. 13, no. 7, pp. 641–659, 2016.

[71] A. G. Pacheco, F. A. Cabello, A. M. Fonoff, P. G. Rodrigues, O. A. Penatti, and P. R. Pinto, "Towards Low-Power Heart Rate Estimation Based on User's Demographics and Activity Level For Wearables," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, pp. 1–5.

[72] D. C. Han, H. J. Shin, S. H. Yeom, and W. Lee, "Wearable human health-monitoring band using inkjet-printed flexible temperature sensor," *Journal of Sensor Science and Technology*, vol. 26, no. 5, pp. 301–305, 2017.

[73] M. B. Alsabek, I. Shahin, and A. Hassan, "Studying the Similarity of COVID-19 Sounds based on Correlation Analysis of MFCC," in *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI).* IEEE, 2020, pp. 1–5.

[74] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–28, 2022.

[75] T. J. Parsons, C. Sartini, P. Welsh, N. Sattar, S. Ash, L. T. Lennon, S. G. Wannamethee, I.-M. Lee, P. H. Whincup, and B. J. Jefferis, "Objectively measured physical activity and cardiac biomarkers: A cross sectional population based study in older men," *International Journal of Cardiology*, vol. 254, pp. 322–327, 2018.

[76] D. Thomas, C. Bouchard, T. Church, C. Slentz, W. Kraus, L. Redman, C. Martin, A. Silva, M. Vossen, K. Westerterp *et al.*, "Why do individuals not lose more weight

from an exercise intervention at a defined dose? an energy balance analysis," *Obesity Reviews*, vol. 13, no. 10, pp. 835–847, 2012.

[77] M. Abu Zaid, J. Wu, C. Wu, B. R. Logan, J. Yu, C. Cutler, J. H. Antin, S. Paczesny, and S. W. Choi, "Plasma biomarkers of risk for death in a multicenter phase 3 trial with uniform transplant characteristics post–allogeneic hct," *Blood, the Journal of the American Society of Hematology*, vol. 129, no. 2, pp. 162–170, 2017.

[78] N. Brasier and J. Eckstein, "Sweat as a source of next-generation digital biomarkers," *Digital Biomarkers*, vol. 3, no. 3, pp. 155–165, 2019.

[79] A. Elhakeem, R. Cooper, P. Whincup, S. Brage, D. Kuh, and R. Hardy, "Physical activity, sedentary time, and cardiovascular disease biomarkers at age 60 to 64 years," *Journal of the American Heart Association*, vol. 7, no. 16, p. e007459, 2018.

[80] E. M. Simonsick, H. C. Meier, N. C. Shaffer, S. A. Studenski, and L. Ferrucci, "Basal body temperature as a biomarker of healthy aging," *Age*, vol. 38, pp. 445–454, 2016.

[81] J. Levy, D. Álvarez, A. A. Rosenberg, A. Alexandrovich, F. Del Campo, and J. A. Behar, "Digital oximetry biomarkers for assessing respiratory function: standards of measurement, physiological interpretation, and clinical use," *NPJ digital medicine*, vol. 4, no. 1, p. 1, 2021.

[82] S. Baliga, S. Muglikar, and R. Kale, "Salivary ph: A diagnostic biomarker," *Journal of Indian Society of Periodontology*, vol. 17, no. 4, p. 461, 2013.

[83] A. H. Gazi, N. Z. Gurel, K. L. Richardson, M. T. Wittbrodt, A. J. Shah, V. Vaccarino, J. D. Bremner, and O. T. Inan, "Digital cardiovascular biomarker responses to transcutaneous cervical vagus nerve stimulation: state-space modeling, prediction, and simulation," *JMIR mHealth and uHealth*, vol. 8, no. 9, p. e20488, 2020.

[84] M. Birch-Machin, E. Russell, and J. Latimer, "Mitochondrial DNA damage as a biomarker for ultraviolet radiation exposure and oxidative stress," *British Journal of Dermatology*, vol. 169, no. s2, pp. 9–14, 2013.

[85] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[86] S. Hagan, E. Martin, and A. Enríquez-de Salamanca, "Tear fluid biomarkers in ocular and systemic disease: potential use for predictive, preventive and personalised medicine," *Epma Journal*, vol. 7, pp. 1–20, 2016.

[87] V. Kozitsin, I. Katser, and D. Lakontsev, "Online forecasting and anomaly detection based on the ARIMA model," *Applied Sciences*, vol. 11, no. 7, p. 3194, 2021.

[88] M. Canizo, I. Triguero, A. Conde, and E. Onieva, "Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study," *Neurocomputing*, vol. 363, pp. 246–260, 2019.

[89] T. Tziolas, K. Papageorgiou, T. Theodosiou, E. Papageorgiou, T. Mastos, and A. Papadopoulos, "Autoencoders for anomaly detection in an industrial multivariate time series dataset," *Engineering Proceedings*, vol. 18, no. 1, p. 23, 2022.

[90] K. Menda, L. Laird, M. J. Kochenderfer, and R. S. Caceres, "Explaining COVID-19 outbreaks with reactive SEIRD models," *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.

[91] R. H. Ramadan and M. S. Ramadan, "Prediction of highly vulnerable areas to COVID-19 outbreaks using spatial model: Case study of Cairo Governorate, Egypt," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 25, no. 1, pp. 233–247, 2022.

[92] A. D. Hossain, J. Jarolimova, A. Elnaiem, C. X. Huang, A. Richterman, and L. C. Ivers, "Effectiveness of contact tracing in the control of infectious diseases: a systematic review," *The Lancet Public Health*, 2022.

[93] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, 1 2009.

[94] P. Espinosa, P. Quirola-Amores, and E. Teran, "Application of a Susceptible, Infectious, and/or Recovered (SIR) Model to the COVID-19 Pandemic in Ecuador," *Frontiers in Applied Mathematics and Statistics*, vol. 6, p. 55, 11 2020.

[95] I. Korolev, "Identification and estimation of the SEIRD epidemic model for COVID-19," *Journal of Econometrics*, vol. 220, p. 63, 1 2021. [Online]. Available: /pmc/articles/PMC7392128//pmc/articles/PMC7392128/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7392128/

[96] R. Schlickeiser and M. Kröger, "Analytical modeling of the temporal evolution of epidemics outbreaks accounting for vaccinations," *Physics*, vol. 3, no. 2, pp. 386–426, 2021.

[97] S. K. Yadav and Y. Akhter, "Statistical Modeling for the Prediction of Infectious Disease Dissemination With Special Reference to COVID-19 Spread," *Frontiers in Public Health*, vol. 9, p. 680, 6 2021.

[98] M. Marzouk, N. Elshaboury, A. Abdel-Latif, and S. Azab, "Deep learning model for forecasting COVID-19 outbreak in Egypt," *Process Safety and Environmental Protection*, vol. 153, p. 363, 9 2021. [Online]. Available: /pmc/articles/PMC8305306//pmc/articles/PMC8305306/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8305306/

[99] R. Kafieh, R. Arian, N. Saeedizadeh, Z. Amini, N. D. Serej, S. Minaee, S. K. Yadav, A. Vaezi, N. Rezaei, and S. H. Javanmard, "COVID-19 in Iran: Forecasting Pandemic Using Deep Learning," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[100] L. Wang, A. Adiga, S. Venkatramanan, J. Chen, B. Lewis, and M. Marathe, "Examining Deep Learning Models with Multiple Data Sources for COVID-19 Forecasting," *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, pp. 3846–3855, 12 2020.

[101] M. Panja, T. Chakraborty, S. S. Nadim, I. Ghosh, U. Kumar, and N. Liu, "An ensemble neural network approach to forecast dengue outbreak based on climatic condition," *Chaos, Solitons Fractals*, vol. 167, p. 113124, 2 2023.

[102] W. Liu, Q. Dai, J. Bao, W. Shen, Y. Wu, Y. Shi, K. Xu, J. Hu, C. Bao, and X. Huo, "Influenza activity prediction using meteorological factors in a warm temperate to subtropical transitional zone, eastern china," *Epidemiology and Infection*, vol. 147, 2019. [Online]. Available: /pmc/articles/PMC7006024//pmc/articles/PMC7006024/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7006024/

[103] S. Engebretsen, K. Engø-Monsen, M. A. Aleem, E. S. Gurley, A. Frigessi, and B. F. D. Blasio, "Time-aggregated mobile phone mobility data are sufficient for modelling influenza spread: the case of bangladesh," *Journal of the Royal Society Interface*, vol. 17, 6 2020. [Online]. Available: /pmc/articles/PMC7328378//pmc/articles/PMC7328378/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7328378/

[104] L. Wang, A. Adiga, J. Chen, A. Sadilek, S. Venkatramanan, and M. Marathe, "CausalGNN: Causal-Based Graph Neural Networks for Spatio-Temporal Epidemic Forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 12 191–12 199, 6 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21479

[105] C. Fritz, E. Dorigatti, and D. Rügamer, "Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly COVID-19 cases in Germany," *Scientific Reports 2022 12:1*, vol. 12, pp. 1–18, 3 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-07757-5

[106] G. Panagopoulos, G. Nikolentzos, and M. Vazirgiannis, "Transfer graph neural networks for pandemic forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 4838–4845, 5 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16616

[107] S. Yu, F. Xia, S. Li, M. Hou, and Q. Z. Sheng, "Spatio-temporal graph learning for epidemic prediction," *ACM Transactions on Intelligent Systems and Technology*, 4 2023.

[108] T. S. Hy, V. B. Nguyen, L. Tran-Thanh, and R. Kondor, "Temporal multiresolution graph neural networks for epidemic prediction," pp. 21–32, 7 2022. [Online]. Available: https://proceedings.mlr.press/v184/hy22a.html

[109] Y. Ma, P. Gerard, Y. Tian, Z. Guo, and N. V. Chawla, "Hierarchical spatio-temporal graph neural networks for pandemic forecasting," *International Conference on Information and Knowledge Management, Proceedings*, pp. 1481–1490, 10 2022. [Online]. Available: https://dl.acm.org/doi/10.1145/3511808.3557350

[110] CDC, "Describing Epidemiologic Data | Epidemic Intelligence Service | CDC," Accessed on: June 30, 2022 [Online]. Available:https://www.cdc.gov/eis/field-epi-manual/chapters/Describing-Epi-Data.html.

[111] N. Waters, "Tobler's first law of geography," *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pp. 1–13, 3 2017.

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

[113] US Census Bureau, “Understanding Geographic Identifiers (GEOIDs),” Accessed on: June 30, 2022 [Online]. Available:https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html.

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

[115] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 7 2017. [Online]. Available: https://arxiv.org/abs/1707.01926v3

[116] H. V. Pham, D. H. Thanh, and P. Moore, “Hierarchical pooling in graph neural networks to enhance classification performance in large datasets,” *Sensors (Basel, Switzerland)*, vol. 21, 9 2021. [Online]. Available: /pmc/articles/PMC8472962//pmc/articles/PMC8472962/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8472962/

[117] J. Lee, I. Lee, and J. Kang, “Self-attention graph pooling,” *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 6661–6670, 4 2019. [Online]. Available: https://arxiv.org/abs/1904.08082v4

[118] A. Das, N. Borisov, and M. Caesar, “Do you hear what i hear? fingerprinting smart devices through embedded acoustic components,” in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 441–452.

[119] J. I. Hall, M. Lozano, L. Estrada-Petrocelli, S. Birring, and R. Turner, “The present and future of cough counting tools,” *Journal of Thoracic Disease*, vol. 12, no. 9, p. 5207, 2020.

[120] E. Por, M. v. Kooten, and V. Sarkovic, “Nyquist–shannon sampling theorem,” *Leiden University*, vol. 1, no. 1, 2019.

[121] T. Smirnova, “Comparative analysis of modern formats of lossy audio compression,” 2020.

[122] F. J. Harris, *Multirate signal processing for communication systems.* CRC Press, 2022.

[123] J. Sterne, *MP3: The meaning of a format.* Duke University Press, 2012.