# PRIVACY AND SECURITY OF EMERGING TECHNOLOGIES IN CHANGING HEALTHCARE PARADIGMS

A Dissertation

by

SHALINI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

| | |
|---|---|
| Chair of Committee, | Nitesh Saxena |
| Committee Members, | Narasimha Annapareddy |
| | Theodora Chaspari |
| | Drew Hamilton |
| Head of Department, | Scott Schaefer |

August 2023

Major Subject: Computer Science

# ABSTRACT

Emerging technologies are revolutionizing healthcare by improving accessibility and enhancing patient-provider interactions. With advancements in medical research and computational resources, diverse healthcare technologies have played a critical role in combating pandemics like COVID-19, especially in situations with restricted access or unavailability of providers. The shift towards telemedicine and Precision Medicine offers new ways to access healthcare, but also raises significant privacy and security concerns.

This dissertation focuses on the vulnerability of predatory research that compromise the integrity of research literature based Medical AI solutions, Mobile Healthcare apps, and Voice-based systems in healthcare. In the changing era of healthcare and social interactions, technological advancements have the potential to bring immense benefits and shape the future of medical research and practices. However, the practical adaptation of these practices is challenging, as users may fear potential violations of their privacy and security. Violations of sensitive mental and reproductive health data may create uncertainty and endanger women's health in the changing legal landscape. Our work is especially crucial with medical practices that are increasingly reliant on technology to manage health and privacy and security concerns are growing more pressing.

Our dissertation contributes to the field of emerging healthcare technologies by analyzing the potential threats and recommended defense in Medical AI solutions, Mobile Health Apps, and Voice-based Systems. Specifically, we identify the vulnerability of Medical AI solutions to predatory research, analyze privacy and security threats in Mobile Health Apps, and identify the vulnerabilities of Voice-based systems to break the speaker's anonymity. Furthermore, our dissertation highlights the criticality of multi-faceted modern medicine practices that use a combination of healthcare technologies to improve patient outcomes. These technologies can help providers make informed decisions and empower patients to stay motivated and feel cared for. By promoting the development of secure and trustworthy healthcare technologies and practices, we aim to enhance patient-provider interactions and improve the overall quality and safety of healthcare delivery.

# DEDICATION

To my parents (Dr. Mahavir Singh and Mrs. Nirmal), family, friends, amazing mentors, inspiring strangers, and daughter Shruti!

# ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Professor Nitesh Saxena, without whom this endeavor would not have been possible. His research expertise, wisdom, guidance and outstanding leadership has helped me to grow as a researcher in the field of cybersecurity. His experience and endless dedication motivated me to push myself further and remain focused on completing quality research. I will be forever indebted for his mentorship and support of my development and growth. I would also like to thank my other committee members Professor Narasimha Annapareddy, Professor Theodora Chaspari and Professor Drew Hamilton. Their valuable feedback and continued positive encouragement greatly helped me for my thesis work. I would also like to thank all of my colleagues in the SPIES research group for their help during my research. Finally, I am sincerely grateful to my family and friends who continued to support me throughout my research and studies. Their constant encouragement gave me the strength I needed to persevere through the rough times and complete this work.

I would also like to acknowledge the National Science Foundation (NSF) for their financial support for this work.

# CONTRIBUTORS AND FUNDING SOURCES

## Contributors

This work was supported by a dissertation committee consisting of Professor Nitesh Saxena (Advisor), Professor Theodora Chaspari, Professor Drew Hamilton of the Department of Computer Science, and Professor Narasimha Annapareddy of the Department of Electrical and Computer Engineering.

All the work conducted for the dissertation was completed by the student independently.

## Funding Sources

This work was partially supported by NSF funding. Graduate study was supported by an assistantship from Texas A&M University, College Station, TX.

# NOMENCLATURE

| | |
|---|---|
| NIH | National Institute of Health |
| NLM | National Library of Medicine |
| CDC | Centers for Disease Control and Prevention |
| CUI | Concept Unique Identifier |
| GUI | Graphical User Interface |
| HPRC | High Performance Research Cluster |
| DNN | Deep Neural Network |
| CNN | Convolutional Neural Network |
| ML | Machine Learning |
| GMM | Gaussian Mixture Model |
| GAN | Generative Adversarial Network |
| PPG | Phonetic Posteriorgram |
| MFCC | Mel-frequency Cepstrum Coefficient |
| FFT | Fast Fourier Transform |
| TTS | Text-to-Speech |
| STT | Speech-to-Text |
| VC | Voice Conversion |
| VCC | Voice Conversion Challenge |
| ASR | Automatic Speech Recognition |
| SR | Speaker Recognition |
| SOTA | State-of-the-art |
| VA | Virtual Assistant |

| | |
|---|---|
| MEDAI | Medical Artificial Intelligence |
| AI | Artificial Intelligence |
| NLP | Natural Language Processing |
| KG | Knowledge Graph |
| PMID | PubMed Reference Index |
| UMLS | Unified Medical Language System |
| OA | Open Access |
| PPP | Predatory Publication Presence |
| APC | Article Processing Charge |
| MMH | Mobile Mental Health |
| Femtech | Female Technology |
| OWASP | The Open Web Application Security Project |
| FTC | Federal Trade Commission |
| HIPAA | The Health Insurance Portability and Accountability Act |
| PII | Personal Identifiable Information |
| MITM | Man-In-The-Middle Attack |

# TABLE OF CONTENTS

# LIST OF FIGURES

FIGURE Page

# LIST OF TABLES

TABLE Page

# 1. INTRODUCTION

A healthcare paradigm shift towards telemedicine and Precision Medicine involves significant technological advancements that offer new ways to interact in patient-provider settings. However, it also raises significant concerns about privacy and security issues, given the sensitive nature of medical information and the potential risks associated with unauthorized access or exposure. The integration of medical AI, mobile health apps, and voice-based technologies in modern medicine can result in a ripple effect of vulnerabilities that compromise patient outcomes and privacy.

Emerging technologies are transforming healthcare by enabling remote access to healthcare services through online platforms or mobile phones, with or without human interactions. Chatbots, automated service calls, and voice assistants are increasingly available to help the general population. With advancements in medical research and computational resources, diverse healthcare applications have played a critical role in combating pandemics like COVID-19. Online alternatives, mobile apps, and telemedicine offer self-guided tests, evaluations, and possible remedies in situations with limited access to providers. In the age of the Internet of Things (IoT), mobile phones are ubiquitous across the globe, making them a valuable resource for reaching out to a much larger and underserved population.

In the dynamic landscape of healthcare and social interaction, our focus is on identifying privacy and security vulnerabilities that can limit the practical applicability of advancements in Medical AI solutions, Mobile Healthcare apps, and Voice-based systems. While technological advancements can bring immense benefits and shape the future of healthcare, their practical adaptation in clinical settings is challenging due to the sensitive nature of healthcare data and the associated vulnerabilities of technological solutions. Therefore, we study and analyze the paradoxes of emerging technologies in healthcare to identify and measure vulnerabilities. By promoting the development of secure and trustworthy healthcare technologies and practices, we aim to ensure the practical applicability of these advancements in clinical settings and enhance the quality and safety of healthcare delivery.

## 1.1 Emerging Technologies in Modern Medicine and Research

Biomedical research aims to achieve affordable, quality healthcare for all, but inferring useful relationships can be challenging due to the vast amount of scattered knowledge in research literature. With technological advancements in data processing, knowledge extraction, and knowledge representation, practical use of research findings may improve overall healthcare.

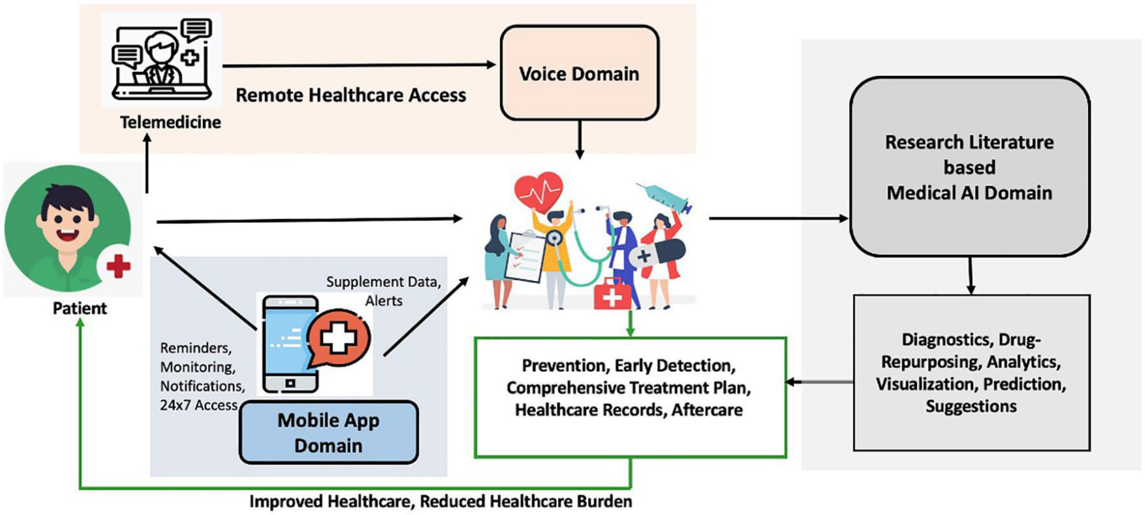

Telemedicine
Remote Healthcare Access
Voice Domain
Research Literature based Medical AI Domain
Patient
Supplement Data, Alerts
Reminders, Monitoring, Notifications, 24x7 Access
Mobile App Domain
Prevention, Early Detection, Comprehensive Treatment Plan, Healthcare Records, Aftercare
Diagnostics, Drug-Repurposing, Analytics, Visualization, Prediction, Suggestions
Improved Healthcare, Reduced Healthcare Burden

Figure 1.1: High-Level Overview of Emerging Technologies in Healthcare.

Figure 1.1 depicts the interdependent usage of emerging technologies in modern healthcare. Advanced AI-based diagnostic tools have the potential to identify rare diseases, while other AI-based solutions can infer knowledge to identify FDA-approved drugs for innovative treatments. Although this can provide patients and providers with confidence in the treatment, the process can be complex and challenging. To support patients, virtual voice-based assistants can keep them informed and motivated to follow through intervention cycles, while mobile healthcare apps can provide 24/7 support through supplement data collection, continuous monitoring, critical reminders, and communication channels.

Validated research builds trust among healthcare professionals and patients when considering interventions, and testing interventions in real-world clinical settings generates additional data that can further validate research hypotheses. This continuous cycle of research and implementation updates can lead to the emergence of new research ideas and questions based on clinical observations and patient data. For instance, telemedicine and voice-inputs generate patient health records that can be used to research unknown diagnoses or treatments, and advanced technologies can be applied to establish new relationships and enrich the data. We can leverage the power of advanced technologies to analyze the data and generate new insights that can improve patient outcomes.

## 1.2 Threats To Emerging Technologies

There is a growing need for the practical implementation of AI-based, voice-based, and mobile-based medical solutions that address the multifaceted nature of modern medical practices. The interconnected nature of modern medicine motivates our work to identify the threats and vulnerabilities in each of these components. While each component has its individual significance and contribution to the modern healthcare paradigm, collectively, their contribution is much more significant, and so is the potential threat to patient care.

Predatory research can pollute the pool of genuine research data and findings, which threatens the trust in overall research inputs used for knowledge extraction and decision-making in healthcare. A manipulated AI-based solution can misrepresent output, leading to incorrect diagnosis and subsequent treatment suggestions. Ensuring research data integrity is critical for advanced technological solutions due to the increasing threat of predatory research-induced data pollution, as even a single unvalidated research finding may have a significant impact on patients, underscoring the importance of validating the authenticity and reliability of research publications.

Mobile-based and voice-based solutions have the potential to transform healthcare by enabling continuous data collection, monitoring, and remote access to healthcare services. However, the misuse or exploitation of these technologies can have detrimental effects on patient outcomes. For example, a malfunctioning or hijacked mobile app can produce incorrect supplement data that affects decisions on drug dosage and frequency. Similarly, missing reminders can lead to missed medical assistance, delayed recovery, and relapses of medical conditions. In addition, a manipulated voice assistant can cause incorrect guidelines or misinterpreted communication, potentially resulting in under or overdose of medicines, missed or incorrect appointment scheduling, misinterpretation, and incorrect classification of voice-notes. These errors can lead to incorrect processing of information, causing patient harm. The potential exploitation of vulnerabilities in these systems can diminish trust in the use of emerging technologies in healthcare, delaying their practical adaptation in clinical settings.

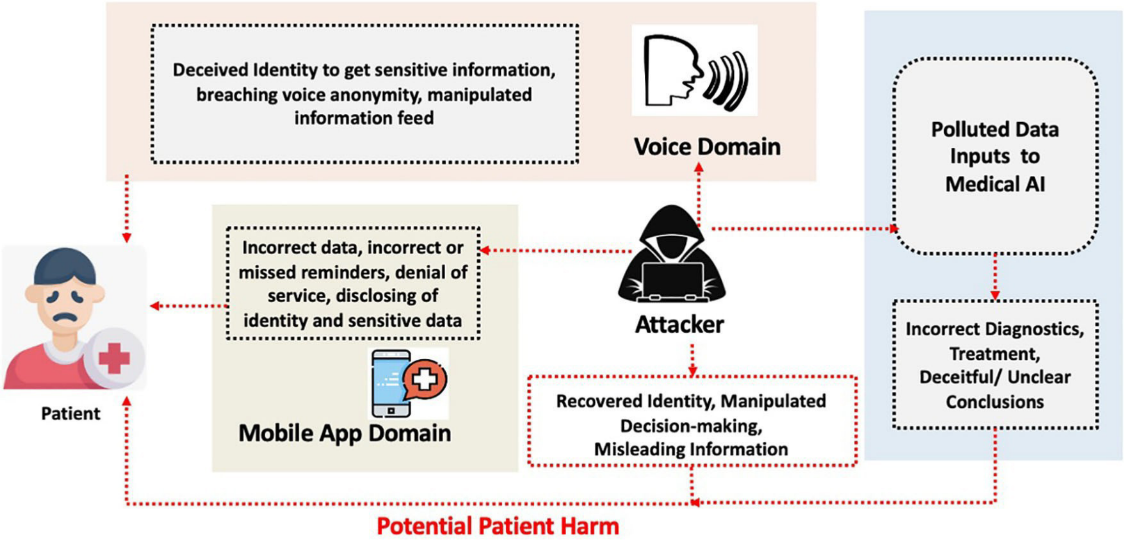

Deceived Identity to get sensitive information, breaching voice anonymity, manipulated information feed
Voice Domain
Polluted Data Inputs to Medical AI
Incorrect data, incorrect or missed reminders, denial of service, disclosing of identity and sensitive data
Attacker
Patient
Mobile App Domain
Incorrect Diagnostics, Treatment, Deceitful/ Unclear Conclusions
Recovered Identity, Manipulated Decision-making, Misleading Information
Potential Patient Harm

Figure 1.2: High-Level Overview of Threat Model: Emerging Technologies in Healthcare.

Figure 1.2 illustrates a threat model that highlights the potential vulnerabilities and risks in the three domains of research literature based MedAI decision-making, mobile-based and voice-based solutions. However, a breach in any of these domains can negatively impact all areas, potentially leading to patient harm since these domains collectively apply to a comprehensive treatment plan. We observe that emerging technologies are becoming essential pillars of modern medicine approaches, such as Precision Medicine and Personalized Medicine. However, privacy and security violations in any of these components can trigger the exploitation of linked medical, personal, social, and financial information, leading to patient harm and potential legal prosecution, especially following changing laws impacting specific and vulnerable populations.

Therefore, it is crucial to identify and address these vulnerabilities to ensure the secure and trustworthy use of emerging technologies in healthcare. By doing so, we can ensure the practical applicability of these technologies and enhance the quality and safety of healthcare delivery, while also protecting patient privacy and data security.

To mitigate these risks, it's essential to develop robust defense mechanisms that minimize the potential damage caused by privacy and security breaches. For example, healthcare organizations must implement strong data protection policies and ensure that all staff are trained in security best practices. Additionally, it's critical to be vigilant in identifying potential vulnerabilities and proactively addressing them to prevent attacks before they occur. By taking these steps, healthcare organizations can harness the power of advanced technologies while safeguarding patient privacy and security, mitigating the risks of unreliable information, and ensuring that future research and healthcare solutions are aligned towards positive outcomes.

## 1.3 Thesis Statement and Main Contributions

In this dissertation, we perform an exposition of privacy and security threats in emerging healthcare technologies. In the era of Precision Medicine, patient-centric tailored medical treatments require the collection and analysis of large amounts of patient data. Research literature based medical AI solutions can assist in finding the right diagnosis and treatment, while the intervention cycle can be facilitated by mobile-based and voice-based solutions. We firstly consider the threat of predatory research that seek to compromise the research literature based medical AI solutions via examining the state-of-the-art real-life Medical AI solutions for Predatory Publications Presence (PPP) in their inputs and output. We propose a defense by potentially classifying predatory publications and minimizing the data-poisoning attacks on Medical AI solutions. We begin by exploring the potential danger of predatory research that aims to undermine medical AI solutions by infiltrating the research literature. We examine real-life medical AI solutions to identify signs of Predatory Publication Presence (PPP) in their inputs and outputs. To defend against this threat, we suggest a strategy for identifying and categorizing predatory publications, which would help to reduce the risk of passive and active data-poisoning attacks on medical AI solutions.

Mobile healthcare apps are playing a vital role in digital healthcare management providing accessibility for needed assistance and continuous monitoring. Our work explores the privacy and security vulnerabilities of mobile mental health (MMH) apps and women's health apps. We analyze the heightened threats to already disadvantaged, underserved, and vulnerable populations. Impacted largely by COVID-19, remote healthcare access is increasingly involving voice-based interaction through computers, smartphones, and digital voice assistants. However, in situations where patient anonymity is crucial, maintaining confidentiality can be challenging, especially with voice recordings. We investigate the risks posed by voice anonymity attacks, which can uncover the speaker's identity and lead to privacy and security attacks.

The objective of this dissertation is to assess the potential success of attacks on medical AI, mobile apps, and voice-based technologies, considering recent advancements in these fields. By examining potential vulnerabilities and assessing the effectiveness of current security measures, this research aims to provide insights into how to effectively protect these technologies from cyber threats, ensuring their continued success in enhancing patient care. Our thesis is organized in three major parts: 1) threats of predatory research compromising research literature based Medical AI solutions, 2) privacy and security vulnerabilities of Mobile healthcare apps impacting vulnerable populations, and 3) Voice-based healthcare solutions compromising voice anonymity and posing privacy and security threats.

This work offers several notable contributions. Firstly, it identifies the potential risk of predatory research that can erode trust in research literature-based Medical AI solutions and proposed a viable defense mechanism to reduce the impact of predatory research-induced data pollution. Secondly, it draws attention to the heightened privacy and security risks posed to women's health and mental health apps in specific contexts. Finally, it uncovers vulnerabilities in emerging voice-based solutions to support the development of robust solutions for sensitive healthcare scenarios.

We present the specific contributions of this dissertation as follows:

### 1.3.1 Threat of Predatory Research to Research Literature-based MedAI Solutions:

- A comprehensive survey of the applicability of medical AI, predatory research, and vulnerabilities highlights a research gap on the impact of predatory research on MedAI solutions.

- We verify the threat of predatory research compromising MedAI solutions by demonstrating Predatory Publication Presence (PPP) in reputable research literature sources, derived databases, and real-world MedAI solutions.

- We propose a machine learning-based defense mechanism to minimize the data pollution caused by predatory research.

### 1.3.2 Privacy and Security Threats to Mobile Healthcare Apps:

- Study of Mobile Mental Health (MMH) apps as highly promising alternative shows the presence of exploitable privacy and security vulnerabilities.

- Demonstrating heightened threat to women's health apps including prosecution for their reproductive choices.

- Our work emphasizes the importance of categorizing mobile healthcare apps based on well-defined privacy and security guidelines, with a focus on accountability.

### 1.3.3 Privacy and Security Vulnerabilities of Voice-based Healthcare Solutions:

- Multiple voice-conversion techniques and speaker recognition methods demonstrate the potential breach of speaker's anonymity.

- Discussing the feasible defense strategies to protect speaker's anonymity.

# 2. BACKGROUND AND RELATED WORK

Emerging technologies are improving patient outcomes, accessibility to healthcare, and reducing healthcare burden by providing remote monitoring, personalized treatment, and real-time communication between patients and healthcare providers. Technological advancements in medical research, AI, data processing, voice synthesis, and smartphones are driving forces for current and future medical practices aiming for improved healthcare and reduced healthcare burden.

In the upcoming sections, we will explore the progression and developments in healthcare technologies and how they support medical practices, research, and patient care. However, it is also important to acknowledge the critical challenges related to the integrity, privacy, and security of these solutions.

## 2.1 Healthcare Revolution through Emerging Technologies

The integration of emerging technologies in the field of AI, telemedicine, mobile apps, wearables, and voice-based solutions is revolutionizing healthcare by facilitating the development of practical alternatives and support systems beyond the research stage. For instance, Medical AI can analyze data collected by mobile health apps and voice-based technologies to identify trends and patterns, which could help healthcare providers make more informed decisions about patient care. Mobile health apps can work in conjunction with voice-based technologies to provide more personalized and convenient healthcare services. For example, a mobile app could be used to collect patient data, which could then be analyzed by medical AI and interpreted by voice-based technologies to provide personalized healthcare recommendations. Voice-based technologies can provide a more natural and intuitive interface, which could make them more accessible to patients and healthcare providers alike.

The undeniable contribution of Artificial Intelligence (AI) in medicine is reaching new milestones. Since 1995, more than 500 AI/ML-enabled medical devices have been approved by the U.S. Food and Drug Administration (FDA), with over half in Radiology. In 2022 alone, FDA approved 91 out of 521 such devices [4]. The *IDx-DR v2.3* is the first FDA-approved human-independent AI system to detect Diabetic Retinopathy, a leading cause of blindness [5]. FDA also permitted marketing of the first mobile women's health app, *Natural Cycles*, as a method of contraception [6]. Google's AI *Inception v3* may assist with the early detection of lung cancer, the deadliest cancer, causing 1.7 million deaths per year globally [7].

AI is transforming drug discovery, and there is compelling evidence that medical AI can play a vital role in enhancing and complementing the *medical intelligence* of the future clinician [8]. According to a recent study, an intelligent machine learning algorithm was able to identify and predict harmful bacteria in blood with 95% accuracy. The study used a convolutional neural network to analyze blood samples and detect the presence of harmful bacteria, demonstrating the potential for machine learning to improve the accuracy and speed of medical diagnoses. [9]. In 2021, a report indicated that 230 startups are utilizing artificial intelligence in drug discovery to improve the success rate significantly [10].

Digital health management, mobile apps, virtual nurses, and voice-based technologies have the potential to enhance patient-provider interaction and provide better healthcare outcomes for individuals with different medical conditions [11]. A mobile health app is a convenient, low-cost, anonymous, and 24/7 resource that can help peoples in need. Interactive or game-based apps can generate user engagement, especially among younger generations, and encourage them to continue using the app [12]. Mental health apps, such as mood tracking or cognitive behavioral therapy apps, can supplement traditional in-person therapy sessions by providing additional data and insights [13]. For voice domain applications, *Molly*, a virtual nurse, engages users via text or voice to meet their needs, while voice-based technology can help visually impaired, physically disabled, speech impaired, hearing impaired, as well as the elderly peoples to communicate [14, 15, 16].

Despite these advances, chronic diseases and mental health issues contribute to 90% of the nation's $4.1 trillion in annual healthcare costs. Chronic diseases are the leading causes of death and disability in the United States [17]. Patients with rare diseases often face incorrect initial diagnoses and long waits, up to 30 years for correct diagnoses, leading to increased healthcare costs and patient suffering [18]. Medical research is crucial to establishing causes and symptoms of diseases, efficient diagnostics, and apt treatments to significantly reduce healthcare costs and improve patient outcomes. The NIH budget for 2024 is approximately $51 billion for medical research, supporting funding almost 50,000 competitive grants to over 300,000 researchers at more than 2,500 institutions throughout the United States of America [19, 20]. PubMed, the NIH-NLM research literature repository, which contains over 35 million citations and abstracts of biomedical literature, plays a critical role in identifying research gaps and shaping future research for healthcare improvements [21].

## 2.2 Threats to Emerging Technologies in Healthcare

The emergence of new technologies in healthcare has brought about significant benefits, including improved patient outcomes, increased efficiency, and reduced costs [20, 7, 5, 14, 6]. However, these technologies also face security and integrity threats that can compromise the quality of patient care and damage healthcare organizations' reputation. As advanced technologies are increasingly applied in healthcare, the need to identify and address privacy and security threats is paramount. While modern healthcare solutions hold great promise in terms of saving time, money, and effort, the cost of relying on *potentially unreliable* information from these solutions is too high to ignore. The consequences of an unreliable decision or erroneous results can be dire in clinical settings including jeopardizing patient safety, negatively impacting healthcare solutions, and misguided future research directions.

As medical AI, mobile health apps, and voice-based technologies become increasingly integrated in modern medicine, vulnerabilities in these technologies can have a ripple effect on each other. A vulnerability in a medical AI system could lead to inaccurate recommendations that are integrated into voice-based technologies, potentially leading to harmful outcomes for patients. Similarly, a vulnerability in a mobile health app that collects patient data could expose sensitive medical information that is used to train medical AI systems or to provide input for voice-based technologies, leading to biased recommendations. A vulnerability in voice-based technologies can compromise the security of patient data in mobile health apps, affecting the accuracy of medical AI systems that rely on that data. Therefore, robust security measures and privacy protections are critical to ensure the safe and effective use of medical AI, mobile health apps, and voice-based technologies in modern medicine.

There are multiple studies to demonstrate how subtle adversarial inputs can potentially change medical decisions significantly [22, 23]. Adversarial attacks on neural networks can cause errors in identifying cancer tumors and damage the confidence in Medical AI [24]. Given any dataset, an attacker can potentially perturb it in a direction that aligns well with the weights of the algorithm and thus amplifies its effect on the output [25]. Gaglio et al. showed that a minimal alteration of the clinical records can subvert predictions with high probability to fool a smart prescription system [26]. Newaz et al. evaluated that, with the partial knowledge of data distribution, model, and algorithm, an adversary can perform both targeted and untargeted attacks to alter patient status [27]. Generic and algorithm-independent attacks pose even a greater threat as those can be applied to a wide range of medical datasets and AI algorithms [2]. NLP is the core process to extract intelligent information from the research literature, a robust and efficient algorithm against NLP adversarial attacks is a necessity rather than preference [28].

The vulnerability of the data shared between mobile apps and the cloud storage can present a threat to privacy and security of users [29]. In a study conducted by O'Loughlin et al. on mobile apps for depression, it was found that 68% of the apps received unacceptable transparency scores on their data security and privacy policies [30]. Another study by Dehling et al. on both Android and iOS apps shows that the majority of apps (95.63%) pose some potential damage because of security and privacy violations [31]. Reardon et al. presented the covert and side-channel attacks on mobile apps by exploiting the Android permission model [32]. In 2023, the case of involving a period-tracking app being prohibited by the Federal Trade Commission (FTC) for selling users' personal data to advertisers without user's consent serves as an example of the potential risks associated with unregulated women's health apps [33]. A study by Alfawzan et al. (2019) observed that all women's health apps allow behavioral tracking and had poor privacy practices [34]. According to a study by Scherwitzl et al. (2017), the Natural Cycles app, the only FDA-approved natural contraceptive method, has a failure rate of 8.3%. Despite being highly effective compared to traditional methods, the app can still produce unintended outcomes [35]. A lack of contextual understanding of the importance of privacy and security may lead to the development of insecure apps. This could result in the exposure of sensitive health information, unintended data sharing, and potential harm to users. For example, Aljedaani et al. shows that 63% of mobile health apps' lack security because of absent security guidelines and regulations for developing secure mobile health apps. Developers of 56% apps may not have knowledge and expertise for secure mobile health app development [36].

In voice domain, voice data privacy and hacking are major concerns for smart speaker users worldwide. A research work found that 45% of users are worried about voice data privacy, and 42% are concerned about voice data hacking. Additionally, 59% of respondents identified privacy as an important consideration when using voice control devices [37]. On voice-based solutions' vulnerability, Wenger et al. found that synthetic speech can fool both humans and machines, and existing defenses against are not equipped to handle it [38]. Using cloud services in voice applications has significant disadvantages related to security, safety, and privacy concerns [39]. Qian et al. proposed *VoiceMask* to protect speech-privacy for cloud-based to mitigate the security and privacy risks exposed in cloud. Reversing attacks are possible in basic voice-conversion samples because the warping functions are invertible [40].

# 3. A SURVEY OF THREATS TO RESEARCH LITERATURE DEPENDENT MEDICAL AI SOLUTIONS*

Medical research is a fundamental component of modern medicine, and emerging technologies are providing researchers with powerful tools to validate research theories through different phases. For instance, AI-based systems can help fine-tune hypotheses and formulate research problems that may not have been previously explored. Advanced tools for data collection, extraction, and transformation, including web interfaces, web crawling, and NLP, are enabling researchers to gather and process large amounts of data efficiently. The analysis is powered by mathematical and statistical tools like MatLab, code libraries, APIs, and cloud-based solutions, with results presented through dynamic visualizations and language models. Sensors, devices, and instruments are increasingly efficient and can rapidly collect and share data to generate customized solutions with reasonable trust. However, the inter-dependency of components can impact the integrity of the output, highlighting the importance of ensuring the integrity of each component in the research process. By utilizing these emerging technologies while maintaining the integrity of each component, researchers can improve the efficiency and effectiveness of medical research, ultimately advancing healthcare for all.

Medical Artificial Intelligence (MedAI) has the potential to address healthcare challenges by harnessing the power of AI algorithms and vast data. However, ensuring the security, integrity, and credibility of MedAI tools is paramount to protect human lives. Predatory research, which exploits the *publish or perish* culture and *pay to publish* model, is a serious threat to the integrity of MedAI inputs, which can impact the trustworthiness of MedAI output. While it's challenging to measure the actual impact of predatory research on data pollution and patient harm, our work shows that the breached integrity of MedAI inputs is a critical issue. We present a comprehensive literature review that addresses the gap in understanding predatory research vulnerabilities affecting MedAI

*Reprinted with permissions from Saini and Saxena. Saini, S., & Saxena, N. (2023). A Survey of Threats to Research Literature Dependent Medical AI Solutions. ACM Computing Surveys. https://dl.acm.org/doi/10.1145/3592597 solutions, including threats of data pollution, feasible attacks on MedAI solutions, and the influence on healthcare. Our contribution is to help develop more robust MedAI solutions in the future by highlighting the importance of addressing predatory research vulnerabilities.

## 3.1 Introduction

Medical Artificial Intelligence (MedAI) for finding a correct diagnosis, treatments, and drug development represents the new age of healthcare. MedAI can be a valuable tool in Precision Medicine and Personalized Medicine, providing data-driven diagnosis and treatment guidance based on enormous amounts of data. Williams et al. (2018) provide compelling evidence that MedAI can help identify previously unknown connections among various medical factors [41]. Additionally, Ramesh et al. (2004) and Krittanawong et al. (2018) suggest that MedAI can enhance and complement the medical intelligence of future clinicians [8, 42].

### 3.1.1 How critical is Biomedical Research in Healthcare

Rare, unknown, or life-threatening diseases have been a great motivation for academic and industrial research for improving patient-centered solutions in research and clinical settings [43, 44]. Many clinical and non-clinical MedAI solutions rely upon scientific research publications in medicine as the primary data source for automated decision-making. *PubMed* [21], maintained by the National Library of Medicine (NLM), is one of the largest medical research literature repositories, comprising more than 30 million citations for biomedical literature. PMID (PubMed ID) is a unique identifier assigned to each article on PubMed. PubMed-derived database SemMedDB, consisting of 96.3 million predications extracted from all MEDLINE citations, integrating PubMed and Unified Medical Language System (UMLS), is an indispensable input to MedAI systems [45, 46, 47, 48, 49, 50, 51, 52]. MedAI solutions harness the power of research through analytical algorithms spanning Heuristics, Neural Networks, Natural Language Processing (NLP), Fuzzy Logic, Semantic Analysis, Knowledge Graphs (KGs), and Machine Learning (ML) to integrate and interpret the complex biomedical and healthcare data [45, 46, 47, 52].

### 3.1.2 How Research gets manipulated, and can Impact the Medical AI and Future Research

Undoubtedly, research literature adds enormous value to accelerate innovations in clinical care and drug discovery. However, research literature repositories can be prone to data abuse through undetected fraudulent research. As research publications serve as a core component of MedAI, it draws attention to research literature and any derived database as a potential attack surface. Invalidated or manipulated academic or industrial research is commonly referred to as *Predatory Science* [53]. In recent years, there has been an unprecedented rise in predatory science publications interfering with genuine research [54, 55, 56, 57, 58, 53]. There may even be a possibility of affecting the actual patient care based on such fraudulent predatory research [59, 60]. Such unreliable data then can be further abused by malicious actors through exploratory or adversarial attacks to compromise the credibility of MedAI solutions [22, 1, 2, 28]. If the inputs are untrustworthy, irrespective of the efficiency of the applied algorithm, the output of such a system cannot be considered as trustworthy [61, 41, 62]. In the case of rare diseases where research is limited, unreliable output from MedAI can be particularly problematic. This can distract service providers, increase costs and effort, and may lead to treatments that are detrimental to patient care, ultimately undermining the purpose of MedAI. Therefore, the accuracy and reliability of MedAI tools must be ensured, especially in cases where reliable research is scarce.

### 3.1.3 Motivation

Despite the increasing significance of research literature-based medical AI solutions in modern healthcare settings, there is a lack of comprehensive literature reviews analyzing the potential threat of predatory science on these solutions, and how it can impact AI usage in medical research and clinical practices. To fill this gap, we are motivated to study the collective threat of predatory research in trusted research literature repositories on medical AI decision-making. Our work aims to provide a better understanding of the impact of predatory science on research literature-based AI solutions and contribute to the development of robust and reliable MedAI tools.

### 3.1.4 Objectives of the survey

The aim of this survey is to conduct a comprehensive study of the impact of predatory research on AI-based medical solutions, with the goal of providing insights and guidance to developers, researchers, data-curators, and clinicians. While cost analysis is outside the scope of this survey, it's important to note that any misdiagnosis or incorrect treatment resulting from predatory research can have a significant financial burden for healthcare providers and patients alike. By identifying and addressing the potential threats of predatory research on AI-based medical solutions, we hope to improve the quality and integrity of healthcare and support more effective decision-making by healthcare professionals.

- First, we establish a background and identify published works that have investigated or evaluated the threat of predatory science undermining the credibility of genuine research in medicine [57, 54, 58, 56, 63, 64, 53, 65, 66, 55, 67, 68, 69, 70, 71, 72, 73, 74, 75].
- Second, we identify a wide range of existing clinical and non-clinical knowledge-extraction, decision-making solutions, and derived databases utilizing the most trustworthy NIH research literature repository PubMed [45, 46, 76, 47, 48, 49, 50, 51, 77, 78, 79, 80, 81, 52].
- Third, we provide a comprehensive study of the direct and indirect impact of predatory science on medicine practices [82, 83, 84, 85, 86, 60, 59, 87]. We study the criticality of potential security, integrity, and safety issues induced by predatory science in clinical and non-clinical MedAI solutions [28, 22, 1, 2, 61, 25, 23, 27, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 68, 99, 100, 101, 102, 103].

### 3.1.5 Contributions of the survey

The results of this survey provide valuable insights for future research in this area, including the development of defensive measures in the event that these attacks prove to be a significant practical concern. This survey is particularly useful for researchers and medical practitioners who are interested in understanding the direct and indirect impact of predatory research on healthcare decisions. By highlighting the potential threats of predatory research on AI-based medical solutions, we hope to inspire further research and development of more robust and reliable AI tools in healthcare, ultimately improving patient outcomes and the quality of care provided by healthcare professionals. To summarize, the key contributions of this survey are:

- Identifying the rising threat of the predatory research and data poisoning in research repositories needs immediate attention of research and medical community to develop practical strategies to minimize the inclusion of predatory research in reputed research repositories.
- Presenting that NIH research literature repository PubMed and PubMed-derived databases as inputs to the state-of-the-art MedAI solutions, will allow researchers, developers, and research repository administration to find mitigating strategies.
- Identifying the threat of predatory publications impacting new-age medical AI solutions will allow the development of more secure and robust MedAI solutions at both the data and the algorithm levels.
- Mitigating the threat of data poisoning into trusted resources will encourage the practical adaptation of research-backed decision-making in clinical settings for improved healthcare.

### 3.2 Taxonomy and Related Work

We examine and compare a wide range of existing works based on their scope and relevance. Each of these categories presents the significance of the area and its relevance under the scope of our work. A deeper analysis further reveals that a common goal is addressing a particular problem of predatory research impacting medical AI integrity and security. Hence, we develop our taxonomy by structuring the related work and expanding our analysis in these categories.

1. *Predatory Research* can be classified as a combination or standalone variation of bogus, duplicate, manipulated, incorrect, and research frauds. If predatory research is mixed with genuine research, overall confidence in the research literature repositories is diminished.

2. *Research literature-based MedAI solutions* use information extracted from the medical research literature repositories like *PubMed*. *Medical AI (MedAI)* solution is the implementation of Artificial Intelligence (AI) in medicine to improve overall healthcare. Advanced technological innovations can gather, analyze, synthesize, and infer meaningful information to reduce the time, effort, and cost involved in complex healthcare solutions. MedAI solutions provide a comprehensive and current knowledge dataset for efficient healthcare decision-making.

3. *The impact of compromised MedAI solutions on medical research and practices* can be significant, particularly in precision medicine scenarios. This can lead to misleading findings and incorrect healthcare practices, with negative consequences for patient outcomes and public health that can last for extended periods of time.

We analyze the existing work about the threat of predatory research, AI in medicine, vulnerabilities of AI-based solutions, and impact of research and predatory research on medical practices and prospective medical research.

### 3.2.1 How do we collect the papers?

We used Google Scholar and PubMed as primary search for papers involving medical AI, threats to medical AI, medical AI solutions based on research literature, and the problem of predatory research. In addition, we also utilized the iris.ai search tool to find a pool of related research papers [51]. We employed Google search to find out the major medical AI milestones, research frauds, the impact of predatory research on public health, and recent work and news regarding medical AI and trends in predatory research. The primary keywords and phrases are predatory research, medical fraud, medical AI, threats to medical AI, and research literature-based medical information extraction tools.

### 3.2.2 Related Work

On the predatory research, a review from Dinis-Oliveira discusses the main characteristics of predatory journals and the impact of predatory research in the context of forensic and legal medicine research and advocates the critical need for education on the threat of predatory research [73]. Mills et al. found that more than half of studied predatory publications are from the nature/biomedical field. Their work focuses on the aspects of shaping publishing motives, decisions, and experiences in predatory publishing [74]. Mertkan et al. found that the highest percentage of their studied predatory journals belong to Medicine (around 39%) [75].

A survey by Ji et al. on Knowledge Graphs (KGs) discusses Graph convolutional networks (GCNs), adversarial training, reinforcement learning, deep residual learning, and transfer learning, and how MYCIN-like systems apply knowledge-based decisions in medicine [103]. The NIH Translational project focuses on providing unified, standard KGs to accelerate knowledge-based medical decision-making tools. Therefore, KGs are expected to play a critical role in new-age MedAI solutions [104, 52]. Secinaro at el. discuss the increasing role of AI in Predictive Medicine, clinical decision-making, patient data, and diagnostics, and making a difference using AI in healthcare management [101]. Srinivasu et al. emphasize the importance of explainable AI in the medical decision-support paradigm to be robust and precise as it deals with human survival [100].

Alshehri and Muhammad discuss the literature in the field of IoMT, AI, medical signals use and fusion, edge and cloud computing, privacy, and security in the smart healthcare domain. Their work covers sensors' interoperability, device and information management barriers, and using AI efficiently. However, building more robust solutions against privacy and security attacks is challenging yet essential [99]. The survey by Zhang et al. covers attacks and defenses on textual deep-learning, presenting how manipulated input data can alter the AI output, exploiting training and influencing output with or without the knowledge of the Model [28]. Their work has close relevance to information-extracting from the research literature as a key pre-processing step before feeding medical information to MedAI.

#### *3.2.2.1 What is the difference between this survey and the former ones?*

There is no broad and precise literature review systematizing all those security and integrity aspects involving predatory research effects on MedAI solutions. Our work investigates the security, integrity, and credibility of MedAI solutions that rely upon research literature repositories as critical data sources.

Table 3.1: Comparison of Recent Relevant Work in Medical AI, Threats, Tools, and Predatory Research

| **Reference** | **MedAI and AI threats** | **MedAI Tools** | **Predatory Research in Medicine** | **Predatory Research affecting MedAI Tools** |
|---|---|---|---|---|
| Alshehri and Muhammad [99], 2020, Survey | ✓ | ✗ | ✗ | ✗ |
| Zhang et al. [28], 2020, Survey | ✓ | ✗ | ✗ | ✗ |
| Dinis-Oliveira [73], 2021, Article | ✗ | ✗ | ✓ | ✗ |
| Ji et al. [103], 2021, Survey | ✗ | ✓ | ✗ | ✗ |
| Mills et al. [74], 2021, Review | ✗ | ✗ | ✓ | ✗ |
| Mertkan et al. [75], 2021, Systematic Review | ✗ | ✗ | ✓ | ✗ |
| Secinaro at el. [101], 2021, Literature Review | ✓ | ✗ | ✗ | ✗ |
| Srinivasu et al. [100], 2022, Case Studies | ✓ | ✓ | ✗ | ✗ |
| Our work, Survey | ✓ | ✓ | ✓ | ✓ |

More specifically, we consider the possibility that research literature repositories may contain predatory content. We also cover that predatory content is navigating from research repositories to MedAI inputs and output, compromising the integrity of these solutions. Table 3.1 shows a brief comparison of the scope of our work with other recent relevant work. We identify a gap in covering all aspects of predatory research, information extraction tools, applied AI methods in healthcare, and associated threats. Identifying non-obvious threats, vulnerabilities, and misuses is essential in designing better defense strategies to protect the integrity and security of current and future MedAI solutions.

## 3.3 Background

The idea of developing machines that can replicate human intelligence, known as Artificial Intelligence (AI), was first introduced in 1956. While AI has the potential to transform medicine, its adoption in day-to-day clinical practices has been cautious due to the complexity of integrating technology with medical, financial, legal, and ethical considerations [105]. [105]. Despite enormous challenges, the rewards of using Medical AI (MedAI) are undeniable, and there has been increasing use of MedAI in robotic procedures, diagnosis, statistics, and human biology, including *omics* [106]. MedAI has opened a new dimension for medicine to harvest the abundance of knowledge scattered in the medical research literature and isolated silos of diseases, drugs, and patient data. Computational advancements have enabled the research-data-AI trio to overcome limitations in precision medicine [107]. However, the cost of unreliable MedAI output is high due to vulnerabilities and threats associated with research misconduct, data flaws, and exploitable algorithms. This can have fatal consequences for patient care and lead to harmful directions in future research and healthcare solutions. By addressing these risks and ensuring the accuracy of MedAI solutions, we can unlock the full potential of precision medicine and improve patient outcomes.

### 3.3.1 MedAI Methodologies and Biomedical Knowledge Representation

#### *3.3.1.1 MedAI Methodologies*

Bayesian methodologies are basic standards for acceptable uncertainty in research data and are widely accepted in the medical research community, and regulatory agencies [44]. Semi-supervised and unsupervised machine-learning techniques are more applicable to developing transformative machine intelligence-based systems for diagnosing and recommending treatments for a range of diseases and health conditions [43]. Artificial Neural networks and fuzzy logic can be combined as a hybrid intelligent system to accommodate common sense, extract knowledge from raw data, and use human-like reasoning mechanisms [8]. A PubMed-based study shows that more than 70% of AI methods applied in medical research are Neural Networks and Support Vector Machines for imaging and genetics [108]. At the same time, NLP is a crucial method to extract information from unstructured data such as research literature, clinical notes, patient reports, etc. [108]. Deep Learning (DL) and Natural Language Processing (NLP) are widely employed to extract meaningful information from the research literature. The intersection of data science, analytics, and Precision Medicine optimizes the tools and information used to deliver improved patient outcomes [109].

#### *3.3.1.2 Biomedical Knowledge Representation*

*Syntactic Analysis* and *Semantic Analysis* are two core operations of NLP to get the structure and the intent of the given text. Semantic analysis is closest to understanding and interpreting information in the right context to mimic human intelligence. Semantic analysis is a multilayered process including filtering, segmenting, encoding, defining, and identifying relationships between objects, linguistic perception, syntactic analysis, pattern classification, data classification, feedback, cognitive reasoning, and data understanding.

The Unified Medical Language System (UMLS) provides semantic knowledge to extract unique associations among diseases, genes, and drugs. Each concept is defined in UMLS by a Concept Unique Identifier (CUI) [110]. Two concepts can be connected with one or more semantic relationships known as predicates. A triplet of object-concept, subject-concept, and relationship carries concise and useful information. For example, drug x (object-concept) treats (relationship-predicate) disease Y (subject-concept). Semantic analysis has been a key component of ML to extract relationships among diseases, diagnoses, and treatments from the research literature [111]. Resource Description Framework (RDF) is a standard language for representing information about resources in the World Wide Web. RDF provides a way to describe resources using a set of triples, which consist of a subject, a predicate, and an object. It is used in applications, such as data integration, knowledge representation, and semantic web technologies [112].

#### *3.3.1.3 Knowledge Graphs in Medical Domain*

Knowledge Graphs (KGs) contain a vast amount of prior knowledge and are widely used in decision-making systems, search engines, and recommendations [113]. In the medical domain and research, KGs are highly applicable, as they can help providers and researchers identify both known and unknown relationships among diagnostics, diseases, and treatments. With knowledge reasoning, logical inference, and probabilistic refinements, intelligent systems can suggest treatment options based on the data in KGs. Furthermore, knowledge reasoning can derive new relationships among the entities, further enriching the KGs and enabling more accurate decision-making [114]. Figure 3.1 shows a visual presentation of knowledge queried from SemMedDB for a drug *Imatinib* to show how it is associated with genes and diseases through different predicates. Nodes represent the concepts, and edges represent predicates. RTX-KG2, a recently developed *Knowledge Provider*, is a biomedical knowledge graph to integrate 70 knowledge sources following the

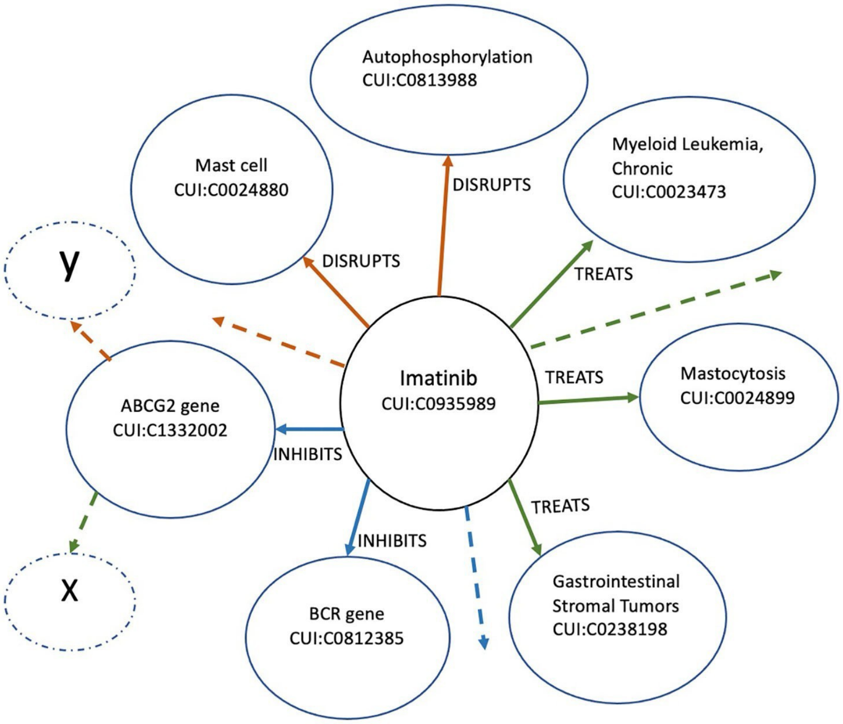


Figure 3.1: SemMedDB Extracted Knowledge Graph- Nodes (Concepts with unique UMLS CUIs); Edges (Predicates)

standard *Biolink* model to maximize interoperability [52, 104]. Biomedical data standardization can provide standardized data to a diverse set of medical applications, reducing the cost of cross-verification and synthesizing among heterogeneous data sources. However, it's even more critical to ensure the integrity of the data, as inaccuracies or inconsistencies can have significant consequences for patient care and research.

### 3.3.2 Security and Integrity of MedAI Solutions

Since 2010, about 40% of 200 new businesses have directed health interventions or predictive capabilities. It is estimated to help with a reduced healthcare spending of around $450 billion if scaled up to mainstream use [115]. However, it also increases the risk of utilizing promising AI methodologies which can be exploited to alter inputs and output through exploratory and adversarial attacks. If the manipulated research publications get included in a reputed research literature repository, this predatory research poses a potential threat to the data integrity of the data source.

However, MedAI's goal of considering all relevant information and filtering out irrelevant information without skipping challenging instances may make it vulnerable to predatory research. In precision medicine, where a single paper on the latest finding can potentially alter a patient's life, it's crucial to validate the accuracy of such papers and ensure they are not predatory.

Data pollution through targeted or untargeted data poisoning can mislead the MedAI algorithm [22]. Even without exploiting the MedAI algorithm, undetected bad inputs can traverse to the MedAI output. Thus, it is important to identify the threat surface and find solutions to mitigate the threat to the minimum possible level. Misclassification in neural networks is a well-known adversarial attack, which is more common to image-based MedAI solutions. However, text-based adversarial attacks are feasible and can be damaging the confidence in MedAI output [24, 22, 2]. Szegedy et al. showed that adversarial inputs need not to appear unusual or pathological, and even slight perturbation can change the outcome completely [116]. An attacker can potentially perturb it in a direction that aligns well with the weights of the MedAI algorithm and thus amplifies its effect on the output [25]. The potential for algorithmic bias may violate beneficence and non-malfeasance medical principles affecting a specific population through incorrect or absent diagnosis and treatment [117]. As NLP is a core process to extract intelligent information from the research literature for MedAI, it is critical to have a defense against adversarial attacks [28].

This work focuses on the threats to MedAI solutions that use research literature as a primary data source, with a particular emphasis on the infiltration of predatory publications that can impact the integrity and security of MedAI solutions. Although AI has the potential to bring significant benefits to healthcare, it can also have profound health effects due to data bias, insufficient sample size, and incorrect results. In human-dependent settings such as therapy sessions, AI may need to be more trustworthy before it can completely replace humans. While AI can reduce costs and improve access to healthcare in rural areas, patients' sensitive information stored in health-related digital records can pose a significant risk to confidentiality. Any manipulation or induced biases in AI can also be a severe problem, potentially harming the patient. By addressing these potential risks and ensuring the accuracy and reliability of MedAI solutions, we can help unlock the full potential of AI in healthcare and improve patient outcomes.

## 3.4 Predatory Science

Medical research has been revolutionary in the past few years, and there is an apparent increase in the number of publications each year. However, innovation is not the only reason for soaring numbers. *Publish or Perish* culture and Open Access (OA) journals are great contributors to an unprecedented increase in research publications. The pressure to publish can lead researchers to bypass rigorous review processes, resulting in non-standard and questionable publishing methods that promote predatory publications with unintended or intended data-poisoning attacks. Research misconduct cases further increase the probability of data manipulation. In this work, we closely examine the publishing methods and motivations behind the rapid increase of predatory publications. Using research literature available on Google Scholar and PubMed, we analyze the definition, motivation, actors, and impact of predatory research. By shedding light on the potential threats of predatory research, we aim to guide future research with trusted integrity of scientific publications. Before the dawn of OA journals around 2000, medicine research publications were under traditional journals where the reader pays for access [118]. The traditional model supports a highly ethical, comprehensive, and robust peer-review system to publish trustworthy and valuable research. Over time, many found the traditional approach too strenuous and limiting to disseminate the research timely, almost delaying the actual benefit and impact of the research findings. A Multi-layered, tedious, mostly yearlong, and complex peer-reviewed process and pay-to-access model pushed researchers to look for alternatives.

Without a robust review process, there is a probability of unauthentic research getting published with flawed findings and conclusions. Such predatory publications can pollute the research literature repository. Figure 3.2 shows a high-level overview of predatory science impacting MedAI solutions and healthcare. Polluted data input can cause an intentional or inadvertent failure of real-world MedAI systems. OA journals may assist researchers to represent their work quickly, with much shorter publication time and free online access to the research community. The APC (Article Processing Charge) may sound necessary to maintain the OA journal’s operations, but that pushed

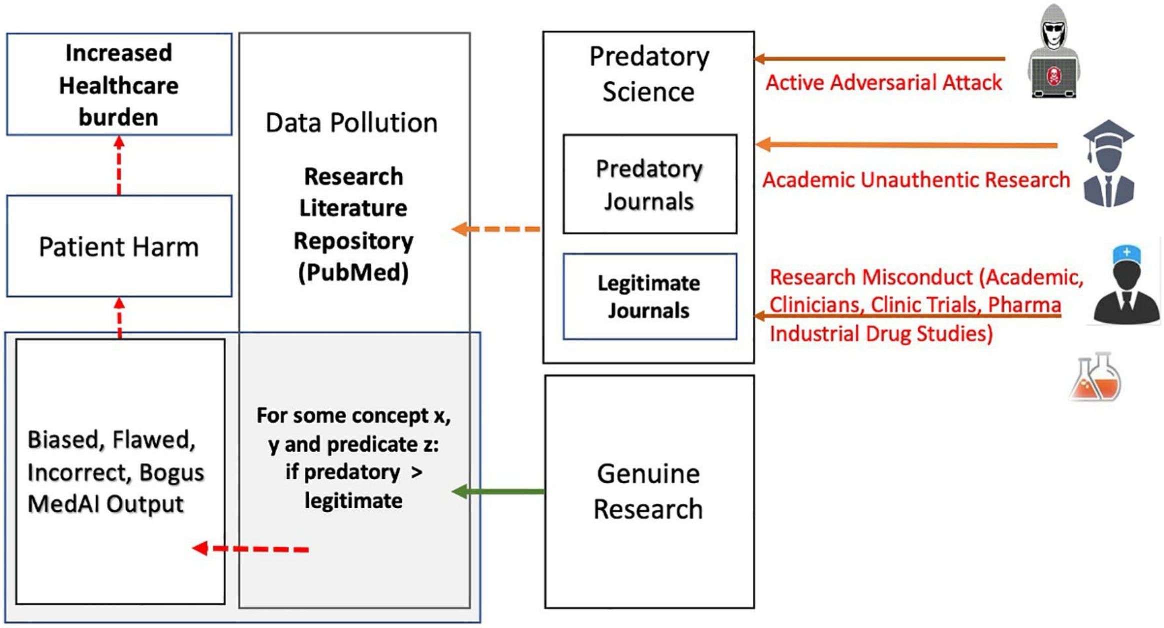

Increased Healthcare burden
Patient Harm
Biased, Flawed, Incorrect, Bogus MedAI Output
Data Pollution
Research Literature Repository (PubMed)
For some concept x, y and predicate z: if predatory > legitimate
Predatory Science
Predatory Journals
Legitimate Journals
Genuine Research
Active Adversarial Attack
Academic Unauthentic Research
Research Misconduct (Academic, Clinicians, Clinic Trials, Pharma Industrial Drug Studies)

Figure 3.2: High-Level Overview of Predatory Science impacting MedAI and Healthcare

to prioritize *pay-to-publish* over scientific integrity [119]. The OA model is here to stay based on the trend over the last decade with the known issues, so the focus has to shift more to what can be done to keep the integrity rather than discrediting OA journals altogether. As per DOAJ there are 15,954 journals from 124 countries in 80 languages [120]. While there is ongoing debate regarding the pros and cons of different publication systems, it's crucial to maintain the scientific integrity of all research publications. The consequences of polluted data can be devastating to science and medicine, making it essential to ensure the accuracy and reliability of research findings.

Research misconduct is another critical component of predatory research. As per the US Office of Research Integrity, research misconduct is, (a) *Fabrication is making up data or results and recording or reporting them; (b) Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record; (c) Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit; and (d) Research misconduct does not include honest error or differences of opinion* [121]. Research misconduct is a less reported, largely undetected problem with a possible long-term impact because of misleading directions based on fraudulent research [95].

Medicine journals are 40% of studied 171 journals, with the highest percentage of one or more predatory publications [75]. Reputable journals are also affected by research misconduct during the COVID-19 pandemic due to panic of finding the solution through rapid research [73, 122]. As we study the scholars addressing the problem of predatory science, including factors, actors, and defense strategies, we see the following trends.

- *What is predatory science?* Predatory journals, publishers with questionable practices, and publications involved with research misconduct [57, 58, 63, 53, 65].

- *What is Predatory Journal?* Mostly OA journals of Low quality, amateurish, and unethical [64] with absent or minimal peer-review [57, 58, 63, 65, 67] , pay-to-publish model with questionable editorial practices [58, 64, 66, 67]. Table 3.2 is a summary of key features of predatory journals and publications, which shows that plagiarism, bias, errors, and fraud are the most discussed issues in studied literature.

- *What are the reasons and motives?* The most common factors are the importance of quantity over quality [67, 71], a critical component for hiring, promotions, funding, and recognition [57, 58, 64, 65, 67, 70], publish or perish [54, 71]. Financial gain, career progression, and finding newness at all costs are primary motives for researchers, clinicians, and pharma industries for predatory research.

- *Who are the actors?* Global issues include world leaders in research and developing countries, Greedy publishers for self-interest, novice researchers because of pressure to produce, desperation after multiple rejects, ignorance, and lack of supervision, experienced researchers for financial gain and fame, or under pressure to produce.

- *What is the damage?* All are agreed that undermining scientific integrity is the most critical danger of allowing predatory science getting mixed with genuine research. A wastage of resources, money, and workforce are other major concerns [65, 66, 55, 71]. Unverified conclusions may harm patients if physicians are unaware of possible pollution [58].

- *What is the extent to harm patients?* Patient harm is not discussed in all papers studied with a focus on predatory journals but is more discussed in retracted research and research misconduct papers. The retracted research may already have done the damage for the period it went undetected, and even after retraction, it may affect physicians and people's perception for a much more extended period [67, 68, 70].
- *What is the defense strategy?* Defining the standard measurable features [58, 63, 53, 55] Awareness and education on identifying predatory journals and publishers [58, 63, 71], data quality checks [70].

Table 3.2: A Summary of Features of Predatory Journals/ Publications

| **Predatory Journals/ Publications Key Features** | **Publications References** |
|---|---|
| Plagiarism, bias, errors and frauds | [58, 64, 55, 68, 69, 70, 71] |
| Absent or minimal peer review process | [57, 58, 64, 66, 67] |
| Distorted editorial and publications practices | [63, 66, 55, 67] |
| Piracy | [58, 64, 71] |
| Pay-to-publish, Drain Money | [58, 63, 67] |
| Aggressive indiscriminate solicitation | [58, 55] |
| Abuse of trust, Concealed conflict of interests | [58, 72] |
| Low quality writing and images | [63, 67] |
| More predatory publications once indexed in reputed repositories | [58, 64] |
| Other: Human activity, Non-relevant scope of interest | [63, 72] |

There is some indication of predatory science presence in reputed databases and how it encourages to publish more predatory publications once a predatory journal is part of such reputed databases [65]. Citations of such unreliable research in reputed journal papers is another concern as there were 389 citations made in the WoS listed journals from 3427 potential predatory papers published between 2010 and 2015 [123]. A more recent work highlights that, in general, current

Table 3.3: Major Impacts of Predatory Science

| **Key Impact of Predatory Science** | **Publications References** |
|---|---|
| Damaging Scientific Integrity | [57, 58, 63, 64, 66, 55, 67, 68, 71, 72] |
| Impact patient care and corrode public health | [58, 63, 55, 67, 68, 70] |
| Pollution in reputed databases | [64, 65] |
| Waste of money, manpower, and resources | [65, 55] |
| Reflecting inadequacies in self-regulation | [71] |

and future African Neurosurgery physicians are unaware of predatory journals and not equipped to identify those [124]. Table 3.3 highlights the significance of predatory science undermining scientific integrity, as mentioned by almost all of the studied literature work.

From the above-discussed literature, there is no standard definition established so far acceptable to the global research community. It is challenging to avoid predatory journals without knowing what can be or cannot be predatory. Most of these journals do not conduct proper peer-review processes and follow questionable practices, including charging a substantial publication fee known as Article Processing Charge (APC). In addition to reviewed papers, we explored a few major predatory journal sites. They charge APC between \$300 and \$3500 with a processing time of 9 days to a couple of weeks, which is much shorter than traditional journals. Without a proper review system, unverified research does not have much credibility. As this kind of published work may have plagiarized, incorrect, unverified, fake data and manipulated results, *Predatory Journals* are increasingly interfering with genuine research [125]. Retracted research and undetected publications with research misconduct make the pollution probability higher and a more significant threat to scientific integrity [126, 127]. By 2015, there were estimated as many as 10,000 predatory journals worldwide. The ultimate risk is the altered results of synthesized knowledge because of rapidly increasing numbers of such predatory publications [53, 57, 55]. We observe that many research publications point out that predatory research can impact patient care and can corrode public health [58, 63, 55, 67, 68, 70]. However, there is no specific work to measure the actual patient harm, caused by predatory research or medical AI. A cost-analysis may provide insights into the net benefits or losses of research investments by NIH [43].

In 2013, John Bohannon's investigative fake medical paper was submitted to many publishers and got accepted by 60% journals including Elsevier [125]. A 2015 "Dr Fraud" experiment exposed the untrustworthy process of hiring editors and reviewers for predatory journals as a fictitious scientist was offered the position by such 40 journals and by 8 Directory of Open Access Journals (DOAJ) [86]. The cancer journal, Tumor Biology suffered retraction of 107 papers following the exposure of a fake peer review process [126]. More recently, 15 papers from Tumor Biology were retracted in 2021 for problems related to image manipulation or misuse [127]. A long-standing issue is becoming even more significant, undermining scientific integrity, and it needs immediate attention from all involved in genuine research efforts [66]. In general, academic institutes have some guidelines to avoid predatory journals, but that did not slow down the growth of predatory journals and predatory publications. There are more suspected predatory journals (10,406) than legitimate journals (10,077) in Cabell's list [55], which indicates the genuine concern of predatory literature polluting research literature repositories.

### 3.4.1 Predatory Science Infiltration in Trusted Resources

There are a couple of outstanding academic databases for biomedical research, medicine, and healthcare to provide credited references. The question is how much these credited resources are already infected by predatory publications. The concern is real as predatory journals are already becoming part of PubMed [64, 65], which is serving as a base source to develop other intermediate resources like NIH SemMedDB and Translational KGs to feed MedAI solutions. More specifically, Fig. 4.1 demonstrates the threat of possible patient harm and increased overall healthcare burden defeating the basic purpose of utilizing AI in medicine. Fig. 4.1 shows that predatory research has the potential to impact MedAI solutions and patient care, as well as influence future research, creating a cycle of increased data pollution in research literature repositories and diminishing confidence in MedAI output. Undetected flawed or bogus research conclusions can have a lasting impact on clinical practices, highlighting the importance of maintaining the integrity and security of research data used as the basis for MedAI in Precision Medicine, and not to let it be Predatory Medicine.

## 3.5 Research Literature based Biomedical Tools

Most biomedical research is conducted and documented in natural language, adding ambiguity, context, synonyms, and variants to the recorded information. Exponentially growing biomedical information, adding over a million publications every year to PubMed, is challenging for manual curation. MedAI can tap the potential of research information in data-centric precision medicine. However, it is necessary to develop methods and tools to comprehend vast biomedical text and extract knowledge in machine-readable form to process and present synthesized and inferred information. PubMed has been a reference point for biomedical research literature. Hence, an apparent primary source for many tools to address the automated curation of biomedical literature [128].

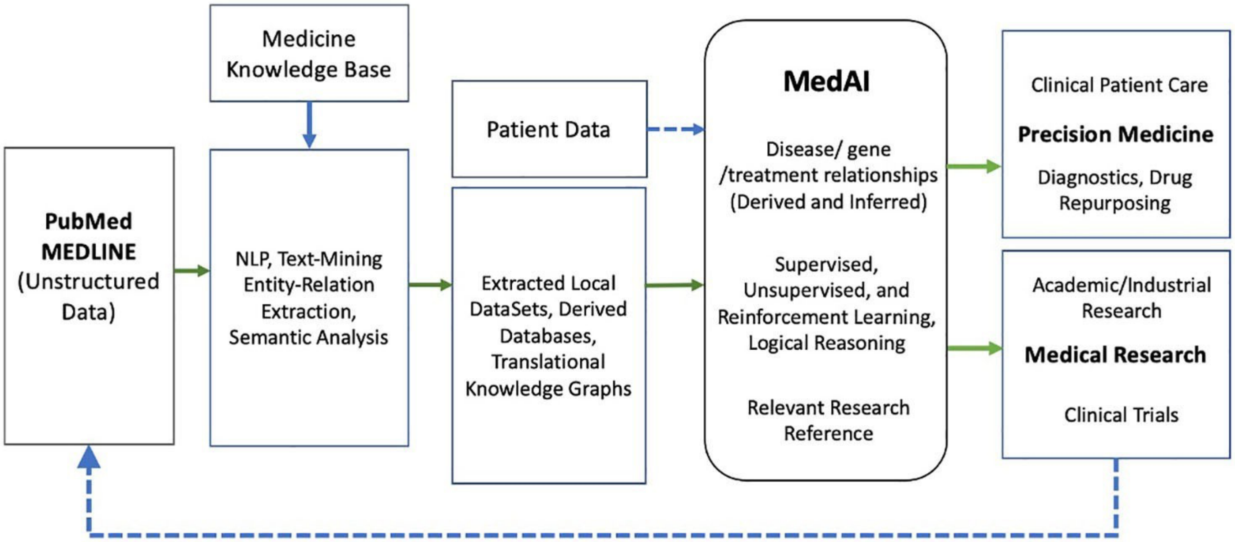

Medicine Knowledge Base
PubMed MEDLINE (Unstructured Data)
NLP, Text-Mining Entity-Relation Extraction, Semantic Analysis
Patient Data
Extracted Local DataSets, Derived Databases, Translational Knowledge Graphs
MedAI
Disease/ gene /treatment relationships (Derived and Inferred)
Supervised, Unsupervised, and Reinforcement Learning, Logical Reasoning
Relevant Research Reference
Clinical Patient Care
Precision Medicine
Diagnostics, Drug Repurposing
Academic/Industrial Research
Medical Research
Clinical Trials

Figure 3.3: Information Flow Among Research Literature, Medicine Knowledge Base, and MedAI

Figure 3.3 demonstrates the basic flow of information extraction from the research repository and how it can impact the MedAI, clinical implementation, patient care, and future research. Text mining through NLP is one core process in extracting knowledge that varies in applied methods, datasets, and User Interface dependent on the tool's goals. Automated curation utilizes various intermediate tools and datasets. Different components add functionality and cross-verification;

it also increases the threat surface to protect and maintain the integrity of these solutions. We searched Google Scholar and PubMed for *biomedical literature-based medical decision-making*, with variations of *medical AI, medical tools, medical solutions, PubMed information extraction tools, and PubMed-based medical AI solutions*. We searched for biomedical research literature-based tools developed in the last 15 years to see how these solutions use the diverse technological developments to apply for information extraction, analysis, and knowledge presentation. We focus more on the adopted AI methods, data sources, scope of usage, and limitations with the current possibilities in the field to expand on. Table 3.4 represents a quick view of methods, data sources, components utilized, and accessibility of studied tools.

SEMANTIC MEDLINE [129] is a web-based application that integrates document retrieval, advanced NLP, automatic summarization, and visualization. Resource Description Framework (RDF) graphs are used in network-based bioinformatics analysis to 1) prioritize the candidate disease genes, 2) propose novel drug targets, 3) discover enriched biological functions/processes in disease-related genes, and 4) identify potential disease relationships within the context of the whole knowledge.

D2R Server is a tool for publishing relational databases on the Semantic Web. It enables RDF and HTML browsers to navigate the content of the database and allows querying the database using the SPARQL semantic query language. Semantic MEDLINE web app shows the relationship between concepts and predicates, including references to the PubMed publication.

PICO provides ubiquitous access to clinical information and knowledge-based resources to answer clinical questions for evidence-based medicine. The PICO representation mainly relies on inherent semantic relationships between concepts to connect different elements. For example, with etiology questions, the connection between interventions and problems is assumed to be causal. Thus, the PICO frame is ill-suited to questions that challenge these implicit relations [47, 130].

GeneView is a fast and powerful tool for navigating biomedical literature for keeping pace with the latest research results. The most important features of GeneView include the possibility to search for articles describing a specific biological entity, flexible ranking of results according

Table 3.4: Comparison of Clinical and Non-clinical Research Literature based MedAI Solutions

| Tool | Year | Methods | Scope of Usage | Intermediate Source, Process or Supporting Tool | Availability |
|---|---|---|---|---|---|
| PICO [47] | 2007 | Text-mining, NLP, ML | Researchers, Clinicians | EBM-Evidence Based Medicine, UMLS, MetaMap, SemRep | Public |
| Semantic MEDLINE [129] | 2011 | Text mining, NLP, Semantic Analysis | Researchers, Clinicians | D2R Server, SPARQL, UMLS, SemRep, Resource Description Framework (RDF) graphs, SemMedDB | Free license for UMLS |
| GeneView [76] | 2012 | ML, NLP, text mining, relation extraction | Researchers | ChemIDPlus, NCBI Taxonomy, DrugBank/PharmGKB, Entrez Gene, Kegg, MeSH, dbSNP, Brno nomenclature, OMIM | Public |
| tmVar [77] | 2013 | CRF based ML, Entity recognition, tokenization, mutation identification (CRF) and regular expression patterns | Researchers, data curators | Tokenization, mutation identification, post-processing- regular expression for matching irregular and rare mention | Public |
| PubTator [79] | 2013 | Text mining, Entity recognition, Dictionary lookup, Annotation, ML | Researchers, data curators | tmVar, GeneTuKit, GenNorm, SR4GN, Dnorm | Public |
| Literome [49] | 2014 | NLP, Text mining, ML | Researchers, data curators | MS SPLAT-Statistical Parsing and Linguistic Analysis Toolkit -Taggers and parsers, sentiment analysis | Public |
| tmVar 2.0 [78] | 2017 | Text-mining, Pattern matching, dictionary lookup | Researchers | GNormPlus, dbSNP, clinvar | Public |
| LitVar [48] | 2018 | ML, Text-mining, Entity recognition | Researchers | tmVar,PubTator, TaggerOne, GNormPlus, SR4GN | Public |
| Iris.ai [51] | 2018 | Classification, word-usage frequencies, ranking algorithm on relevance | Researchers | Document Grouping, Core, Arxiv | Free Basic, Paid-Commercial |
| PubTerm [80] | 2019 | Co-occurrence and statistics of occurrences, annotation | Researchers, data curators | PubTator, DataTables, | Public |
| CancerMine [45] | 2019 | NLP, ML, text mining-genotype-phenotype relationship, word frequencies, semantic features | Researchers, Clinicians | Kindred relation classifier, PubTator, GNormPlus, tmVar, Dnorm | Public |
| mediKanren [46] | 2020 | Logical Reasoning, heuristics, indexing | Researchers, Clinicians | miniKanren, Racket, Knowledge graphs, SemMedDB, Precision Medicine | Public- on Github, proof of concept |
| LitSuggest [81] | 2021 | NLP, ML, Ridge classifier, elastic net classifier, Logistic Regression Classifier | Researchers, Curators | Pubmed, PubTator | Public, NIH Web-based tool |
| RTX-KG2 [52] | 2022 | Extract-Transform-Load approach, Biolink [104] Schema | Researchers, Clinicians | UMLS, SemMedDB, ChEMBL, DrugBank, Reactome, SMPDB, and 64 other knowledge sources | Public, code on Github, API |

to the users' needs using optimized ranking algorithms, and intuitive visualization of semantically annotated texts. GeneView can considerably reduce the necessary effort for searching, reading, understanding, and annotating biomedical articles [76].

tmVar is a text-mining tool based on a conditional random field for extracting a wide range of sequence variants described at protein, DNA, and RNA levels. tmVar 2.0 implements the future direction from original tmVar research to improve the clinical relevance of dbSNP reference variants (RS) by text-mining PubMed [77, 78]. PubTator is a web-based text mining tool for assisting biocuration, providing automatic annotations of biomedical concepts such as genes and mutations in PubMed abstracts and PMC full-text articles [79]. PubTerm is a simple system to acquire, curate, annotate, and categorize not only abstracts but also genes, diseases, species, drugs, sequence variants, journal, and author-related information [80].

The Iris.ai tool suite is explicitly aimed at researchers in the early phase of a new project [51]. They are especially suitable for interdisciplinary projects where the combination of knowledge from various research fields will be vital to the project's success. Iris.ai searches from a paper of the user's choice or a self-written problem statement. Iris.ai search is based on machine-extracted keywords, contextual synonyms, and hypernyms against more than 200 million Open Access papers, patents, and even EU-funded research projects.

Literome provides a cloud-based knowledge base for genomic medicine, featuring knowledge automatically curated from PubMed abstracts by an NLP system. It offers powerful search and exploration capabilities and a feedback mechanism to improve annotation and extraction continuously. Literome focuses on entities and relations most pertinent to genomic medicine. Users can browse and search the resulting knowledge base through the Literome website, which gets updated as new abstracts become available [49].

Litvar is a novel web-based tool that combines robust and advanced text mining with data integrating from PubMed, dbSNP, and ClinVar for an accurate search of variants and related gene, disease, and drug information. In addition, variants are often asserted with different names in publications; thus, a search in PubMed using only one name usually cannot retrieve all relevant articles [48]. LitVar uses tmVar, a high-performance variant name disambiguation engine, to normalize different forms of the same variant into a unique and standardized name so that all matching articles can be returned regardless of the use of a specific variant in the query. LitVar leverages the state-of-the-art literature annotation tool, PubTator, to provide critical biological relations among variations, drugs, genes, and diseases. LitVar supports searches by variant or variant with a gene found in the title, abstract, and full texts, including supplementary materials. For relation extraction, LitVar currently relies on sentence co-occurrence, and results may include false positives.

CancerMine is an automated approach using text-mining of the database of drivers, oncogenes, and tumor suppressors in different types of cancer [45]. CancerMine is using an ML approach logistic regression classifier on word frequencies and semantic features. A known relationship between genes and their role in a certain kind of cancer can help with early diagnosis and timely treatment. CancerMine can reduce manual effort and save time and cost to extract this information from the research literature.

mediKanren is a MedAI employing reasoning over the NIH SemMedDB knowledge base, using logical reasoning, heuristics, and indexing [46, 131]. mediKanren is a combination of miniKanren (Logic Programming Language), Racket (General-purpose Programming Language), a database SemMedDB, Knowledge Graphs, and a graphical user interface (GUI) to simplify data exploration for drug repurposing to assist Precision Medicine.

A more recent NIH tool, LitSuggest, can find and rank publications from research conducted in the Computational Biology Branch, NCBI/NLM, using advanced machine learning and information retrieval techniques. LitSuggest can automatically scan the literature weekly for new publications relevant to a user-specified topic [81].

RTX-KG2 is presented as the first open-source knowledge graph that integrates UMLS, SemMedDB, ChEMBL, DrugBank, SMPDB, and 65 additional knowledge sources within a knowledge graph that conforms to the Biolink standard for its semantic layer and schema at the intersections of these databases. The current version of RTX-KG2 contains 6.4M nodes and 39.3M edges with a hierarchy of 77 relationship types from Biolink [52, 104].

### 3.5.1 Comparative Analysis of Tools based on Biomedical Literature

We performed a manual analysis to compare Methods, Scope of Usage, Nature of Tool (As

Source, process, intermediate or supporting tool), and Availability. We compare these tools based on their objectives and scope, type of input data source and the data format, their limitations, and challenges, and if they consider predatory research in source data or during the information extraction/presentation process. From the review of these tools, it is evident that there is an undeniable need to efficiently curate the abundant knowledge from the research literature to enhance decision-making in research and clinical settings. It also clarifies that NLP and text-mining algorithms are basic yet critical components in data curation to make it machine-readable for any further utilization. It is interesting to observe that over the years, basic tools developed for information extraction [79, 78] have been used by more complex data curation with targeted approach [45, 48, 80]. Many of such tools utilize entity and relation extraction to assist automatic biocuration [76, 77, 78, 79, 48, 45]. Most of the tools use machine learning for data training and predictions, while mediKanren [46] uses logical reasoning, heuristics, and indexing. Supervised machine learning is the most commonly applied AI method using classifiers on keywords, co-occurrence, pattern matching, and dictionary lookup to extract and normalize entities and relationships. Few of these tools utilize similar intermediate systems like UMLS, GNormPlus, ClinVar, DNorm, and dbSNP for more efficient and comprehensive knowledge extraction from the bio-literature. GNormPlus is an end-to-end system that handles gene/protein name and identifier detection in biomedical literature, including gene/protein mentions, family names, and domain names. ClinVar aggregates information about genomic variation and its relationship to human health. dbSNP contains human single nucleotide variations, micro-satellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic. The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote the creation of more effective and interoperable biomedical information systems and services. A recent development with integrating datasets to develop more consistent inputs is prone to the same threat of predatory data-induced data pollution if any component dataset is already exposed to such a threat. For example, RTX-KG2 is based on 70 databases with SemMedDB as a primary source. The most significant contributing knowledge source for RTX-KG2.7.3pre is

SemMedDB, which has 19.3M edges, about one-third of the total edges. SemMedDB is PubMed derived database, and PubMed is vulnerable to predatory research [52, 64]. We compare these MedAI solutions, broadly on their objectives and scope, their data sources, and data formats. We also discuss the limitations and challenges of these tools that can affect the efficiency and integrity of these solutions.

#### *3.5.1.1 Objectives and Scope*

Some tools are more focused on providing ready support to clinicians based on derived or inferred relationships between diseases, genes, and drugs [47, 129, 45, 46] while others are primarily for data curation [77, 78, 79, 80]. The Iris.ai tool is explicitly aimed at researchers to provide relevant research literature based on research problem formulation, especially for interdisciplinary projects where the combination of knowledge from multiple research fields can be vital [51]. LitSuggest can help with the biomedical literature recommendations and curation [81]. These tools can be used independently, and more advanced systems utilize the work done through modular tools as subsystems.

#### *3.5.1.2 Input Data Sources and Data Formats*

PubMed as a primary data source was the key criterion to study these tools, but this is worth noting that these tools work with different data formats. For example, many of them extract information from PubMed directly and incorporate other curated domain-specific datasets for genetics, diseases, and drugs to present known and inferred knowledge. Semantic MEDLINE and mediKanren utilize PubMed-derived SemMedDB and Knowledge Graphs. Some other tools employ other databases in combination with PubMed data. Most of these tools are web-based, with regular updates to the data with growing biomedical literature. mediKanren is still evolving but currently can operate on a moderately advanced laptop with a local GUI interface with local datasets on disk to assist physicians in real-life clinic scenarios [131]. RTX-KG2 is using the largest set of KGs among current solutions and NIH translational data project is expected to be a common uniform standard dataset for future MedAI solutions [52, 104]. LitSuggests is directly extracting information from

PubMed, and it can have a direct impact on increasing predatory publications [81].

#### *3.5.1.3 Limitations and Challenges*

Though all tools acknowledge the significance of research literature and data curation, many of these tools face the challenges and limitations of data and algorithms used in the process. The sheer volume of the medical literature and the high cost of expert curation force current curated variant information in existing databases to be incomplete and out of date [78]. Many of the studied tools work with only titles and abstracts to extract the information. However, more information exists in full text, and these tools may have incomplete or missing information. These tools are also bound to the accuracy of the applied text mining algorithms, which are known to be imperfect in both entity recognition, and relation extraction [48]. As tools are primarily trained on abstracts for entity tagging, their full-text results may be inferior due to their structure and complexity. Another common limitation to using these tools is the domain knowledge, especially with PICO and Iris.ai; results are dependent on how well the problem is formed. Other tools' outputs may also be more valuable to the researchers and clinicians well-versed with background knowledge to know what to look for even if the information exists.

All these data-driven knowledge extraction tools heavily depend on the integrity of the research publications hosted by PubMed. Possible pollution in terms of bogus, invalid, manipulated research either through predatory journals or research misconduct cases may cripple the chain of trust, and results cannot be trusted. Any vulnerabilities in lower-level tools may impact the functioning of complex tools using these tools in the process.

This review of research literature-based tools clarifies that none of the studied literature on tools discusses the possibility of data pollution in PubMed or any other intermediate data sources. As there is no current consideration of input data pollution, that confirms the potential threat that predatory research mixed with genuine research in a research repository may impact the output of these tools. To build more robust MedAI tools, it is important to acknowledge the threat and then build the defense to mitigate the threat of predatory research-induced data pollution.

### 3.6 Threats and Attacks on AI-based Healthcare Solutions

On April 11, 2018, the first AI computer vision diagnostic system without a human clinician was approved by U.S. Food and Drug Administration (FDA) [5], and FDA approved 29 AI-based medical systems between 2016 and early 2020. There has been an immense interest in MedAI since 2010 with a 20-fold increase in AI/ML-based research publications from 2010- 2019 [132]. On the one hand, this acknowledged the significance of AI in future medicine; on the other hand, it raised the known concerns of adversarial examples even higher, especially in the absence of clear ethical and legal consequences when a machine fails.

We present a basic understanding of the most common threats and attacks which can impact MedAI solutions, and then we compare the studied works, based on their scope, AI method(s), attack type, and key impact. Studied papers are, either specific to healthcare or generalized but can be highly damaging in healthcare [22, 1, 92].

#### 3.6.1 Vulnerabilities of MedAI Solutions

Adversarial examples are known threats, especially in computer vision to alter the output with unsuspecting minor pixels manipulations [25], a similar approach is expandable in text-based AI algorithms and systems. This somewhat new threat has not been mapped yet to the depth of the possible destruction it may cause when applied to the medical research literature. Appropriately peer reviewing is essential in validating and approving medical diagnostics research, which may require public access to the architecture and methods to maintain transparency, but that provides an opportunity for more targeted adversarial attacks [22]. Medical staff may not have advanced knowledge to build secure systems, but internal medical staff may develop applications to keep usability simple with no inbuilt defense mechanism. It is vital to know how model deployment can enable end-users to submit data into a running ML algorithm to subtly influence its behavior without ever knowing or accessing the model or the hosting IT environment [22].

While image-based adversarial attacks are well discussed in the literature, now more studies are finding vulnerabilities to exploit in text-based adversarial attacks. Our understanding is that if the input is polluted but undetected, it is difficult to identify the possible attacks and the defense. Finlayson et al. showed how text manipulation can turn a high-risk observation into a low-risk event, which will hamper the needed attention to the patient [22]. Especially, in the case of opioid abuse, it can be life-threatening in the absence of due medication assistance. The other scenario presents a motivation to get the reimbursement approved, which can be achieved through the adversarial addition of similar yet different medical codes as those are interpreted differently by the AI [1].

It is difficult to define perturbations in symbolic and discrete research literature texts. If an attacker can map the discrete to continuous data, attacks from computer vision may be applicable to the text. The unperceivable textual adversaries are complex as even minor changes may be noticeable. Also, they can be tracked or detected through many languages reviewing tools through spelling-check or grammar-check. The semantics of a word or sentence may be changed in context and sentiment even by small text changes that are not preferred for designing adversarial attacks. Due to these differences, current state-of-the-art textual attackers either carefully adjust the known computer vision-based methods by enforcing additional constraints or propose novel methods using different techniques. Subtle, undetected, and unsuspected input pollution is necessary for the success of an adversarial attack [133].

Li et al. show how Deep Neural Network (DNN)-based text interpretation using NLP can be manipulated by evasive, minimal yet well-crafted adversarial examples to get completely different output from the sentiment analysis. The changes look like typos with less suspicion, changing only a digit or letter strategically. However, these changes replace a 100% negative result with a 89% positive one [28, 23].

### 3.6.2 Type of Attacks Impacting MedAI Solutions

An AI pipeline generally consists of data collection, data pre-processing, model training, model inference, and system integration [134].

Each phase of the AI pipeline can be vulnerable to various security threats. In this survey, we are focusing only on data collection and data preprocessing phases to study if it can impact the MedAI output. The major security risks in the data collection phase include data biases, manipulated data, and data breaches. In terms of research repositories, the data collection represents the published research articles and data presented in these publications. Every time a predatory publication is added to a reputed research repository, the data pollution may have a larger influence to impact overall conclusion based on data extracted from these research repositories. The data pre-processing phase covers any derived datasets from research repositories to use as MedAI input.

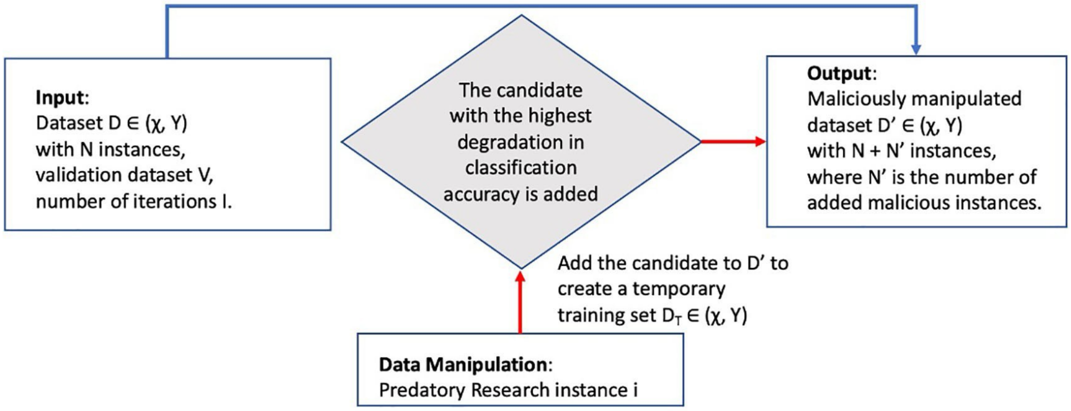


Figure 3.4: Algorithm-Independent Predatory Research Data Poisoning Attack - Adapted from Mozaffari et al. [2]

Fig. 3.4 is adapted from Algorithm 1 of Mozaffari et al. and simplified to demonstrate the threat of algorithm-independent attack through input data manipulation [2]. The original dataset D belongs to (X,Y) with N instances. Where X represents an instance's attributes and Y is the class label. The attack adds N' malicious instances to the original dataset to develop a manipulated dataset D' belongs to (X,Y) with N + N' instances. For each candidate, the algorithm evaluates its classification accuracy on the validation set. The candidate with the highest degradation in classification accuracy is selected and added to the dataset.

This generic and algorithm-independent attack can be applied to a wide range of medical datasets and AI algorithms. The robust and modified algorithm may defeat algorithm-specific attacks, but algorithm-independent attacks pose a larger threat. An attacker may not need to know any specifics of the applied algorithm. However, knowledge of the algorithm can increase the efficacy of such attacks [2]. NIH-developed translational databases are to provide extracted medical research information in a standard machine-readable format so that more applications can utilize available information without redoing the extraction process each time separately [104]. In this case, if information extracted from predatory publications is part of such intermediate datasets, it has an even larger scope to impact MedAI solutions.

Adversarial attacks can be categorized based on methodology, application, algorithm, and data. Major methodological categories of Attack Methods are Black-Box and White-Box. We briefly discuss the attack categories more specifically towards text-based as applicable to extract information from research literature and how predatory science can exploit these methods to impact research and clinical MedAI systems.

#### *3.6.2.1 Poisoning Attacks/ Training Phase Attacks/ Causative Attacks*

During the training, an adversary carefully manipulates the training data to compromise the learning process. An adversary can change the value of the input data to a certain threshold using data injection, modification, and logic corruption methods to manipulate the training data. Data injection pollutes the training data, while the modification is poisoning the data before the training, and logic corruption can temper the model itself. This way, an adversary can bias the overall learning process of the ML model applied to MedAI to misdiagnose the test data to mislead the suggested treatment, which may harm the patient [27].

#### *3.6.2.2 Evasion Attacks/ Testing Phase Attacks/ Exploratory Attacks*

In an evasion attack, the adversary tries to deceive the MedAI by enforcing adversarial samples during the testing phase. An adversary does not influence the training data but can access the ML model to obtain sufficient information.

As a result, it attacks the ML model and manipulates it to misclassify the patient status in a MedAI [88]. White-box attacks and black-box attacks are two major categories of evasion attacks.

#### *3.6.2.3 White-Box Attack*

White-Box attack requires access to the model's complete information, including architecture, parameters, loss functions, activation functions, input, and output data. White-box attacks typically approximate the worst-case attack for a particular model and input, incorporating a set of perturbations. *Logical attack* is a white-box attack corrupting the algorithm itself for generating malfunctioned output. White-box adversary strategy is often very effective as an attacker may know inputs, model, and algorithms. ML models used in different medical applications can be vulnerable to White-Box attacks impacting accuracy drop and attack success rate [27].

#### *3.6.2.4 Black-Box Attack*

Black-Box attack does not require the details of the algorithm, but it can access the input and output. This type of attack often relies on heuristics to generate adversarial examples. Concatenation, Edit, Paraphrased-based, GAN-based, substitution, and reprogramming adversaries can be applied to target characters, words, sentences, or APIs to impact different models [28]. It is more practical as in many real-world applications, the detail of the system is a black box to the attacker. Black-Box methods are further classified as MRC: machine reading comprehension; QA: question answering; VQA: visual question answering; DA: dialogue generation; TC: text classification; MT: machine translation; SA: sentiment analysis; NLI: natural language inference; TE: textual entailment; MD: malware detection. Each of these Black-Box attacks can impact MedAI input data, and information extracted in the process, and have the potential to influence the MedAI output.

Table 3.5: A Summary of Attack Types, Affected AI Methods, Applications and Key Impacts

| Work | Application | AI Method(s) | Attack Type | Key Impact |
|---|---|---|---|---|
| Mozaffari-Kermani et al. [2], 2015 | Healthcare, Biomedicine | Naive Bayes, Nearest-neighbor | Poisoning Attack | Distrust in MedAI, Patient distress. |
| Papernot et al. [89], 2018 | Clinical data, Network Intrusion, Autonomous vehicles | SVM, Random Forest Classifier | Exploratory attacks, injection, modification, input poisoning | Confidentiality, Integrity, and Availability |
| Biggio et al. [90], 2018 | Computer Vision and Cyber Security | ML, Deep Learning | Evasion attack, Poisoning Attack | Integrity, Availability, Privacy/ Confidentiality |
| Duddu [92], 2018 | Cyber Warfare | Supervised, Unsupervised, and Reinforcement Learning | Evasion attacks, Poisoning Attacks, Black Box attacks | integrity, availability, and privacy |
| Finlayson et al. [22], 2018 | Computer vision and medical imaging | Deep Learning, Neural Networks | Projected Gradient Descent (PGD), Black-box, white box | Misclassification by accurate classifiers |
| Sun et al. [88], 2018 | Health Informatics | Deep Learning, Predictive Modeling | Iterative attacks, optimization-based attacks, FGSM | Misclassification from alive to deceased |
| Finlayson et al. [1], 2019 | Medical Diagnostics and Decision Support | Deep Learning | Evasion attack, Poisoning Attack | Insurance Frauds, Biased drug trials |
| Ngiam et al. [93], 2019 | Big Data and ML in clinical healthcare | Deep Learning, Deep Neural Networks | Neural Networks, Supervised ML | Data privacy and security, Fear of replacing humans |

In an active adversarial attack scenario, predatory publications with a targeted approach can be produced automatically in bulk through new-age NLP text-generator tools like GPT-3 [135, 136]. MIT *SCIgen* project verifies a similar threat in the Computer Science field where users could generate the paper, and it got accepted in multiple conferences [137]. A targeted adversarial attack on a rare disease can make it look valid while injecting fake data and findings with alternative conclusions.

The attacker can approach multiple predatory research publishing venues to publish much work validating these new findings. Based on the current exploratory threat, this active attack is viable and capable of shifting the weight on the concept understood by the MedAI algorithm. Compromised data resulting from predatory publications can have significant consequences, particularly for rare and unknown diseases where data and information are limited, and cross-validation is challenging.

As we study the different attacks and vulnerabilities of MedAI solutions, we selected diverse research work yet relevant to the scope of our work. Table 3.5 provides a summary of the literature studied on attack types that can impact the integrity and security of AI-based solutions, especially, applicable in healthcare. Mozaffari-Kermani et al. and Sun et al. presented how different attack types can affect multiple AI algorithms causing distrust in MedAI solutions and patient distress [2, 88]. Finlayson et al. demonstrated that medical diagnostics and decision support using Deep Learning is vulnerable to poisoning, evasion, black-box, and white-box attacks causing insurance frauds and biased drug trials [22, 1]. Vulnerability to exploratory, injection and input poisoning attacks threatens the integrity, privacy, availability, and security of MedAI solutions using a range of AI algorithms and methods, including supervised, unsupervised, reinforcement learning, ML, and DL [89, 90, 92, 93].

The most common adversarial attack is on classification algorithms to mislead the system to misclassify by polluted inputs, corrupted predictive models, or exploiting the algorithm's vulnerabilities. Other primary application text-specific categories are Machine Translation, Machine Comprehension, Text Summarization, Text Entailment, POS Tagging, Relation Extraction, and Dialogue System. To adopt the most effective strategy for target data, model, and algorithm, the adversary can fine-tune the attacks by *Iterative attacks* or *Optimization-based attacks* by generating specific adversarial examples for focused attacks on predictive models [88]. In the case of a research literature-based MedAI system, even subtle pollution can be critical in the era of Precision Medicine with a very targeted approach. The risks of having any possibility of manipulated biased results can be devastating in actual patient care. The following section discusses some compelling cases highlighting predatory science's short and long-term impact on medical practices.

## 3.7 Impact on Research, Tools, and Medical Practices

It is worth looking at some high-profile frauds in medicine to get a sense of how technological advances and manipulation of such systems can impact medical practices and how critical it is to ensure the integrity of MedAI solutions and define liabilities.

### 3.7.1 Impact on Clinical Research and Research Repositories

A study shows that over a decade, 9,189 patients were treated in 180 retracted primary studies. Not only were these patients put at risk, but these studies influenced other secondary trials to recruit more patients [69]. The annually increasing number of predatory publications increases the probability of data pollution in trusted research repositories. It is also observed that once a predatory venue is part of the reputed research repository, it publishes more papers, which allows predatory research to slip through minimal or no review and get mixed with genuine research [64]. This trend puts a damaging distrust in research literature repositories, with not knowing how much fraudulent research literature exists from legitimate or predatory journals.

### 3.7.2 Impact on Advanced MedAI Solutions

We observe that various medical AI tools use PubMed as a primary data source directly or as an intermediate resource. When such a trusted source gets infiltrated with unreliable data, it is necessary to evaluate the impact of data pollution on MedAI functioning. Apparently, without having awareness and acknowledgment of predatory science, MedAI may not have any defense mechanism. In the absence of any defense against predatory science, the integrity of MedAI is questioned to provide reliable output.

### 3.7.3 Impact on Healthcare Decision-Making

It is interesting that the motivation in research frauds is irrespective of the attacking method like performing unnecessary dermatological surgeries for money, which is image-based but can also be applied as text-based [94]. For example, a research paper from a predatory journal claims that some rare skin cancer can be treated by the drug ‘d’. If that paper is part of a trusted repository like PubMed by crawling through PubMed Central [64], tagged by MEDLINE, and moved to an intermediate database like SemMedDB or translational database mapping and eventually being used as part of inputs to a MedAI system. If there is no cross-validation, it is a high probability that the system will pick the reference. Being desperate for any health aid, a patient may trust this new possible intervention to give it a try.

### 3.7.4 Impact on Clinical Care and Public Health

There are many known cases to promote a drug to cure some disease but not disclosing outliers in the trials or making false claims through fraudulent clinical trials [95, 96]. Anesthesiologist Scott Reuben and Family Medicine physician Anne Kirkman Campbell are few among many others to commit such frauds for money, fame, and position, causing direct harm to patients including loss of lives [59, 97]. Misleading conclusions can deprive the community of needed preventive care, which can have a long-term impact on public health. Multiple studies about certain vaccines causing disorders in children baffled the medical community, as well as the general public for more than a decade [87, 98].

## 3.8 MedAI Survey Discussion

In medicine, ground truth is often ambiguous, and patient care is still primarily provider-dependent and often subjective even after all the assisting tools, including MedAI solutions. In such trust-based settings, mistrust can damage the patient and provider relationship and affect patient care. We understand that utilizing AI to tame the scientific knowledge for Precision Medicine can work against it entirely if the medical and computer science community is not proactive in acknowledging and developing a robust defense against predatory science invasion in the research literature. Medical AI systems are vulnerable to adversarial attacks because of sensitive and complex healthcare data from heterogeneous medical and technical sources.

As discussed in this work, predatory science is on the rise, and that can make the infiltration even higher in PubMed-like repositories. Higher data pollution will pose a higher probability of impacting MedAI tools using the research literature. It is challenging to keep up with the marking or removal of predatory research from trusted repositories. With the other technological advancements like *Fog Computing, and Edge Computing*, there are newer open challenges to utilize the power of these solutions, specific to medical data handling in real-time [138]. However these solutions are still developing and in the future, an analysis of applications and security analysis would be relevant to identify solution-specific threats.

There is an apparent fear of relying more on machines and ignoring the value of the human component. However, even with all the advancements, MedAI solutions have a long way to establishing the needed trust and addressing financial, ethical, and legal aspects. On the positive side, the contribution of MedAI is getting well-acknowledged and providing an exciting opportunity for physicians, researchers, and computer scientists to work together, unwinding new possibilities in healthcare. It is essential to establish trust in MedAI, especially the research literature data source, as no algorithm can extract the trusted output if inputs are polluted and not trustworthy [61].

We plan to look at other security vulnerabilities to develop a better defense mechanism against data pollution in the era of a plethora of information. Studying how predatory research information is navigating through publication channels, MedAI solutions, and eventually altering the clinical decisions affecting patient care will be the focus of our future work. MedAI solutions need to take advantage of every available information, especially the open-access research and social media. However, handling such information platforms, the future trustworthy information extraction will have greater challenges and higher stakes. From a broader perspective, the survey indicates the possible failure of any state-of-the-art MedAI logic to deliver the intended output if the inputs are polluted and left unidentified.

### 3.9 MedAI Survey Conclusion

We address the research gap in identifying the impact of predatory science on MedAI solutions and patient care. Though it may be challenging to measure the extent of actual patient harm as a direct impact of predatory science, it is intended to demonstrate the possible damage undermining the trust in MedAI solutions and their practical adaptation in clinical care. A few key findings from this survey are as follows:

- In the absence of a defined strategy to avoid predatory or retracted research publications, numbers are on the rise, and so they serve to pollute MedAI inputs.

- No existing standard guidelines to mitigate the threat of predatory research for extracting information or preparing intermediate research databases, which are being utilized by a wide range of MedAI solutions.

- Most of the work discussing security and integrity talks more about the possibility of exploiting vulnerabilities to fool AI algorithms into misclassifying but none of the tools or attacking studies discuss the possibility of data pollution in research literature data sources.

- Medicine practices and trends are highly influenced by medical research, and undetected predatory research can harm public health for a prolonged period.

We summarize the contributions of our work as follows:

- **Contribution to the Theory:** We create greater awareness of the predatory research-induced data pollution in a trusted research repository. We believe that our work provides a good base reference to study the problem, its significance, and the potential threat.

- **Contribution to the Practice:** We explain how medical research influences medical practices, and how medical AI is necessary for modern medical practices. We present that if the core data input gets manipulated, it can impact the medical AI performance, medical AI adaptation, future research, and overall influence the medical practices.

- **Contribution to the community:** If the threat is not mitigated, targeted attacks can turn into severe threats putting patients' lives at risk. Once trust is diminished in research literature-based medical AI solutions, that will hinder the adaptation of these solutions in practical settings and thus, not able to provide the intended service to the community. More robust future MedAI can provide great assistance to modern medicine and the community.

From our analysis of the literature on the predatory science, medical AI tools based on the research repositories, and AI vulnerabilities, we can see how technological advances help medical research, curating knowledge, and then using this knowledge to advance further in medical decision-making. However, the rising threat of predatory research can pollute these trustworthy research literature repositories, and that can potentially impact all those medical AI solutions which are using PubMed as a critical and trusted data source. Our work raises awareness about missing defense against predatory research-induced data pollution, and we expect that our work will initiate the difficult discussion at a larger level and help to develop feasible defense strategies. We are confident that verifying the threats of any sort early in the process will only help develop a more robust MedAI solution for taking Precision Medicine to the next level in broader settings.

# 4. ASSESSING RISK OF PREDATORY RESEARCH TO MEDICAL AI SOLUTIONS*

Medical Artificial Intelligence (MedAI) for diagnosis, treatment options, and drug development represents the new age of healthcare. The security, integrity, and credibility of MedAI tools are paramount issues because human lives are at stake. MedAI solutions are often heavily dependent on scientific medical research literature as a primary data source that draws the attacker's attention as a potential target. We present a first study of how the output of MedAI can be polluted with Predatory Publications Presence (PPP).

We study two MedAI systems: *mediKanren* (Disease independent) and *CancerMine* (Disease-specific), which use research literature as primary data input from the research repository *PubMed*, PubMed derived database *SemMedDB*, and NIH translational Knowledge Graphs (KGs). Our study has a three-pronged focus: (1) identifying the PPP in PubMed; (2) verifying the PPP in SemMedDB and the KGs; (3) demonstrating the existing vulnerability of PPP traversing to the MedAI output. Our contribution lies in identifying the existing PPP in the MedAI inputs and demonstrating how predatory science can jeopardize the credibility of MedAI solutions, making their real-life deployment questionable.

## 4.1 Introduction

The undeniable contribution of Artificial Intelligence (AI) in medicine is reaching new milestones. In 2018, FDA approved *IDx-DR* as the first human-independent AI system to detect Diabetic Retinopathy, a leading cause of blindness [5]. Google's AI *Inception v3* could aid in the early detection of lung cancer, which causes 1.7 million deaths globally every year [7].

*Reprinted with permissions from Saini and Saxena. Saini, S. and Saxena, N., 2022. Predatory Medicine: Exploring and Measuring the Vulnerability of Medical AI to Predatory Science. arXiv preprint arXiv:2203.06245. CHIL

### 4.1.1 MedAI Relevance to Precision Medicine

*Precision Medicine* is an innovative approach based on an individual's genetics, health history, environment, and diet. Biomedical research-driven Precision Medicine is promising to provide improved healthcare and lower the overall burden of unknown, delayed, or incorrect diagnosis and treatment [139, 140, 141]. Precision Medicine relies heavily on data and analytics for its adoption into healthcare [109]. MedAI assists Precision Medicine by processing a large amount of information from heterogeneous sources in no time.

### 4.1.2 Impact of Research on MedAI

National Institute of Health (NIH) is the largest public investor of biomedical research globally, investing more than $30 billion a year aiming to provide improved and affordable healthcare [20]. NIH-National Library of Medicine (NLM) maintains the comprehensive research publication source *PubMed* comprising more than 32 million citations for biomedical literature [21]. Many MedAI systems rely upon scientific research publications in medicine as the primary data source for knowledge extraction.

If research gets manipulated through bogus, plagiarized, biased, or fraudulent conclusions, it turns into predatory research, potentially harming patients directly or indirectly. As per Wang et al., around 15% of retraction in biomedical Open Access Journals is due to fraudulent data [142]. Anesthesiologist Scott Reuben and Family Medicine practitioner Anne Kirkman Campbell are few among many others to commit such frauds for money, fame, and position, causing direct harm to patients including loss of lives [143, 144]. For over a decade, multiple studies suggested a potential link between certain vaccines and disorders in children, causing widespread concern and confusion among the medical community and the public [87, 145].

### 4.1.3 Data Pollution in MedAI Solutions

MedAI solutions, which utilize research literature as primary data input, are prone to data pollution. Passive data pollution attacks and active adversarial attacks can impact the integrity and security of MedAI solutions. Untargeted predatory publications induce passive data pollution, while an active adversarial attack is a targeted approach with deliberately poisoning the publication databases through specific predatory journals. Targeted and untargeted predatory publications can inject new data to poison the input dataset. We focus on existing and presumably untargeted data pollution, which can potentially influence the output of the MedAI.

### 4.1.4 Our Contributions

In this paper, we report on a novel case study of passive data pollution in MedAI solutions mediKanren [46], and CancerMine [45] to verify the vulnerability of research publications. To the best of our knowledge, this is the first study to explore the existing data pollution and demonstrate the threat in real-world MedAI solutions.

Our work casts serious doubt on whether research literature-based MedAI solutions are reliable enough to take chances with human lives. Moreover, our results show that MedAI may have no built-in logic to mitigate such threats.

### 4.1.5 Why is this a Security Study?

Predatory research infiltration in MedAI solutions and its impact on MedAIs' output is an exploitable vulnerability, which is new to the security community. While our study is an interdisciplinary effort, we believe that the security community should firsthand know the threat of research literature-based data pollution impacting MedAI solutions. The current and future MedAI systems may avoid these pitfalls if aware of the threats. We also contacted mediKanren team highlighting the underlying vulnerability.

## 4.2 Background

*Artificial Intelligence* (AI) in medicine is being employed in robotic procedures, diagnostics, statistics, and human biology, including *omics* [106]. Though it is still a far-fetched idea to replace the human touch in medicine, MedAI has opened possibilities to save on manual efforts and time to provide faster and adequate decision-making.

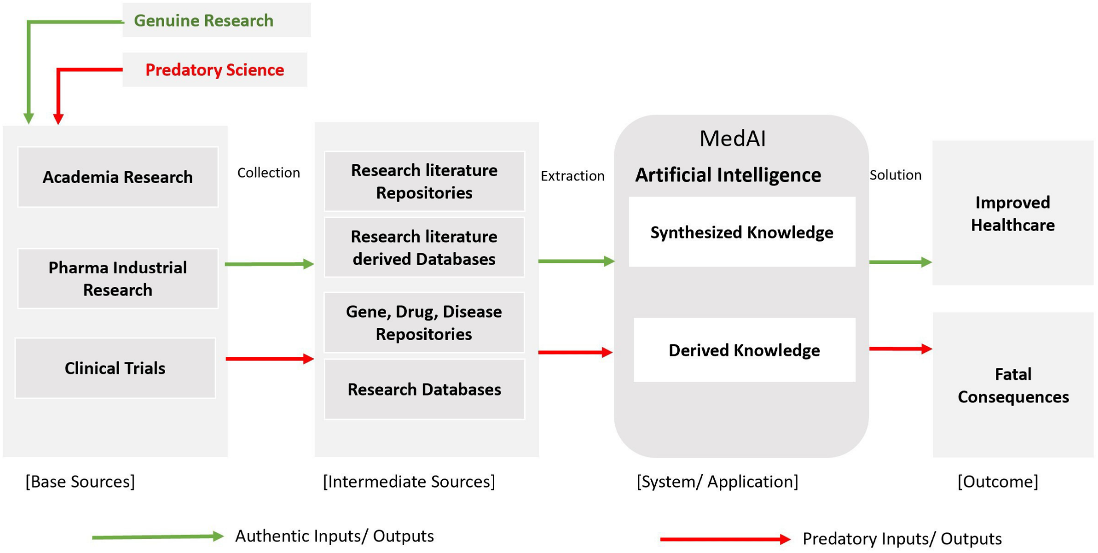


Figure 4.1: High-Level Overview of MedAI Components and Involvement of Predatory Science

### 4.2.1 Revolutionizing Healthcare with MedAI

Around one in 10 Americans is affected by some rare disease, and 80% of around 7000 known rare diseases are genetics-based [146]. Orphan disease diagnosis may take from 1 to 5 years [141] while rare diseases patients suffer from 40% wrong initial diagnosis and 5 to 30 years wait for correct diagnosis [18].

A 2021 data shows that 230 Startups are using AI in drug discovery to improve success rate significantly [10]. A recent effort reported finding successful novel FDA-approved therapeutic recommendations for disorders ranging from undiagnosed and purely symptomatic diseases to genetically diagnosed metabolic disorders [147, 46]. In 2016, Wang et al. reported that pathology image-based MedAI could correctly identify metastatic breast cancer with 92% accuracy. A human expert could determine with 96% accuracy but applying both led to 99.5% accuracy [148]. There is compelling evidence that MedAI can play a vital role in enhancing and complementing the ‘medical intelligence’ of the future clinician [8, 146].

### 4.2.2 Key Methodologies Adopted by MedAI

With advances in computational power and big data, machine learning (ML) is the most widely used AI component in MedAI solutions. Deep Learning (DL) and Natural Language Processing (NLP) are widely employed AI methods to extract meaningful information from the research literature. Artificial Neural networks and fuzzy logic combined as a hybrid intelligent system can accommodate common sense, extract knowledge from raw data, and use human-like reasoning mechanisms [43]. KGs are highly applicable in the medical domain and research as knowledge reasoning can find relationships among diagnoses, diseases, and treatments. Logical inference and probabilistic refinements can develop intelligent systems to suggest treatment options.

### 4.2.3 Security and Integrity of MedAI

Modern Healthcare solutions using AI look promising to save time, money, and effort significantly, but the cost of “trusted but manipulated" information from such MedAI solutions is too high to ignore. Figure 4.1 depicts a high-level overview to show how predatory inputs can compromise a MedAI. An unreliable MedAI output can be fatal in clinical settings, and erroneous results can misalign the overall cycle of future research and healthcare solutions towards the harmful direction. Adversarial attacks on neural networks can cause errors in identifying cancer tumors and damage the confidence in MedAI output [24]. Szegedy et al. showed that very subtle adversarial inputs, which may not appear as pathological, can potentially change the output [149]. A critical insight into why adversarial examples exist is that, given any dataset, the attacker can potentially perturb it in a direction that aligns well with the weights of the MedAI algorithm and thus amplifies its effect on the output [25]. Extracting intelligent information from research literature is a core process in MedAI, and NLP plays a crucial role in this. However, NLP algorithms are vulnerable to adversarial attacks, making it essential to develop robust and efficient algorithms that can protect against such attacks. Ensuring the reliability of MedAI requires a strong focus on protecting against NLP adversarial attacks, and this should be a necessity rather than a preference [28]. MedAI aims to bring all the relevant data together and filter out irrelevant information without skipping more challenging instances. The inclusion of the latest findings from a single paper could transform the treatment options for a patient. However, it is essential to ensure that the data is not only inclusive but also validated to prevent potential harm from predatory research.

A real challenge is maintaining the integrity and security of research data as the basis of MedAI in Precision Medicine and not letting it be Predatory Medicine. Our study highlights the vulnerability of polluted input through predatory publications that may have the potential to generate unreliable MedAI output, especially in finely targeted scenarios of Precision Medicine.

## 4.3 Predatory Research

Medical research has undergone a revolution in recent years, with significant innovation driving a high volume of research publications. However, the *Publish or Perish* culture places enormous pressure on budding scientists and researchers to publish their research, potentially leading to a proliferation of low-quality studies [54]. Publications and citations are being used as a metric for progressing towards the doctorate, employment, promotions, and grants/ funding by state and federal agencies. Opportunists may exploit these trends for their benefit to lure easy targets looking for some publication credits [54]. Research misconduct is an even more significant threat to pollute the research repositories [72, 150, 142]. Misleading conclusions may go undetected for an extended period and may affect clinical practices before being retracted [151]. Accommodating information from heterogeneous sources is vital in MedAI decision-making; it also enables predatory research to get mixed with the authentic inputs.

### 4.3.1 What is a Predatory Journal?

Currently, the characteristics of predatory journals have not been standardized nor broadly accepted by the research community. The majority of potential predatory journals have absent or minimal peer-review process [57, 58] and follow questionable practices focusing on ‘pay to publish’ model [58]. With a possibility of having plagiarized, incorrect, fake data and manipulated results, predatory journals are, in fact, increasingly interfering with genuine research. Jeffrey Beall was the first to raise the concern around 2008 and maintained a list of possibly predatory journals.

DOAJ, Cabell’s list, and other independent online resources are reference points to identify potentially predatory journals [152, 55, 120]. In the absence of any standard definition, we rely on the current list of potential predatory journals from these resources to apply in this work.

### 4.3.2 How Big is Predatory Research?

Beall identified a few potential predatory publishers in 2011, and by 2015, there were estimated as many as 10,000 predatory journals worldwide [153]. The ultimate risk is the altered results of synthesized research because of rapidly increasing numbers of such predatory publications [53, 57]. There have been efforts to expose such practices of accepting fake papers, and recruiting fake editors, but numbers are rising every year [125, 86]. A genuine concern is that there are more suspected predatory journals (10,406) than legitimate journals (10,077) in Cabell’s list [55].

### 4.3.3 PPP Infiltration in Trusted Resources

The possibility of predatory research infiltrating credited research resources like PubMed raises concerns about the reliability of research findings. Manca et al. have reported on the potential infiltration of predatory research in PubMed, highlighting the need for vigilance in ensuring the accuracy and reliability of research publications. [64]. European databases also carry predatory journals and research shows that predatory journals get even more publications than non-predatory journals after being listed in a reputed database [65].

## 4.4 Studied MedAI Solutions

In this work, we focus on research literature-based clinical MedAI solutions, which can impact patients directly. We study two MedAI solutions *mediKanren* and *CancerMine* with different AI approaches, data processing, and output representation [46, 45]. Both studied MedAI solutions heavily rely on research literature inputs, primarily from PubMed. mediKanren has a broad scope of drug repurposing to treat newly diagnosed or unknown diseases based on inferred relationships. CancerMine aims to help with the early diagnosis of cancer types based on genetic profiles.

Table 4.1: Selected 25 Concepts, covering pandemics, common, and rare diseases, to analyze mediKanren prototype 'code' vulnerability to PPP

| Concept Name | Query Term | Category | Type |
|---|---|---|---|
| ADNP | ADNP, Helsmoortel-VanDerAa Syndrome, HVDAS | very rare | disease |
| Adenoid Cystic Carcinoma | Adenoid Cystic Carcinoma, cylindroma | rare | disease |
| BCR | BCR | common | gene |
| Cervical Cancer | Cervical cancer | rare | disease |
| Colorectal Cancer | Colorectal cancer, colon cancer, rectal cancer | common | disease |
| Coronavirus | Coronavirus , HCoV-OC43, SARS-CoV-2 | pandemic | virus |
| Curcumin | Curcumin, Diferuloylmethane, Turmeric | common | substance |
| Dravet Syndrome | Dravet Syndrome, Severe myclonic epilepsy of infancy, SMEI | very rare | disease |
| Ebola | Ebola, rVSV-ZEBOV, ebola virus | pandemic | virus |
| Epithelial Mesenchymal Transition | Epithelial-mesenchymal transition, epithelial mesenchymal transition, | common | cell |
| Gastric Carcinoma | Gastric Carcinoma, Gastric Cancer, Stomach Cancer | rare | disease |
| Imatinib | Imatinib, Imatinib mesylate | common | drug |
| Ischemic Stroke | Ischemic stroke | common | disease |
| Malaria | Malaria, Plasmodium berghei, Plasmodium falciparum | rare | disease |
| Methyltransferas | Methyltransferase | common | disease |
| MIR-200 | MIR-200, MicroRNA-200 | common | gene |
| Neuroblastoma | Neuroblastoma, mycn | very rare | disease |
| Neuroendocrine Prostate Cancer | neuroendocrine prostate cancer | rare | disease |
| Non-Small Cell Lung Cancer | Non-small cell lung cancer, NSCLC | common | disease |
| Ovarian cancer | Ovarian cancer, germ cell cancer, germ cell tumor | rare | disease |
| Pancreatic Cancer | Pancreatic cancer, Pancreatic neuroendocrine | rare | disease |
| Prostate Cancer | Prostate cancer | common | disease |
| T-Cell | T-Cell, t cells, Chimeric Antigen Receptor | common | cell |
| Triple Negative Breast Cancer | Triple negative breast cancer | rare | disease |
| Tyrosine Kinase | Tyrosine, Tyrosine Kinase | common | Protein |

### 4.4.1 mediKanren

*mediKanren* is a MedAI employing logical reasoning over the NIH SemMedDB and translational knowledge graphs to reduce the cost of drug discovery and repurposing [154]. mediKanren is an implementation of miniKanren (Logic Programming Language), Racket (General-purpose Programming Language), Pubmed-derived SemMedDB, NIH translational KGs, and a graphical user interface (GUI) to simplify data exploration to assist Precision Medicine [46]. mediKanren also utilizes KGs with the standardized structure to improve interoperability, ingesting and processing new data from different sources. CURIE (Compact Uniform Resource Identifier), Concept Normalization, and KGs are the basis of mediKanren's functionality.

We study state-of-the-art mediKanren prototypes, *code* and *biolink* [155]. mediKanren is currently under real-world stress testing, which makes it relevant to study for the existing exploitable vulnerabilities [147].

### 4.4.2 CancerMine

*CancerMine* is an automated text-mining approach for extracting gene-disease relationships from PubMed literature to reduce the manual effort and cost of providing timely diagnosis and treatment [156]. CancerMine provides a database of drivers, oncogenes, and tumor-suppressors for different types of cancer. CancerMine extracts information from the PubMed, PubMed Central Open Access (PMCOA) subsets, and PubMed Central Author Manuscript Collection (PMCAMC). CancerMine uses the supervised machine learning approach using the Logistic Regression classifier on word frequencies and semantic features [45]. CancerMine is currently incorporating information from 35,623 PubMed publications. We study the publicly available CancerMine data downloaded in January 2021 to verify how predatory research is navigating to the output of this ML-based MedAI. Without exploiting any ML logic, the focus of this work is to identify and verify the existence of predatory research in the MedAI output.

## 4.5 Preliminaries: Resources and Setup

The study environment was built on Linux OS using Singularity container 2.6.1 utilizing High-Performance Cluster (HPC). We reconstructed the NIH SemMedDB tables for mediKanran analysis using Open GPLv2 MariaDB ver 10.3.10 MySQL. CancerMine inputs and outputs in TSV/CSV file formats were analyzed using MS Excel.

### 4.5.1 Key Terms and Definitions

Each article on PubMed has a unique identifier called *PMID* (PubMed ID). In this work, we are using the SemMedDB (semmedVER40_R), with PubMed data processed up to June 30, 2018. NIH defines a rare disease as a condition that affects fewer than 200,000 people in the US. A rare disease is known as an orphan disease if drug companies are not interested in developing treatments [146, 141]. A *Concept* is defined as a unique medical term as per NIH Unified Medical Language System (UMLS) Metathesaurus, and *CUI* is the Concept Unique Identifier [110]. *PREDICATE* is the representation of the relationship between any two medical concepts identified as SUBJECT and OBJECT. For example, Imatinib (SUBJECT) Treats (PREDICATE) Mastocytosis (OBJECT). Compact Uniform Resource Identifiers (CURIEs) serve as machine-readable markers for different databases. Graph vertices represent medical concepts in the KGs, while directed graph edges depict relationships between concepts. Edges also show metadata about source and evidence backing the represented relationship [46].

## 4.6 Data Extraction

We extracted a set of predatory journals based on *Beall's list of potential predatory journals and publishers* , and compared with more current list. We created an advanced query with the list of existing predatory journals in PubMed, and downloaded the metadata for 47,051 predatory publications. We extracted 8,289 retracted publications from PubMed with "Retraction of publication [Publication Type]", a query adopted from [151].

`https://beallslist.net/`

`https://web.archive.org/web/20211220083526/https://predatoryjournals.com/journals/`*Web Archives-predatoryjournals.com.*

### 4.6.1 mediKanren Data Inputs

We study 25 concepts to study in this work under common, rare, very rare, and pandemic categories to see the extent of PPP in diverse scenarios. For each of these 25 concepts, we extracted the set of 200 rows (100 rows of predatory PMIDs and 100 rows of non-predatory PMIDs) from the PREDICATION table of SemMedDB. A .csv file was prepared for each concept except for a few rare concepts with less than 200 publications. Table 4.1 provides details of the selected concepts.

Prototype biolink is currently utilizing data inputs from four NIH translational knowledge graphs *RoboKop*, *SemMed*, *Orange*, and *Rtx*. mediKanren team provided the local copy of NIH KGs employed in biolink. We applied python scripts to extract PMIDs from applied NIH Translational KGs to study the presence of predatory PMIDs.

### 4.6.2 CancerMine Data Inputs

We utilized data from `https://zenodo.org/record/4304808#.X_cwITSSlPb` *CancerMine Zenodo repository*. The primary raw input cancermine_unfiltered.tsv is processed to create other two main inputs as cancermine_collated.tsv and cancermine_sentences.tsv. cancermine_collated.tsv contains the cancer gene roles supporting citation counts. cancermine_sentences file carries the sentences for the cancer gene roles. This SENTENCE file contains information on the source publication (e.g., journal, publication date, etc.), the actual sentence, and the cancer type, gene, and role extracted from the sentence.

## 4.7 Passive Data Pollution Verification

Based on our hypothesis, PubMed has existing data pollution with the predatory publications, and MedAI solutions use this polluted dataset. Other than predatory journal publications, the PubMed-derived database also carries the retracted research publications. To verify PPP in SemMedDB tables, we queried restructured SemMedDB on HPC to find predatory PMIDs.

For mediKanren prototype *Code*, we run each concept-specific CSV file through Racket commands to pre-process the input. We execute all the end-user queries on Racket ver 7.4 and gui-simple.rkt file as working GUI for code prototype of mediKanren. *Biolink* prototype works with

Table 4.2: PPP in SemMedDB Tables

| SemMed Table | Total Rowcount | PPP Rowcount |
|---|---:|---:|
| PREDICATION | 97,772,561 | 256,641 |
| SENTENCE | 187,449,479 | 357,384 |
| ENTITY | 1,369,837,426 | 2,735,289 |
| CITATION | 29,137,782 | 44,670 |

gui_simple_v3.rkt GUI on KGs, and we run sample queries to verify the data pollution in the input and output of the mediKanren.

For CancerMine, we verified the overall PPP in provided datasets from their repositories, queried online tool, and downloaded data to verify the PPP in tool-provided subsets. We utilized the same set of predatory PMIDs extracted from PubMed for mediKanren and CancerMine PPP verification.

### 4.8 Results

Our results show that predatory publications have a significant presence in the research literature repository PubMed (47,051 predatory publications). Results also validate that predatory research can traverse from PubMed to MedAI output. For detailed analysis, we organize our results into five sub-sections: (1) PPP in SemMedDB; (2) Retracted publications in SemMedDB; (3) PPP in NIH Translational KGs; (4) PPP in mediKanren GUI output; and (5) PPP in CancerMine.

#### 4.8.1 PPP in SemMedDB

The PREDICATION table is carrying concept, sentence ID, predicate, SUBJECT_CUI, and OBJECT_CUI. Although predatory PMIDs numbers seem smaller compared to the huge dataset size, the fraction holds a large number (44,665) of predatory publications. Table 4.2 shows the overall PPP infiltration in SemMedDB tables.

In this work, we further verified the presence of PPP in SemMedDB for 25 selected concepts, highlighting the potential risks associated with relying on unvalidated data sources in MedAI. PPP count varies from under 5 (very rare diseases: Dravet Syndrome, ADNP) to around 5000 or

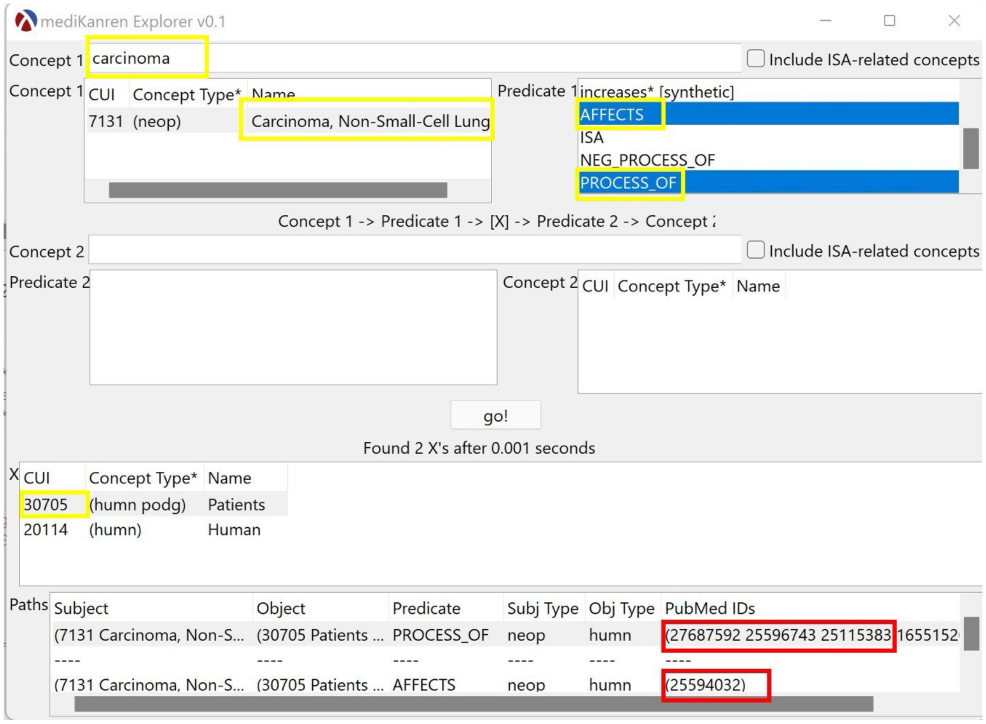
Table 4.3: SemMedDB: PPP For Studied 25 Concepts

| Concept | Category | PPP |
|---|---|---:|
| ADNP | very rare | 5 |
| Adenoid Cystic Carcinoma | rare | 51 |
| BCR | commom | 659 |
| Cervical Cancer | rare | 1,005 |
| Colorectal Cancer | common | 5,510 |
| Coronavirus | pandemic | 6 |
| Curcumin | common | 500 |
| Dravet Syndrome | very rare | 1 |
| Ebola | pandemic | 15 |
| Epithelial Mesenchymal Transition | common | 679 |
| Gastric Carcinoma | rare | 3,727 |
| Imatinib | common | 424 |
| Ischemic Stroke | common | 266 |
| Malaria | rare. | 252 |
| Methyltransferase | common | 57 |
| MIR-200 | common | 300 |
| Neuroblastoma | very rare | 1,269 |
| Neuroendocrine Prostate Cancer | rare | 11 |
| Non-small cell Lung Cancer | common | 4,143 |
| Ovarian Cancer | rare | 2,718 |
| Pancreatic Cancer | rare | 2,139 |
| Prostate cancer | common | 3,849 |
| T-Cell | common | 4,378 |
| Triple Negative Breast Cancer | rare | 246 |
| Tyrosine Kinase | common | 2,061 |

more(common cancer: Non-small-cell Lung Cancer, Colorectal Cancer). We can see that PPP is higher, in general, for common diseases, and lower for rare and very rare diseases, but also higher PPP for rare diseases like *Gastric Carcinoma*. We did not analyze the percentage of concept-specific PPP for these 25 concepts as any presence may be damaging for a specific scenario. Table 4.3 shows the PPP count in SemMedDB for all 25 concepts studied in this work.

### 4.8.2 Retracted Publications in SemMedDB

There are 8,289 retracted publications found on PubMed. An increase from 311 in 2010 to 860 in 2019 indicates the rising problem of retracted publications in PubMed. These retracted articles



Concept 1: carcinoma

□ Include ISA-related concepts

Concept 1

| CUI | Concept Type* | Name |
|---|---|---|
| 7131 | (neop) | Carcinoma, Non-Small-Cell Lung |

Predicate 1
- increases* [synthetic]
- AFFECTS
- ISA
- NEG_PROCESS_OF
- PROCESS_OF

Concept 1 -> Predicate 1 -> [X] -> Predicate 2 -> Concept 2

Concept 2:

□ Include ISA-related concepts

Predicate 2

Concept 2

| CUI | Concept Type* | Name |
|---|---|---|

Found 2 X's after 0.001 seconds

X

| CUI | Concept Type* | Name |
|---|---|---|
| 30705 | (humn podg) | Patients |
| 20114 | (humn) | Human |

Paths

| Subject | Object | Predicate | Subj Type | Obj Type | PubMed IDs |
|---|---|---|---|---|---|
| (7131 Carcinoma, Non-S... | (30705 Patients ... | PROCESS_OF | neop | humn | (27687592 25596743 25115383 1655152 |
| ---- | ---- | ---- | ---- | ---- | ---- |
| (7131 Carcinoma, Non-S... | (30705 Patients ... | AFFECTS | neop | humn | (25594032) |

Figure 4.2: mediKanren ‘code’ GUI [Concept, Predicate and Object- yellow; Predatory PMIDs - red]

also exist in SemMedDB with their original PMIDs and with retraction notice PMIDs. A query with “Retracted:” returns 538 rows from the SENTENCE table, and the PREDICATION table returns 398 rows.

### 4.8.3 PPP in NIH Translational KGs

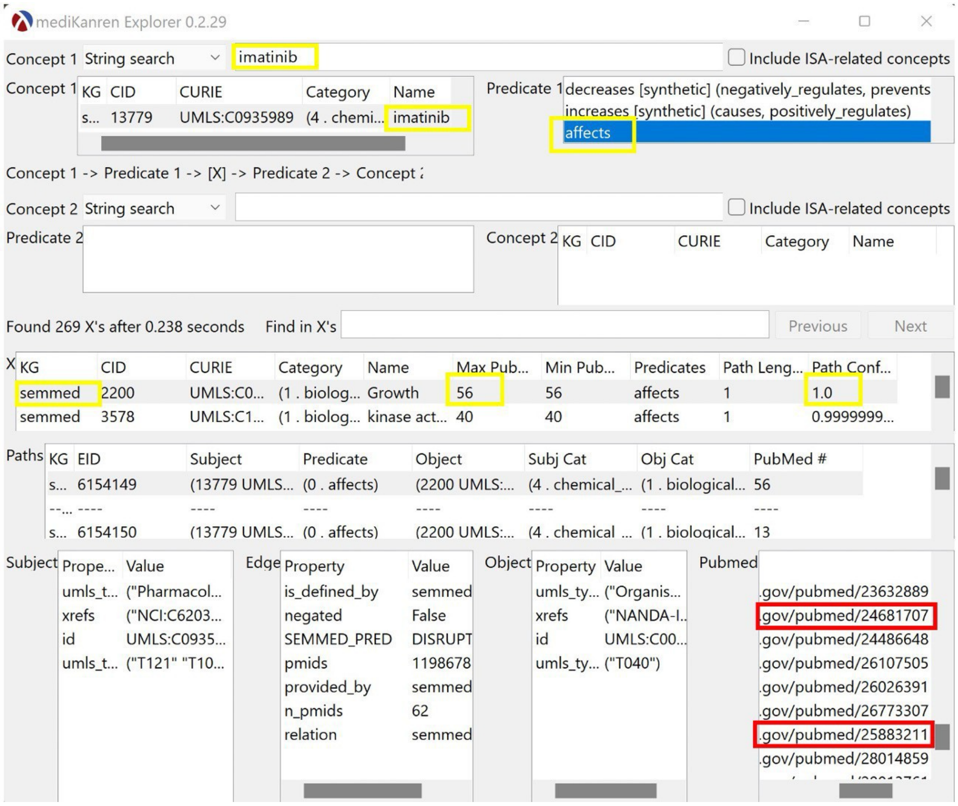
mediKanren prototype ‘biolink’ can create and run queries on translational KGs to deliver the suggestions in a comprehensible manner [154]. Table 4.4 shows the PPP in all four translational KGs. The percentage of PPP in these KGs does not appear huge, but PPP is significant, especially in the largest KG SemMed. For example, KG Rtx is with the most significant percentage of 1.3 with a smaller PPP (99). KG SemMed shows a smaller percentage (.37) with a much higher PPP (164,324). This observation also validates that even a lower number with a high percentage on any concept will have a higher probability to appear in MedAI output.

Table 4.4: PPP in NIH Translational KGs

| Database | Total Rowcount | PPP Rowcount |
|---|---:|---:|
| Robokop | 5,367,905 | 5,676 |
| SemMed | 44,245,576 | 164,324 |
| Orange | 836,118 | 287 |
| Rtx | 7,664 | 99 |

### 4.8.4 PPP in mediKanren GUI Output

Specific input datasets for 25 concepts are fed independently to mediKanren prototype *code* to analyze targeted queries to follow the Precision Medicine approach. We executed multiple queries



mediKanren Explorer 0.2.29

Concept 1 String search: imatinib ☐ Include ISA-related concepts

Concept 1

| KG | CID | CURIE | Category | Name |
|---|---|---|---|---|
| s... | 13779 | UMLS:C0935989 | (4 . chemi... | imatinib |

Predicate 1
decreases [synthetic] (negatively_regulates, prevents
increases [synthetic] (causes, positively_regulates)
affects

Concept 1 -> Predicate 1 -> [X] -> Predicate 2 -> Concept 2

Concept 2 String search ☐ Include ISA-related concepts

Predicate 2

Concept 2

| KG | CID | CURIE | Category | Name |
|---|---|---|---|---|

Found 269 X's after 0.238 seconds Find in X's

X

| KG | CID | CURIE | Category | Name | Max Pub... | Min Pub... | Predicates | Path Leng... | Path Conf... |
|---|---|---|---|---|---|---|---|---|---|
| semmed | 2200 | UMLS:C0... | (1 . biolog... | Growth | 56 | 56 | affects | 1 | 1.0 |
| semmed | 3578 | UMLS:C1... | (1 . biolog... | kinase act... | 40 | 40 | affects | 1 | 0.9999999... |

Paths

| KG | EID | Subject | Predicate | Object | Subj Cat | Obj Cat | PubMed # |
|---|---|---|---|---|---|---|---|
| s... | 6154149 | (13779 UMLS... | (0 . affects) | (2200 UMLS:... | (4 . chemical_... | (1 . biological... | 56 |
| --... | ---- | ---- | ---- | ---- | ---- | ---- | ---- |
| s... | 6154150 | (13779 UMLS... | (0 . affects) | (2200 UMLS:... | (4 . chemical ... | (1 . biological... | 13 |

Subject

| Prope... | Value |
|---|---|
| umls_t... | ("Pharmacol... |
| xrefs | ("NCI:C6203... |
| id | UMLS:C0935... |
| umls_t... | ("T121" "T10... |

Edge

| Property | Value |
|---|---|
| is_defined_by | semmed |
| negated | False |
| SEMMED_PRED | DISRUPT |
| pmids | 1198678 |
| provided_by | semmed |
| n_pmids | 62 |
| relation | semmed |

Object

| Property | Value |
|---|---|
| umls_ty... | ("Organis... |
| xrefs | ("NANDA-I... |
| id | UMLS:C00... |
| umls_ty... | ("T040") |

Pubmed
.gov/pubmed/23632889
.gov/pubmed/24681707
.gov/pubmed/24486648
.gov/pubmed/26107505
.gov/pubmed/26026391
.gov/pubmed/26773307
.gov/pubmed/25883211
.gov/pubmed/28014859

Figure 4.3: mediKanren prototype 'biolink' GUI [Concept, Predicate and Object, PubMed articles, and Path confidence- yellow; Some of Predatory PMIDs - red]

Table 4.5: PPP in MediKanren ‘code’ GUI Output with Predatory and Non-predatory PMIDs

| **Concept Name** | **(Subject_CUI, Oject_CUI, Predicate)** | **Pred** | **Non-Pred** | **Total** |
|---|---|---|---|---|
| ADNP | (1334473, 3394, INTERACTS_WITH) | 1 | 0 | 1 |
| Adenoid Cystic Carcinoma | (10606, 30705, PROCESS_OF) | 3 | 2 | 5 |
| BCR | (812385, 935989, COEXISTS_WITH) | 1 | 0 | 1 |
| Cervical Cancer | (7847, 30705, PROCESS_OF) | 4 | 6 | 10 |
| Colorectal Cancer | (7102, 27651, AFFECTS) | 1 | 0 | 1 |
| Coronavirus | (10076, 15576, PROCESS_OF) | 1 | 0 | 1 |
| Curcumin | (10467, 6826, AFFECTS) | 1 | 0 | 1 |
| Dravet Syndrome | (3064, 15827, LOCATION_OF) | 1 | 0 | 1 |
| Ebola | (282687, 30705, PROCESS_OF) | 1 | 3 | 4 |
| Epithelial Mesenchymal Transition | (14609, 346109, LOCATION_OF) | 1 | 0 | 1 |
| Gastric Carcinoma | (24623, 30705, PROCESS_OF) | 2 | 5 | 7 |
| Imatinib | (935989, 23437, TREATS) | 5 | 3 | 8 |
| Ischemic Stroke | (948008, 30705, PROCESS_OF) | 7 | 4 | 11 |
| Malaria | (24530, 21311, ISA) | 1 | 2 | 3 |
| Methyltransferas | (25831, 11315, PART_OF) | 1 | 0 | 1 |
| MIR-200 | (1537839, 13081, AFFECTS) | 1 | 0 | 1 |
| Neuroblastoma | (27819, 27651, ISA) | 2 | 2 | 4 |
| Neuroendocrine Prostate Cancer | (936223, 30705, PROCESS_OF) | 1 | 1 | 2 |
| Non-Small Cell Lung Cancer | (7131, 30705, NEG_PROCESS_OF) | 7 | 4 | 11 |
| Ovarian cancer | (29925, 1520166, PROCESS_OF) | 1 | 1 | 2 |
| Pancreatic Cancer | (235974, 30705, PROCESS_OF) | 2 | 2 | 4 |
| Prostate Cancer | (376358, 30705, PROCESS_OF) | 5 | 6 | 11 |
| T-Cell | (279592, 30705, PROCESS_OF) | 1 | 0 | 1 |
| Triple Negative Breast Cancer | (6142, 30705, PROCESS_OF) | 3 | 3 | 6 |
| Tyrosine Kinase | (206364, 27651, ASSOCIATED_WITH) | 1 | 0 | 1 |

on each concept for different triples of Subject_CUI, Object_CUI, and Predicate. We found that the majority of the queries show predatory PMID(s) in mediKanren GUI output. Figure 4.2 shows non-predatory PMIDs appearing in mediKanren prototype *code*. For example, the only PMID returned for the concept *Non-small-cell lung cancer* and predicate *AFFECTS* is predatory PMID, and for predicate *PROCESS_OF* output contains both predatory and non-predatory PMIDs.

Table 4.5 shows 25 representative cases for 25 concepts validating the problem of PPP traversing from PubMed to mediKanren. Data demonstrates that mediKanren may pick predatory PMIDs for common as well as rare diseases if predatory PMIDs are present in the inputs.

PPP in SemMedDB can affect the MedAI outcome even with a low number of predatory PMIDs. If a particular case has only a few publications on a specific concept, and there are more predatory than non-predatory, that may mislead the decision. For example, the concept *Imatinib* for predicate *TREATS* returns five predatory and three non-predatory PMIDs from a pool of 200 rows (100 predatory PMIDs and 100 non-predatory PMIDs). MediKanren GUI showed clear evidence of picking up predatory PMIDs in output with targeted input datasets.

For mediKanren prototype *biolink*, we queried the whole input for all four KGs without any specific test sampling. Figure 4.3 shows that predatory/ non-predatory PMIDs are appearing in prototype biolink's output. We did not verify the presence of retracted PMIDs in mediKanren GUI output in this work. However, based on the working logic, MedAI is expected to pick any present publication (predatory or non-predatory) for a specific user query.

### 4.8.5 PPP in CancerMine

We analyze CancerMine data directly downloaded from the data repository and downloaded output files from online queries on the CancerMine web tool. Table 4.6 shows that overall predatory PMIDs have above 5% in raw data and extracted Sentence CSV file.

Table 4.6: PPP in CancerMine Dataset

| | **unfiltered tsv** | **sentences tsv** |
|---|---:|---:|
| Total PMIDs | 172,443 | 55,881 |
| PPP | 9,479 | 2,873 |
| PPP % | 5.50% | 5.14% |
| Unique PMIDs | 73,099 | 35,623 |
| Unique PPP | 3,384 | 1,637 |
| PPP % | 4.63% | 4.60% |

We observe that 51.64% of predatory PMIDs have a prediction probability of 0.7 and higher (max is .9997), indicating the role of these publications on overall prediction probability. We

Downloaded raw data from the CancerMine repository to carry 455 cancer types, and 150 out of 455 types show predatory PMIDs. Breast cancer shows the highest number of predatory PMIDs. We cross-referenced the specific gene-cancer-role details from the CancerMine data repository and its web application output queries. We observe PPP in multiple subsets of targeted queries regarding a particular cancer type, gene, and role of the gene. Table 4.7 is presenting PPP for few other cancer-gene-role triples of selected cancer types. We selected cancer type similar to some concepts explored for other MedAI *mediKanren* in this work. PPP can be much higher on a particular cancer-role-gene triple. For example, Neuroblastoma-Driver-MYCN filters a set of 164 PMIDs, and 20.12% (33) PMIDs are predatory. Another case of Stomach cancer-Tumor_Suppressor-TFF1 is a much smaller set of 16 PMIDS with 25% (4) predatory PMIDs.

We successfully show these predatory PMIDs in the input sources (SemMedDB and CancerMine datasets) and the output (mediKanren GUI output and CancerMine website downloaded output), which verifies the existing threat in real-life research-literature based MedAI solutions. Though we are not exploring the threat of ML manipulation in this work, it may be possibile to analyze predatory publications' impact on shifting prediction probability.

Table 4.7: CancerMine Predatory and Non-Predatory PMIDs for Cancer-Role-Gene Triple

| Type of Cancer | Role | Gene | PMIDs | Non-Pred | Pred |
|---|---|---|---:|---:|---:|
| Glioblastoma | Oncogene | CDK4 | 5 | 4 | 1 |
| Lung small cell cancer | Driver | ALK | 1 | 0 | 1 |
| Colorectal cancer | Oncogene | KRAS | 167 | 158 | 9 |
| Prostate cancer | Tumor Suppressor | PTEN | 137 | 128 | 9 |
| Breast cancer | Oncogene | FOXM1 | 8 | 7 | 1 |
| Pancreatic cancer | Driver | KRAS | 171 | 160 | 11 |
| Neuroblastoma | Driver | MYCN | 164 | 131 | 33 |
| Stomach cancer | Tumor Suppressor | TFF1 | 16 | 12 | 4 |
| Pituitary cancer | Oncogene | PTTG1 | 17 | 15 | 2 |
| Ovarian cancer | Tumor Suppressor | BRCA1 | 123 | 119 | 4 |

## 4.9 Limitations and Challenges

Though our work clearly shows the infiltration of PPP in PubMed and affecting the output of mediKanren GUI, there are several factors to limit the outcome of this study. One of the challenges is determining the valid list of all known predatory journals at any given point in time.

The frequency of derived database updates also can affect the PPP. All the mediKanren queries for the prototype *code* are executed with targeted much smaller datasets for selected concepts to highlight the threat. We believe that results will be comparable even if we execute queries on the whole dataset. This work only focuses on the existing PPP verification without any current analysis of data training and prediction probability affected by PPP.

## 4.10 Discussion and Future Work

As observed in our study, MedAI relies on the credibility of the reputed data source. There is no in-built defense logic in both the studied MedAI solutions to minimize the threat of predatory research influencing the output. Our work indicates the need to have a better defense system at the source level to minimize the threat of data pollution.

We shared our findings and concerns with the mediKanren team, and they have acknowledged the threat. Though this work is specific to research literature-based clinical MedAI solutions, we looked at other current state-of-the-art solutions for existing defense strategies towards predatory science, if any exists. Iris.ai is specific for researchers to provide relevant research literature based on the research hypothesis. We observed predatory publications in iris.ai results as well, and the iris.ai team also confirmed not to have any current defense mechanism to filter out potential predatory research [51]. Though currently not explored but in a possible adversarial attack scenario, in near future, it can be viable to produce targeted predatory publications in bulk through new-age NLP text-generator tools like GPT-3 [136, 135]. A targeted adversarial attack on a rare disease can make it look valid while injecting fake data with alternative conclusions. This approach may impact rare and unknown disease scenarios more as there is little available research to cross-validate. Approximately 50% of medical providers showed concern about producing fatal errors and technical/operational glitches. These concerns resonate with the known ethical and regulatory challenges with MedAI solutions involving privacy, data integrity, accessibility, accountability, transparency, and liability [157, 158]. In the age of social media and web-based information, the future trustworthy information extraction will have more significant challenges and higher stakes. Future work will further study the impact of predatory research, including retracted publications on MedAI solutions. We plan to look at other security vulnerabilities to better defend against information pollution.

### 4.11 Conclusion

Our work concludes that polluted inputs can cause a possible failure of any MedAI to deliver the intended output. Predatory research is on the rise and may further degrade research literature-based MedAI solutions' credibility. Studied MedAI solutions treat all research data as trusted and do not consider predatory research-induced data pollution. In the absence of any defense, MedAI solutions may produce unreliable output. Existing data pollution fuels motivation for more targeting attacks. Our study shows clear evidence of how predatory research information is navigating through publication channels and eventually may alter patient care decisions if used in the clinical settings without resolving the existing threat of predatory research intrusion. We are confident that verifying the vulnerabilities early in the process will contribute in developing more robust solutions for taking Precision Medicine to the next level in broader settings.

## 4.12 A Proposed Defense Mechanism: Automated Predatory Research Classification

Predatory research induced data pollution can impact any advanced technological solution which is either using the research literature repository data or its derived databases as primary inputs. Based on the nature of AI applied in such tools, it can be algorithm-dependent or algorithm-independent problem, which may impact the outcome. In case of ML-based tools, passive poisoning of training occurs when malicious or biased examples inadvertently contaminate the training data, resulting in biased or harmful models. Predatory research publications can contribute to this problem by spreading misinformation or publishing flawed studies that include inaccurate or biased data. This can negatively impact the performance and fairness of ML models, leading to unintended consequences in real-world applications.

Predatory journals lack a widely accepted, objective set of features for identification, but are commonly characterized by lack of peer review, spamming for submissions, poor website quality, questionable publishing standards, hidden fees, and fake claims about impact factor (IF) and being indexed in reputed research repositories.

In this work, we present a defense strategy to identify potential predatory research. Our goal is to minimize the infiltration of predatory research into trusted research literature repositories and dependent derived databases. Figure 4.4 illustrates the possibility of identifying predatory research at different stages, such as before adding to PubMed, existing PubMed publications, information transfer between PubMed and derived databases, or verifying existing derived databases.

### 4.12.1 Background

It has been observed that research-oriented universities provide guidelines to help researchers avoid potential predatory venues [159, 160]. Additionally, there are studies that document observations and provide pointers for identifying potentially predatory publishing venues [161, 162]. These guidelines, most commonly, do refer to *Beall's List*, and other online resources, [119].

Beall's List- `https://beallslist.net/`
`https://www.openacessjournal.com/blog/predatory-journals-list/`
`https://predatoryreports.org/news/f/list-of-all-mdpi-predatory-publications`

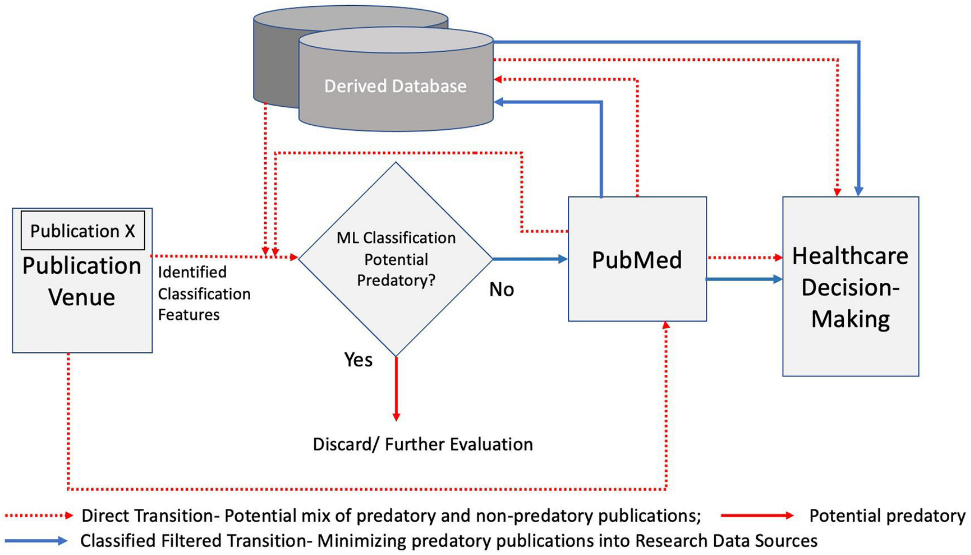

Derived Database
Publication X
Publication Venue
Identified Classification Features
ML Classification Potential Predatory?
No
Yes
Discard/ Further Evaluation
PubMed
Healthcare Decision-Making
Direct Transition- Potential mix of predatory and non-predatory publications;
Potential predatory
Classified Filtered Transition- Minimizing predatory publications into Research Data Sources

Figure 4.4: Applying ML Algorithms to Classify and Predict Potential Predatory Research

Web platforms like *Clarivate*, present the publishing data of the journals indicating their years of operation, Impact Factor (acknowledged credibility), number of citations (relevance and potential influence on other research works), scope, publishers' details etc, and can be considered a resource to verify details of a legitimate journal. DOAJ is another web resource to find details about Open Access (OA) journals [120]. As per DOAJ, quality of a journal is accessed by mostly business transparency, but peer review process remains as one of the key criteria to evaluate the quality of a journal. However, DOAJ article explains that blacklists like Beall's list or other lists maintained can never be exhaustive, and researchers should pay more attention to whitelisted journals [163].

### 4.12.2 A Preliminary Analysis of Classifying Predatory Research Venues

### 4.12.3 Methodology and Data Collection

In this work, we identify a list of features which can be a measurable criterion in defining the quality of a publishing journal. We selected twenty each of verified journals and potentially predatory journals indicated by either/or being present in the above discussed blacklists and/or being absent in the whitelists. As per our observations and discussed features in the research

Cabelles Predatory Report- `http://www2.cabells.com/predatory` literature, we categorize journal's features in four major categories under *Identification*, *Presence*, *Performance*, and *Validation*. We studied the feasibility of data collection and identified 32 features to consider in this work. Table 3.2 presents the features we use to analyze the probability of predatory classification in this work. There are 10 features as Identification, 7 features under Presence, 6 under Performance, and 9 under Validation categories.

We collect the data manually from multiple web resources, as well as studying the journal's and publisher's websites for all the identified features in this work. We manually label each instance of training data based on available information to best identify predatory journals. We then apply machine learning algorithms to classify the predatory publishing venues. Data was analyzed by ML algorithms, and we compare the model's performance using machine learning evaluation matrices shown in Figure 4.5.

### 4.12.4 ML Classification Algorithms

We analyzed the journals' data through multiple ML algorithms. We provide a brief description of the used classification methods as following.

A *RandomForest* is an ensemble learning method that combines multiple decision trees to create a more accurate and robust model. It generates a large number of decision trees using random subsets of the training data and features, which helps reduce overfitting and improve generalization. Random forests are widely used in image classification, text classification, bioinformatics, and other domains for their high accuracy, scalability, and interpretability.

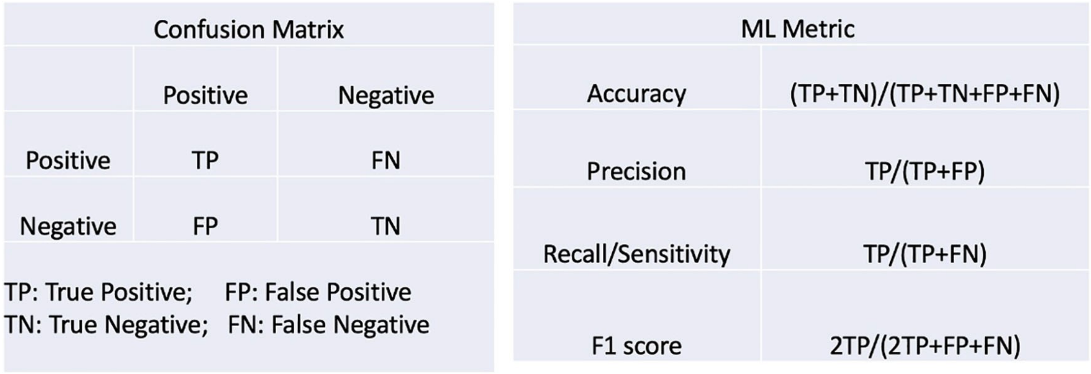
The *Naïve Bayes classifier* is a probabilistic classifier that assumes all features are independent of each other, given the category variable. The NB classifier performs well in many real-world applications, such as text classification, keyphrase extraction, and medical diagnosis [164].

*Support Vector Machines (SVM)* is a supervised learning algorithm that can be used for classifying journals as predatory or non-predatory based on measurable features such as publication fees, editorial board composition, journal metrics, and indexing and archiving status. SVM can

https://jcr.clarivate.com/jcr/home
https://www.ncbi.nlm.nih.gov/nlmcatalog/

Table 4.8: Journal Features for Predatory Classification; JCR (Journal Citation Reports)

| Feature Name | Feature Category |
|---|---|
| Country of Publication | Identification |
| Editorial policiesAvialble | Validation |
| eISSN | Identification |
| IsOpenAccess | Presence |
| ISSN | Identification |
| Journal Abbr | Identification |
| Journal Age | Presence |
| Journal Citation Index (JCI) | Performance |
| Journal Editor's Contact | Presence |
| Journal Impact Factor (JIF) | Performance |
| Journal Title | Identification |
| Journal-Total Citations | Performance |
| Journal's Website Quality | Validation |
| DOA Indexed | Validation |
| PubMed Indexed | Validation |
| Manuscript Guidelines | Presence |
| Submission Method | Presence |
| Is Claimed JIF same as in JCR? | Validation |
| MEDLINE Indexed | Validation |
| NLM ID | Identification |
| Peer-Reviewed | Validation |
| Publication Frequency | Performance |
| Publication Start Year | Presence |
| Publication Type | Identification |
| Publisher City | Identification |
| Is Publisher same as in JCR? | Validation |
| Publishing Fee | Presence |
| Submit-to-publish Period | Performance |
| Web of Science Categories | Identification |
| Website Language Efficiency | Validation |
| Website Claimed JIF | Performance |
| websitePublisher | Identification |



| Confusion Matrix | | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

TP: True Positive; FP: False Positive
TN: True Negative; FN: False Negative

| ML Metric | |
|---|---|
| Accuracy | (TP+TN)/(TP+TN+FP+FN) |
| Precision | TP/(TP+FP) |
| Recall/Sensitivity | TP/(TP+FN) |
| F1 score | 2TP/(2TP+FP+FN) |

Figure 4.5: ML Evaluation Metrics

handle imbalanced datasets and is useful for identifying rare predatory journals. However, SVM may not perform well in datasets with high levels of noise or outliers, which can be present in datasets of predatory journals.

ML algorithms are useful for identifying predatory journals, especially when combined with feature selection, cross-validation, and expert judgment. It is important to carefully evaluate the model’s performance using metrics such as precision, recall, and F1 score, and to validate it on new, unseen data for generalizability.

# 5. PRIVACY AND SECURITY THREATS OF MOBILE MENTAL HEALTH APPS*

Mental health is an extremely important subject, especially in these unprecedented times of the COVID-19 pandemic. Ubiquitous mobile phones can equip users to supplement psychiatric treatment and manage their mental health. Mobile Mental Health (MMH) apps emerge as an effective alternative to assist with a broad range of psychological disorders filling the much-needed patient-provider accessibility gap. However, it also raises significant concerns with sensitive information leakage. The absence of a transparent privacy policy and lack of user awareness may pose a significant threat to undermining the applicability of such tools. We conducted a multifold study of - 1) Privacy policies (Manually and with *Polisis*, an automated framework to evaluate privacy policies); 2) App permissions; 3) Static Analysis for inherent security issues; 4) Dynamic Analysis for threat surface and vulnerabilities detection, and 5) Traffic Analysis.

Our results indicate that apps' exploitable flaws, dangerous permissions, and insecure data handling pose a potential threat to the users' privacy and security. The Dynamic analysis identified 145 vulnerabilities in 20 top-rated MMH apps where attackers and malicious apps can access sensitive information. 45% of MMH apps use a unique identifier, *Hardware Id*, which can link a unique id to a particular user and probe users' mental health. Traffic analysis shows that sensitive mental health data can be leaked through insecure data transmission. MMH apps need better scrutiny and regulation for more widespread usage to meet the increasing need for mental health care without being intrusive to the already vulnerable population.

## 5.1 Introduction

As per the Centers for Disease Control and Prevention (CDC), from mid-2020 to early 2021 (after starting of COVID-19), there is an increase in anxiety or depressive disorder from 36.4%

*Reprinted with permissions from Saini et al. Saini, S., Panjwani, D. and Saxena, N., 2022, August. Mobile Mental Health Apps: Alternative Intervention or Intrusion? In 2022 19th Annual International Conference on Privacy, Security & Trust (PST) (pp. 1-11). IEEE.

to 41.5%. The percentage of unmet mental health care needs is increased from 9.2% to 11.7% [165]. Depression, anxiety, and other mental disorders can cripple everyday functioning and even can claim lives as suicide is the tenth leading cause of death in the U.S. [166]. Winkler et al. found that COVID-19 increased the prevalence of major depressive disorder and suicide risk three times and almost doubled the current anxiety disorders [167]. However, a shortage of providers may encourage people to seek alternatives more often for immediate help [168]. Ubiquitous mobile devices can help to bridge the patient-provider gap and health divide for underserved and hard-to-reach populations to manage their mental health needs [169]. As per Wang et al., MMH apps have the potential to improve mental health, but most of the currently available apps lack clinical evidence to support their efficacy [170]. Mainstream clinical practice incorporating MMH apps is also challenging as research is open to ensure that technological vulnerabilities do not compromise the privacy and safety of patients [171, 172, 173].

Though there are privacy and security vulnerabilities which may impact any mobile app, but Mental health apps often collect sensitive data such as patients' psychological symptoms, diagnoses, and treatment history, which can be used to identify and target vulnerable individuals. Misuse of such sensitive info can discriminate already struggling population in employment or insurance decisions. If the information is leaked or sold to third parties, it may lead to potential harm such as stigmatized or ostracized by others, leading to social isolation with increased risk of depressive episodes, anxiety attacks, and even suicide. Mobile app developers may need standard guidelines to develop a secure MMH app with the help of the medical community on the usability [174]. Selecting a secure and reliable MMH app from a vast pool of available MMH apps may be daunting for the end user. A well-defined privacy policy may be the starting point for app users in making decisions balancing information-sharing and preserving privacy. No explicit declaration on apps' permissions, information collected, and intended usage of collected information makes it difficult to grade the apps on security and privacy parameters.

There is a lack of reproducible rigorous scientific testing to support rapidly developed MMH apps and quick launches in awaiting market. Currently, there is no national standard for evaluating the effectiveness of the available hundreds of mental health apps [13]. It is essential to maintain transparency in privacy policies to build the needed trust for the broader usage of MMH apps. Our contribution is manifold through this study as follows:

- Analyzing the availability, accessibility, and deficiencies of MMH apps' privacy policies.
- Exploring threat surface, dangerous permissions, and injection vulnerabilities.
- Analyzing the apps' code files exposing major exploitable security and privacy flaws.
- Network traffic payload analysis to study insecure data transactions.

Our results show that 95% of 20 Highly Popular (HP) apps provide a privacy policy compared to 65% of 20 Less-Popular (LP apps). From 65% of the analyzed HP apps, unique device information can tie users to a specific mental disorder. *Polisis* analysis found that 9 out of 10 apps are collecting Device IDs and IP Addresses, which can pose a potential threat to users' privacy and security. Dynamic analysis reveals that 17.67% of all 20 HP apps permissions are categorized as dangerous by the Android framework. 11 out of 20 HP apps use one or more dangerous permissions. In static analysis, 45% of 20 HP MMH apps show *Hardware Id Usage*, a unique permanent device Id, and thus poses a privacy risk. 20% of apps show *Insecure TLS/SSL Trust Manager*, which may cause insecure network traffic to leak sensitive data.

To the best of our knowledge, there is no existing work with a comprehensive analysis for privacy and security concerns specific to MMH apps targeting vulnerable populations with a higher possibility of sharing personal and sensitive mental health information.

The following section presents the background of the issue and the need for the study. Then, we discuss the related work done in the field, followed by our study design and goals. After that, we present the methodology and data collection, results, and challenges. We conclude the study findings with a discussion and future directions.

## 5.2 Background

Mental illnesses impact one in five United States (U.S.) adults with having anxiety disorders as the most common mental illness. Every year, around 40 million U.S. adults of age 18 and older are affected by anxiety disorders. Anxiety disorders affect 25.1% of U.S. teenagers. Even though anxiety disorders are considered highly treatable, treatment covering only 36.9% indicates a substantial gap [175]. A study of depression data from 1990 to 2017 by Liu et al. found an overall increase in depression globally and staying as a global public health issue [176]. Studies have reported that in the U.S., nearly $193 billion in lost earnings each year is attributed to mental health disorders [177]. Overburdened and overwhelmed providers and patients are not able to afford the services because of limited insurance participation by providers and unclear or no coverage by insurance providers [178, 168].

### 5.2.1 MMH Apps as Mental Health Intervention

MMH apps can assist large, underserved populations with a broad spectrum of mental disorders, from a simple reminder to take medicine on time to sending automatic signals outside regarding a predictable near-future crisis [179]. The MMH app is convenient, low-cost, anonymous, free from human biases, and provides round-the-clock assistance to reach out to people in need. Interactive or game-based app therapies can invoke users' interest, especially in a younger generation, and can help them to follow through [12]. MMH app can be a great supporting tool to conventional in-person therapy sessions by providing supplement data [13].

A large pool of MMH apps is available through iOS and Android app stores. If app stores can identify the scientific research-based MMH apps for users, It may add confidence to use the app [180]. User ratings are helpful but not necessarily valid for clinical usefulness backed up by research. Online resources like *PsyberGuide* may help provide objective and actionable information for publicly available MMH apps [172].

The privacy and security concerns are potential factors to hinder the widespread acceptability, and use of MMH apps [181]. A real-life patient's perspective suggests that apps can be utilized in psychiatric treatment with flexible use of apps rather than relying on a single condition-specific MMH app [182].

### 5.2.2 MMH Apps Security and Privacy Issues

Users' privacy and security is significant concern regarding MMH apps. As per O'Loughlin et al., 68% of apps received unacceptable transparency scores on data security, and privacy policies of studied mobile apps for depression [30]. A lack of industry regulations and standards makes it very difficult to evaluate these apps on security, privacy, and effectiveness. Non-standard development practices may produce an insecure, easily exploitable app that malicious attackers can target to breach. An absence of a privacy policy in almost half of the studied apps shows a significant concern of not informing users about data collected, data security, and data sharing [183].

Leaked information can harm users financially, mentally, physically, or emotionally, even life-threatening to already overwhelmed users who do not know how to handle any additional mental burden. Most of these policies use legal "boilerplate" without saying much about what the app does precisely and what is involved in data collection, sharing, or app permissions. The typical average user is not even familiar with legal or technical jargon and has little or no ability to understand to be agreed on the details given.

In the absence of HIPAA protection, the app company may collect and share healthcare-related data for usage that the patient never imagined. Although the proliferation of mental health technologies like condition-specific smartphone apps and wearable sensors continues to increase, evidence for their clinical utility, efficacy, and safety is generally lacking. All MMH apps need proper identification if falling under compliance or outside the scope. At present, data brokers may end up indefinitely owning the patient's data and using it for a variety of purposes like the generation of FICO Medication Scores, targeted advertisements, or more considerable profiling efforts [184]. Based on the highly sensitive nature and broad scope of usage, MMH apps need more rigorous and continuous evaluation to ensure the privacy and security of these apps' users.

## 5.3 Related Work

Mobile app security and privacy issue are well-acknowledged by healthcare, legal authorities, and researchers. A 2017 study in rural India shows that privacy policy complexity may be a barrier to informed decision making [185]. Reardon et al. presented how mobile apps can gain access to protected data without user consent by using both covert and side channels through exploiting the Android permission model and posing a threat to users' privacy and security [32]. Device ID is reported only for 0.2% of tested apps (172/88113) by [32], but our work shows that 45% of top-rated MMH apps use Hardware ID, which indicates a higher risk for the target population. We present all observed dangerous permissions for the tested apps and their possible misuse, while Reardon et al. analyzed only a few dangerous permissions uncovering covert and side channels. Au et al. found a fundamental trade-off between the stability of the permission specification and enforcing least-privilege with fine-grain permissions, which may impact developers' choices of designing MMH apps and allowing unnecessary permissions bundled with the necessary ones [186].

Giota et al. describe how using mental health apps can be risky compared to consulting the therapist daily, including loss of data and theft through insecure devices and communication channels [180]. Robillard et al. focus on privacy policies and terms of agreements covering 319 MMH apps for both Android and iOS platforms. Their work shows that only 18% of iOS and 4% of Android apps have privacy policies with collecting user information by 92% of studied privacy policies [187]. O'Loughlin et al. highlight the absence of privacy policy, consent, transparency, data sharing, and difficulty of readability for the general population [30]. This work emphasizes the absence of privacy policy in nearly half of 116 studied apps and the absence of not covering what personal information is being collected [30]. Parker et al. conducted an empirical study for 61 apps, identifying that malicious apps or attackers can exploit 'dangerous' app permissions to access sensitive information [183].

Significant recommendations are about the transparency of privacy policies on data sharing, allowing users to opt-out from data collection, improved user interface, clinical trials, rigorous evaluation, and integration with EHR [188]. Parker et al. studied the MMH apps' privacy policies aligned with the local government policy. Interestingly, the suggested solution is for users to pay more attention to selecting safe and efficient apps [189]. Papageorgiou et al. presented that manual, static and dynamic analysis provides better insight into the apps' state regarding privacy and security vulnerabilities. The significance of this work is more practical to show that developers improved the apps after being informed of the issues [190]. Another study on both Android and iOS apps shows that the majority of apps (95.63%) pose some potential damage because of security and privacy violations [31].

MMH App as an alternative health intervention is a topic of great interest, but mostly it is limited to manual studies, identifying concerns, and suggesting general improvements. Papageorgiou et al. analyzed the security and privacy of mobile health apps covering Manual, Dynamic, Static, and Traffic analysis [190], which is closest to our work but has different methods and analysis. For example, their work examines app permissions through manual analysis, but we performed manual and dynamic analysis to analyze app permissions. Similarly, their dynamic analysis evaluated apps based on data transmission observation over the internet. However, we analyze any exploitable app permissions, attack surface, injection, and other vulnerabilities on Android devices at runtime. Our static analysis uncovers Hardware ID Usage and other vulnerabilities different from their analysis. We also analyzed privacy policies through DL-based tool *Polisis* in addition to a manual analysis.

## 5.4 Study Goals and Design

We aim to evaluate Android MMH Apps to keep their security and privacy practices transparent to the users. We also look at the vulnerabilities associated with these apps, leading to security and privacy exploitation. App permissions are studied to understand the necessity of these permissions for app functionality and users' options to opt out.

### 5.4.1 Study Goals

The study presents a comprehensive picture of security and privacy practices, issues, and vulnerabilities to exploit by studying privacy policies, terms and conditions, static analysis, and dynamic analysis of MMH Apps. We outline our higher-level study goals as follows:

- Manual and *Polisis* Analysis: Identify the deficiencies in privacy policies and challenges for users and the developers.
- Dynamic Analysis: Identifying the threat surface and runtime vulnerabilities.
- Static Analysis: Identifying the security and privacy threats in apps' code.
- Traffic Payload Analysis: Identifying the insecure data transactions.

We comprehensively study privacy and security issues in MMH apps by combining observations from manual app usage from a regular user's perspective, privacy policy analysis, and dynamic and static analysis. We look at the exploitable app permissions levels, injection vulnerabilities, security, and privacy settings through static and dynamic analysis.

### 5.4.2 MMH App Inclusion Criteria

As many MMH Apps are available, we study 40 MMH Apps fulfilling the following inclusion criteria. We searched online for the available mental health apps in the Android/Google Play Store. All selected MMH apps are in the English language and free to download. The primary criteria for inclusion is based on the app ratings, the number of downloads, and the number of reviewers.

We grouped apps into Highly Popular (HP) and Less Popular (LP) based on their ranking and number of downloads following inclusion criteria. Table 5.1 shows 20 HP apps in the range of ranking from 3.3 to 4.8, and the number of downloads varies from 100,000 to 10,000,000. For LP Apps, Table 5.2 shows that ratings are in the range of 1.0 to 3.2, and downloads vary from 1,000 to 100,000. Higher ratings and number of downloads were combined with a higher number of reviewers. Minimum number of reviewers for HP apps are 575, while the lowest LP app, with a rating of 1, has only 3 reviewers.

## 5.5 Data Analysis and Results

We summarize our findings under manual analysis, privacy policies through web-based DL tool *Polisis*, dynamic analysis, static analysis, and traffic analysis to cover privacy and security

Table 5.1: Twenty Highly-Popular (HP) MMH Apps with Category, Downloads, Rating, and Rated By

| Apps | Categories | Downloads(+) | Ratings | Rated By |
|---|---|---:|---:|---:|
| 7 Cups | Depression | 1,000,000 | 4.3 | 18,215 |
| Anxiety Relief Hypnosis | Anxiety/ stress | 100,000 | 4.2 | 1,375 |
| BetterHelp | Anxiety/ stress | 500,000 | 4.5 | 10,290 |
| Breathe2Relax | PTSD | 100,000 | 3.3 | 1,082 |
| Calm | Mindfulness/ Meditation | 10,000,000 | 4.4 | 271,581 |
| CBT Thought Record Diary | Anxiety/ stress | 100,000 | 4.7 | 1,431 |
| eMoods | Bipolar Disorder | 100,000 | 4.6 | 4,133 |
| Happify | Depression | 1,000,000 | 4 | 2,408 |
| Headspace | Mindfulness/ Meditation | 10,000,000 | 3.5 | 133,877 |
| MindDoc: Mood Tracker for Depression & Anxiety | Depression | 1,000,000 | 4.5 | 35,016 |
| MindShift | Anxiety/ stress | 100,000 | 4.1 | 1,227 |
| MoodSpace - Stress, anxiety | Anxiety/ stress | 100,000 | 4.7 | 2,915 |
| MoodTools | Depression | 100,000 | 4.3 | 3,108 |
| PTSD Coach | PTSD | 100,000 | 4.6 | 575 |
| Sanvello | Anxiety/ stress | 1,000,000 | 4.6 | 17,005 |
| Self-Help for Anxiety Management (SAM) | Anxiety and stress | 500,000 | 3.9 | 2,957 |
| Super Better | PTSD | 100,000 | 4.4 | 5,999 |
| Ten Percent Happier | Mindfulness/ Meditation | 500,000 | 4.8 | 10,541 |
| What’s Up | Suicide Prevention | 500,000 | 4 | 3,221 |
| Wysa: stress, depression | Anxiety/ stress | 1,000,000 | 4.7 | 51,746 |

Table 5.2: Twenty Less-Popular (LP) MMH Apps with Downloads, Rating, and Rated By

| **Apps** | **Categories** | **Downloads(+)** | **Ratings** | **Rated By** |
|---|---|---:|---:|---:|
| Anger Management & stress relief game (pstd) | Anxiety/ stress | 100,000 | 2.9 | 592 |
| Bipolar Test | Bipolar Disorder | 10,000 | 3.1 | 41 |
| Brain Manager by UPMC | Depression | 1,000 | 2.1 | 18 |
| Course of Cognitive Behavioral Therapy | Depression | 5,000 | 3.0 | 33 |
| Daylight - Worry Less | Anxiety/ stress | 10,000 | 2.5 | 40 |
| Depression | Depression | 1,000 | 1.0 | 3 |
| Depression: The Game | Depression | 10,000 | 3.1 | 418 |
| EAP In Your Pocket | Anxiety/ stress | 5,000 | 2.3 | 16 |
| iPrevail: Anxiety & Depression | Anxiety/ stress | 10,000 | 2.9 | 130 |
| Mental Health - psychologist | Depression | 5,000 | 3.0 | 10 |
| NarcStop - Narcissistic abuse and recovery guide | Depression | 1,000 | 2.7 | 14 |
| PTSD Aid | Depression | 1,000 | 2.8 | 14 |
| R U Suicidal? | Suicide Prevention | 1,000 | 2.6 | 33 |
| Real Antistress Stress Relief: Relaxing games | Anxiety/ stress | 100,000 | 3.2 | 1080 |
| SafetyNet: Your Suicide Prevention App | Suicide Prevention | 1,000 | 2.8 | 11 |
| Stress relief ducky: antidepressant & anti anxiety | Anxiety/ stress | 1,000 | 3.0 | 10 |
| Suicide Prevention | Suicide Prevention | 1,000 | 3.2 | 13 |
| Talkspace Counseling & Therapy | Depression | 100,000 | 2.1 | 2774 |
| VA Health Chat | PTSD | 10,000 | 2.7 | 73 |
| Waver - Meet others with same Mental Health Issues | Depression | 5,000 | 3.1 | 172 |

issues exhibited by studied MMH apps. We study 40 MMH apps with varying ratings and downloads classified under six categories. *Anxiety and Stress* (14) and *Depression* (13) are two major categories. *Suicide Prevention* and *PTSD* each have 4 apps, 3 *Mindfulness and Meditation* apps, and 2 apps for *Bipolar Disorder* are studied.

### 5.5.1 Manual Analysis

We document the availability and accessibility of the Privacy Policy for the studied MMH apps. We also compare our HP (20) and LP (20) app groups to see any identifiable differences. Based on ratings and the number of downloads, results show that 95% of HP apps have a defined PP compared to 65% of LP apps. From the privacy policies of 20 HP apps, we have the following observations:

#### *5.5.1.1 Privacy Policy*

Sixty-five percent of 20 HP apps collect device information, while 40% of apps collect the most common data points like name, email, phone, and address. 55% of apps collect email addresses. A declaration to collect location or IP address is stated only by 15% of apps. 5% of apps are collecting medication information as well as counseling sessions.

On personal data security and privacy, only 30% apps maintain de-identified data, 10% apps declare not de-identifying data, and 60% apps do not provide information on de-identifying data. Most HP apps share data with third-party providers, and 20% of apps declare to share data with sales and legal teams. only 10% of apps share data with therapists or healthcare providers. 60% of apps store data with the app provider, 30% of apps store on the user's device, and 5% of apps store data on the cloud. 50% of apps allow users to opt-out, 15% apps do not allow it, and 35% of apps do not provide any information on the data sharing opt-out option. 40% of apps offer security through account ID and password, while 20% of apps provide encryption and firewalls. However, few apps mention providing secure access with the paid version of the app. One app mentions *commercial means security*, but no further detail is provided.

Few app-specific privacy policies indicate threats to the privacy and security of the app user.

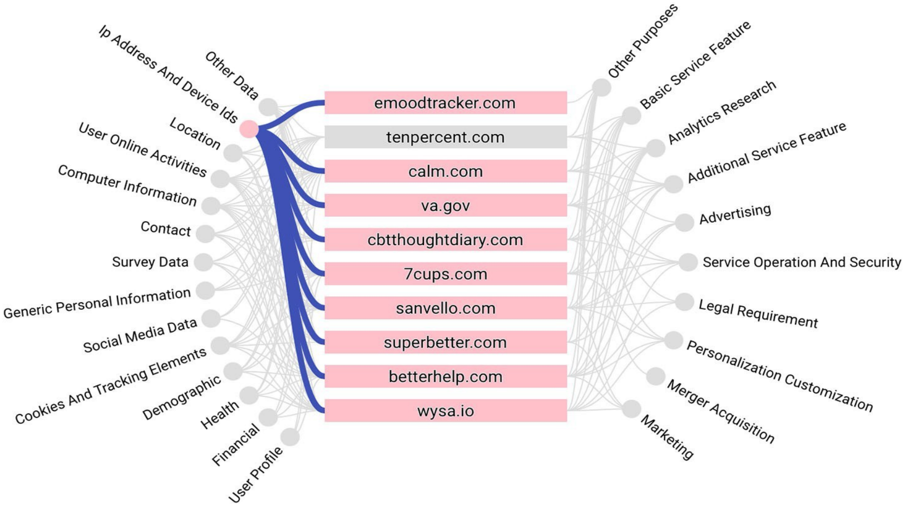


Figure 5.1: Polisis [3] Analysis- Comparison of HP MMH Apps Privacy Policies

For example, *7 Cups* says, “While we generally do not monitor transcripts of chats between users and Listeners and Therapists, we may occasionally review the chat transcripts to conduct quality control, address potential safety issues, and prevent misuse of our platform, if certain suspicious or potentially harmful activity is detected.” App also says that the ‘Do Not Track’ feature is not supported, so users cannot opt out. App *Calm* also suggests that the app will retain some information as required or permitted by law for a certain period even after the user requests to cancel or delete the account.

### 5.5.2 Privacy Policy Analysis using Polisis

Polisis allows up to 10 privacy policies to compare, and for 10 HP MMH apps’ privacy policies, we found that 65% of the applications do not discuss Secure Data storage, Privacy Security Program, and Security Data program. 60% of the applications have not mentioned data access limitations and secure user authentication. Also, 40% of the applications are unclear about their security measure, and 20% of them use many generic statements, which are confusing for the readers. Fig. 5.1 demonstrates the Polisis comparison of 10 MMH apps’ privacy policies on key data collection (Left side) and data usage (Right side) categories. For example, 9 out of 10 HP MMH apps are collecting Device IDs and IP Addresses which can be exploited to violate users’ privacy and security. Only 5 out of 10 apps use collected information for service operations and security. All of the apps have a date category as *Other Data*, and all the apps may use the collected information for *Other Purposes*, which hinders the transparency about what is being collected and what it can be used for. Fig. 5.2 represents the details of a specific app (*Betterhelp*) regarding information collection (left side) and how the collected information can be utilized (right side). We can see that 'Other Data' goes to many different usages, including 'Other Purpose', making it difficult for users to understand the associated risks. We observe that 9 out of 10 compared apps pay attention to children's privacy as they represent a more vulnerable population. However, 8 out of 10 do not talk about sharing personal information with third parties, and two apps show concern with warning signs, as per Appendix Figure 5.1, stating that "Several types of personal information types are shared with third parties."

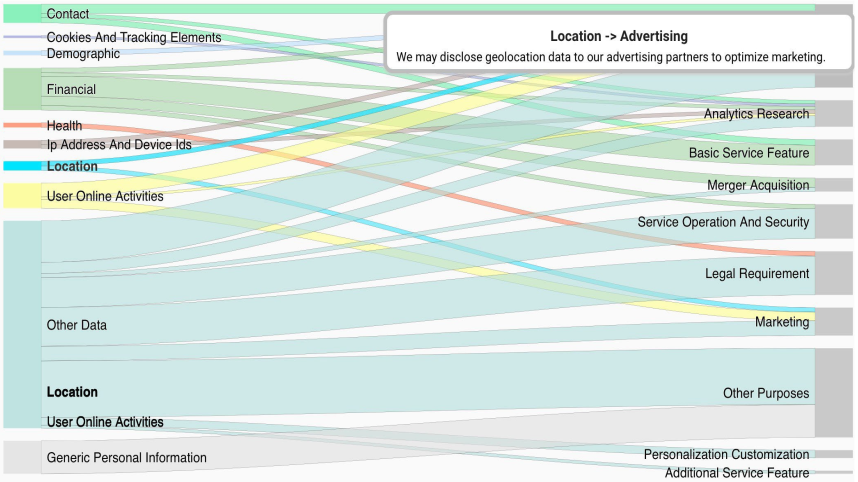


Figure 5.2: Polisis Analysis of MMH App *BetterHelp*- Data Collection and Data Usage

### 5.5.3 Static Analysis

We performed a Static Analysis on HP MMH apps studied in our work. In addition, we study the security issues and vulnerabilities existing in code analysis that may be exploited in security and privacy attacks.

Table 5.3: Static Analysis: Major Security and Privacy Vulnerabilities in MMH Apps

| App | Hardware ID Usage | Insecure TLS/SSL | Potential Multiple Certificate Exploit | File Readable |
|---|---|---|---|---|
| 7 Cups | √ | √ | √ | √ |
| Anxiety Relief Hypnosis | ✗ | ✗ | ✗ | √ |
| BetterHelp | ✗ | ✗ | ✗ | ✗ |
| Breathe2Relax | √ | ✗ | √ | ✗ |
| Calm | ✗ | √ | ✗ | ✗ |
| CBT Thought Record Diary | √ | ✗ | √ | ✗ |
| eMoods | √ | ✗ | √ | √ |
| Happify | √ | √ | √ | √ |
| Headspace | √ | ✗ | √ | √ |
| MindDoc | ✗ | ✗ | ✗ | ✗ |
| MindShift | ✗ | ✗ | ✗ | ✗ |
| MoodSpace - Stress, anxiety | ✗ | ✗ | ✗ | ✗ |
| MoodTools | √ | ✗ | √ | ✗ |
| PTSD Coach | ✗ | ✗ | ✗ | ✗ |
| Sanvello | ✗ | ✗ | ✗ | ✗ |
| Self-Help for Anxiety Management (SAM) | √ | ✗ | √ | ✗ |
| Super Better | √ | √ | √ | √ |
| Ten Percent Happier | ✗ | ✗ | ✗ | √ |
| What’s Up | ✗ | ✗ | ✗ | ✗ |
| Wysa: stress, depression | ✗ | ✗ | √ | ✗ |

Broadly, as categorized in Android Studio 3.6.1, we observed errors and warnings for five categories of *Accessibility*, *Correctness*, *Performance*, *Usability*, and *Security*. We further focused on identifying major security flaws. Table 5.3 shows four studied categories of security and privacy features that can potentially be exploited. Thirteen out of twenty HP MMH apps are vulnerable to at least one category, and three apps show vulnerability to all four categories. Our results show that 45% of HP MMH apps show *Hardware Id Usage*, a unique permanent ID (Android hardware ID can get reset with a factory reset option) for the device and thus can be associated with a particular user. User Hardware ID can be accessible to all apps installed on the device as per the Android framework scope, which increases the risk of privacy violation [191].

50% of apps indicate *Potential multiple certificate exploits*, showing that App signatures can be exploited if not validated properly. 30% of apps can read the files through *File.setReadable()*, which is used to make file word-readable. 20% apps show *Insecure TLS/SSL Trust Manager*, which can cause insecure network traffic disclosing sensitive information. TrustManager can implement custom certificate validation strategies, and an insecure trust manager implementation makes an application vulnerable to Man-In-The-Middle attacks. Malicious entities may intercept an app's data over the network with insecure TLS/SSL and can violate the users' privacy and security [191].

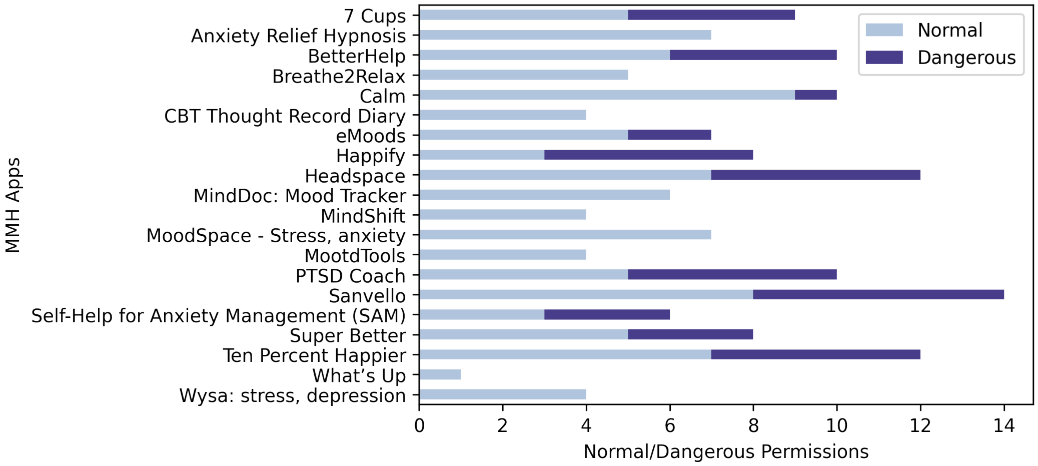


Figure 5.3: Dynamic Analysis- Permissions Used by HP MMH Apps

### 5.5.4 Dynamic Analysis

Dynamic analysis is conducted on twenty HP apps to analyze the app permissions, attack surface, and exploitable vulnerabilities on Android devices at runtime.

#### *5.5.4.1 Permissions*

App permissions analysis can identify the potential risk of violating users' privacy and security. Standard permissions allow access to data and actions which present minimal risk to the user's privacy. Runtime permissions allow additional access to restricted data and may significantly perform

Table 5.4: Dynamic Analysis: Dangerous Permissions

| App Permission (Category- Dangerous) | HP Apps |
|---|---:|
| android.permission.WRITE_EXTERNAL_STORAGE | 9 |
| android.permission.READ_CONTACTS | 3 |
| android.permission.READ_EXTERNAL_STORAGE | 9 |
| android.permissions.RECORD_AUDIO | 4 |
| android.permission.READ_CALENDAR | 1 |
| android.permission.WRITE_CALENDAR | 1 |
| android.permission.READ_PHONE_STATE | 5 |
| android.permission.GET_ACCOUNTS | 2 |
| android.permission.CAMERA | 4 |
| android.permission.ACCESS_FINE_LOCATION | 2 |
| android.permission.ACCESS_COARSE_LOCATION | 2 |
| android.permission.BODY_SENSORS | 1 |

restricted actions that impact the system and other apps. For example, runtime permissions may access private user data, and sensitive information like user's location and contact information.

Drozer identified that more than 50% of HP apps have 1 to 6 permissions per app defined as *Dangerous* according to Android developer guidelines [191]. There are 12 different types of dangerous permissions identified in studied apps. From 148 standard and dangerous permissions, 29.05% of permissions fall under the dangerous category. For example, permission to read contacts threatens privacy and security as a malicious app can use this data without the user's knowledge. Based on the permissions count of all twenty HP apps permissions, 18.38% of apps are listed as dangerous, while 44.87% of apps fall under standard permissions. The rest of the permissions consist of obsolete and other permissions. Fig. 5.3 shows the distribution of app permissions, while Table 5.4 shows the specific dangerous permissions used by studies apps. These highly popular MMH Apps do carry dangerous permissions, which can be exploited to evade patients' privacy. Therefore, it is critical to define who may access such data and its maintenance.

We also manually observed all apps installed on Android phone and requested permissions. Table 5.5 presents eight major permissions.

Table 5.5: Manual Analysis: Dangerous Permissions Requested for 20 HP MMH Apps

| **App** | **Camera** | **Location** | **Body sensors** | **Mic** | **Contacts** | **Calendar** | **Phone** | **Storage** |
|---|---|---|---|---|---|---|---|---|
| 7 Cups | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Anxiety Relief Hypnosis | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| BetterHelp | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ |
| Breathe2Relax | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Calm | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| CBT Thought Diary | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| eMoods | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Happify | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ |
| Headspace | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ |
| MindDoc | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| MindShift | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |  | ✗ |
| MoodSpace | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| MootdTools | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| PTSD Coach | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ |
| Sam | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| Sanvello | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ |
| SuperBetter | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| Ten percent happier | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| What’s up | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| Wysa | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Total= | 4 | 2 | 1 | 5 | 5 | 1 | 3 | 10 |

Results show that 12 out of 20 HP MMH apps carry 31 dangerous permissions. Storage Access permission is most common, with ten apps requesting it, while body sensors and calendar access are the least requested (1 app each) permission. Commonly requested permissions are for Camera (4), Microphone (5), and Contacts (5). Other requested permissions are Phone (3) and Location (2). These findings are per dynamic analysis identified dangerous permissions. There are two differences observed between manual and dynamic analysis. First is that app *BetterHelp* does not show read or write storage permissions in Drozer output, but it is requesting ‘Storage’ permission on an Android phone. Second is the app *Wysa* which does not show ‘RECORD_AUDIO’ permission in Drozer output, but on the phone, ‘Microphone’ permission is requested.

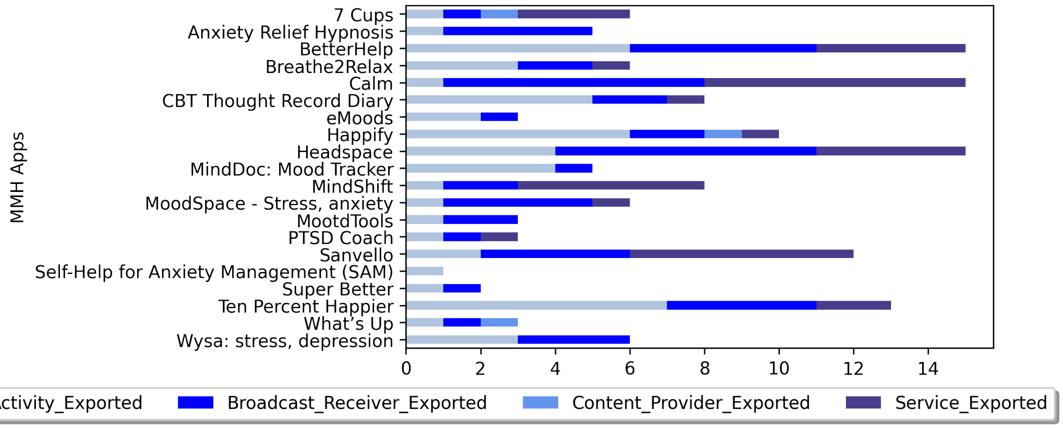


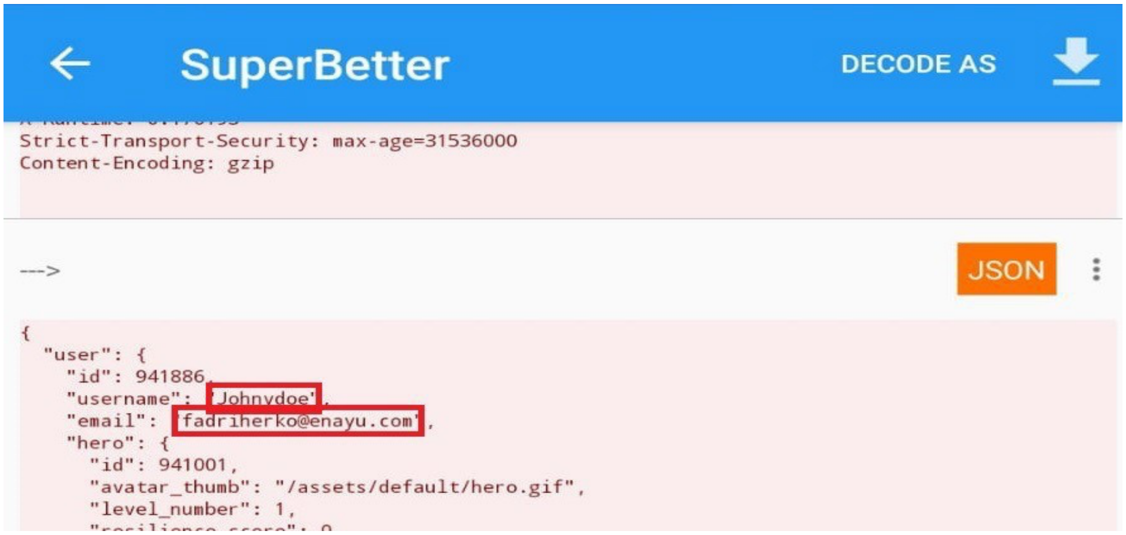
Figure 5.4: Dynamic Analysis - HP MMH Apps Attack Surface

#### *5.5.4.2 Attack Surface*

As per Drozer output, all HP MMH apps show exposure as an attack surface. Attack surface is categorized as 'Activities Exported', 'Broadcast Receiver Exported', 'Content Providers Exported', and 'Service Exported'.

Fig. 5.4 shows how each HP app has entry/exit points as attack surfaces under these four categories that can be exploited by a malicious user or a malicious app. There are a total of 52 activities exported for 20 studied apps. Another query in Drozer indicates if any permission is required or not-required to perform some app activity. If no permission is required to export some activity, any application will launch the activity, allowing a malicious application to gain access to sensitive information. In the context of a malicious application, it is possible to modify the internal state of the application or to deceive the user into interacting with the victim application while believing that they are still interacting with the malicious application. [192].

As per our results, there are 36 services exported, and three content providers are exported. If access to an exported Service is not restricted, any application may start and bind to the Service. The exposed functionality may allow a malicious application to perform unauthorized actions like gaining access to sensitive information or corrupting the application's internal state. If access to a Content Provider is not restricted to only the expected applications, then malicious applications



```
Strict-Transport-Security: max-age=31536000
Content-Encoding: gzip
```

```
--->
```

```
{
  "user": {
    "id": 941886,
    "username": 'Johnydoe',
    "email": 'fadriherko@enayu.com',
    "hero": {
      "id": 941001,
      "avatar_thumb": "/assets/default/hero.gif",
      "level_number": 1,
      "resilience_score": 0
```

Figure 5.5: *SuperBetter*: Traffic captured through 'pcap' shows username and email in plaintext

might be able to access the sensitive data [192].

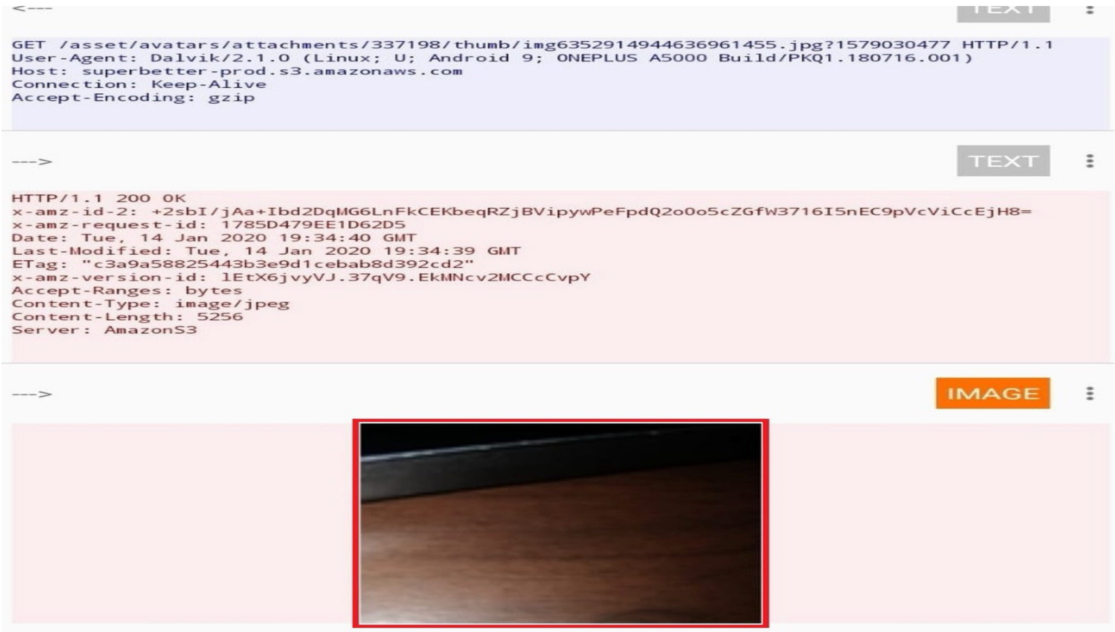
Our findings also show 54 *Broadcast Receiver Exported*, which is another exploitable vulnerability. Additionally, a malicious application can send an explicit intent to the victim app, which may assume that any received intent is valid, which may cause unintended app behavior.

### *5.5.4.3 Injection and Other Vulnerabilities*

We scanned all 20 HP apps for injection vulnerabilities, but Drozer did not identify any app as vulnerable. However, if the content provider is exported and *grantUriPermissions* is set to true, that can make an app vulnerable to injection attacks. Drozer indicated three such apps where we have null permissions and content provider is exported, but all three apps have *grantUriPermissions* as false. We scanned 20 HP MMH apps to find any vulnerable provider, but Drozer did not identify any vulnerable provider for any app. Furthermore, Drozer did not identify any vulnerable URI paths for the 20 HP MMH apps.

### 5.5.5 Traffic Payload Analysis

We ran each application separately and used the packet capturing tool *pcap* and monitored the packets. However, currently, we could only find defects in one application, i.e., *Superbetter*.



```
<---
GET /asset/avatars/attachments/337198/thumb/img6352914944636961455.jpg?1579030477 HTTP/1.1
User-Agent: Dalvik/2.1.0 (Linux; U; Android 9; ONEPLUS A5000 Build/PKQ1.180716.001)
Host: superbetter-prod.s3.amazonaws.com
Connection: Keep-Alive
Accept-Encoding: gzip
```

```
--->
HTTP/1.1 200 OK
x-amz-id-2: +2sbI/jAa+Ibd2DqMG6LnFkCEKbeqRZjBVipywPeFpdQ2o0o5cZGfW3716I5nEC9pVcViCcEjH8=
x-amz-request-id: 1785D479EE1D62D5
Date: Tue, 14 Jan 2020 19:34:40 GMT
Last-Modified: Tue, 14 Jan 2020 19:34:39 GMT
ETag: "c3a9a58825443b3e9d1cebab8d392cd2"
x-amz-version-id: lEtX6jvyVJ.37qV9.EkMNcv2MCCcCvpY
Accept-Ranges: bytes
Content-Type: image/jpeg
Content-Length: 5256
Server: AmazonS3
```

```
--->
```

Figure 5.6: *SuperBetter*: Traffic captured through pcap shows Unencrypted Image

Superbetter is an application for teens dealing with stress, anxiety, depression, or other mental challenges. Though we could find only one app to show this behavior, it does reflect the discrepancies between the app's privacy policy and app behavior. As per the privacy policy of Superbetter, data is stored and encrypted, but we could capture the plain text in the pcap file as shown in Fig. 5.5. Fig. 5.6 shows the unencrypted image in the captured traffic. Our experiments show that an MMH app transmits text data and images over the network without any encryption. However, information transferred over the network unprotected can be intercepted and threaten users' privacy and security.

### 5.5.6 Summary of Results

We can summarize our results as follows:

- Manual Analysis: 20 HP MMH apps show that 95% of HP apps provide a privacy policy compared to 65% of LP apps. 65% of apps collect device information, while 40% of apps collect the most common data points like name, email, phone, and address. 5% of apps are collecting medication information and counseling sessions. Polisis analysis demonstrates that apps collect identifiable information, share it with third parties, and retain some data indefinitely. 12 out of 20 HP MMH request ‘dangerous’ permissions on Android phones.

- Dynamic Analysis: 60% of 20 HP MMH apps use 12 dangerous permissions. Attackers and malicious apps may exploit identified 145 vulnerabilities to access sensitive data.

- Static Analysis: 45% of MMH apps use a unique identifier, *Hardware Id*, which can connect particular users to their mental health issues, thus violating privacy and security. 20% of apps indicate *Insecure TLS/SSL Trust Manager*, which may leak sensitive data through insecure network traffic.

- Traffic Payload Analysis: Captured traffic unveils the unencrypted text and image in one app while the app’s privacy policy indicates data encryption.

## 5.6 Challenges and Limitations

In the era of ever-changing information at a fast pace, MMH apps are also subject to rapid changes in usage, ratings, downloads, and reviews from the user population. Reading privacy policies is a time-consuming process but more difficult to analyze and compare without a standard format followed by the app developers. Though tools like *Polisis* can help to extract usable information automatically from the Privacy Policy, inconsistent content covered in privacy policies may limit the comparison analysis. This study is also limited to the Android platform only. Frequent changes in app usage and availability forced us to recollect some data.

It is also worth mentioning that we were limited to observing the complete app functionality in unpaid versions. A complete analysis of all the app’s features, especially regarding audio, video, call, and message capabilities, can provide more details on how some app utilizes or exploits app capabilities and permissions.

## 5.7 Conclusion

Studied MMH Policies do not indicate app permissions or the necessity of permissions for intended app functionalities. Not all policies mention collecting location data, but it can be collected indirectly from the device information, cookies, and web beacons. Data sharing with many other parties is inevitable as different parties maintain the app and store data in the cloud. As a result, the app can get geolocation that is not transparent to the user, posing a privacy violation. Users may have no control over some in-app settings to opt out from specific data sharing as suggested in their privacy policies. Users may want to discontinue using the app and delete the account(s), but data can be retained for a more extended period as covered by different clauses.

Results show that manual, static, dynamic, and traffic analysis identify exploitable vulnerabilities in MMH apps, posing privacy and security threats. More transparent privacy policies and standard secure development guidelines are needed to ensure that users and providers trust the apps. Otherwise, MMH apps can be intrusive to users' privacy and security, negatively affecting their intended purpose of providing alternative interventions to improve their mental health.

# 6. NAVIGATING PRIVACY AND SECURITY THREATS TO WOMEN'S HEALTH APPS*

Female Technology (Femtech) has emerged as a growing phenomenon in the global mobile app market, offering the potential to manage women's health digitally and support safe and informed family planning. However, the privacy and security vulnerabilities of period-tracking and fertility-monitoring apps pose significant risks to women's health, potentially leading to unintended pregnancies and legal prosecution.

Our analysis of period-tracking and fertility-monitoring mobile apps employed a multilayered approach that included manual observations of privacy policies and app permissions, as well as dynamic and static analysis with multiple evaluating frameworks. Our research findings indicate that a significant number of these apps collect personally identifiable information (PII) and sensitive healthcare data. Additionally, 75% of the studied apps allow for geographical tracking, which poses significant privacy risks to users. Despite the sensitive nature of health data, our analysis shows that around 85% of app privacy policies did not mention any security measures. Moreover, our study identified that 61% of the code vulnerabilities present in the apps belong to the top-10 Open Web Application Security Project (OWASP) vulnerabilities. These findings highlight the urgent need for stricter regulations and increased transparency for reproductive health apps.

Overall, our research underscores the importance of addressing the privacy and security vulnerabilities of period-tracking and fertility-monitoring mobile apps. By raising awareness of these critical privacy and security risks, we aim to spark a much-needed conversation and promote greater accountability and transparency in the development and use of digital tools for women's health and encourage the industry to prioritize user privacy and security.

## 6.1 Introduction

Digital reproductive health management meets modern medicine needs for better accessibility and continuous monitoring. Female Technology (Femtech), including period-tracking and fertility

*Submitted work, Under Review-2023

monitoring apps, addresses women's unique health needs [193]. The femtech market has grown tenfold in the last decade and is expected to reach $50 billion by 2025 [194].

Smartphone accessibility has changed healthcare by enabling remote recording, sharing, and access to medical assistance, especially during the COVID-19 pandemic. In North America, 88% of the population (329 million mobile users) and 75% of networked devices/connections will be connected over Wi-Fi [195]. Mobile technologies have the potential to provide sexual and reproductive health education, assist vulnerable populations, and optimize reproductive health through fertility tracking apps, contraceptive counseling, and birth control reminders [196, 197, 198]. However, reproductive health apps may leverage users' needs and interests to encourage them to share sensitive data and personal identifiable information (PII), which can be exploited for surveillance capitalism [199, 200].

With Roe v. Wade overturned, period-tracking and pregnancy-managing mobile apps may be vulnerable to exploitation for tracking and identifying vulnerable populations for legal prosecution [201, 202]. Though the U.S. Department of Health and Human Services (HHS) issued guidelines to protect users of reproductive healthcare mobile apps, apt regulatory standards are needed to ensure their protection. The Federal Trade Commission (FTC) protects sensitive health data in all cases, even if not regulated by The Health Insurance Portability and Accountability Act (HIPAA) [203, 204]. The recent case of a period-tracking app being barred by the Federal Trade Commission (FTC) for selling personal data to advertisers is just one example of the exploitation of unregulated reproductive healthcare apps [33].

Mobile apps face challenges in validating integrity, reliability, effectiveness, privacy, security, and accountability, unlike regulated medical devices. The FDA approved *Natural Cycles* in 2018 as the first mobile application for contraception [6]. However, it has a typical use failure rate of up to 8.3% [205]. If an app fails to prevent pregnancy, it can result in lawsuits and financial claims [206]. Without strict regulation and compliance, app developers may exploit the freedom to maximize data monetization, even if it may jeopardize users' privacy and security. Certain default settings, bundled permissions, and data sharing can lead to such exploitation [207].

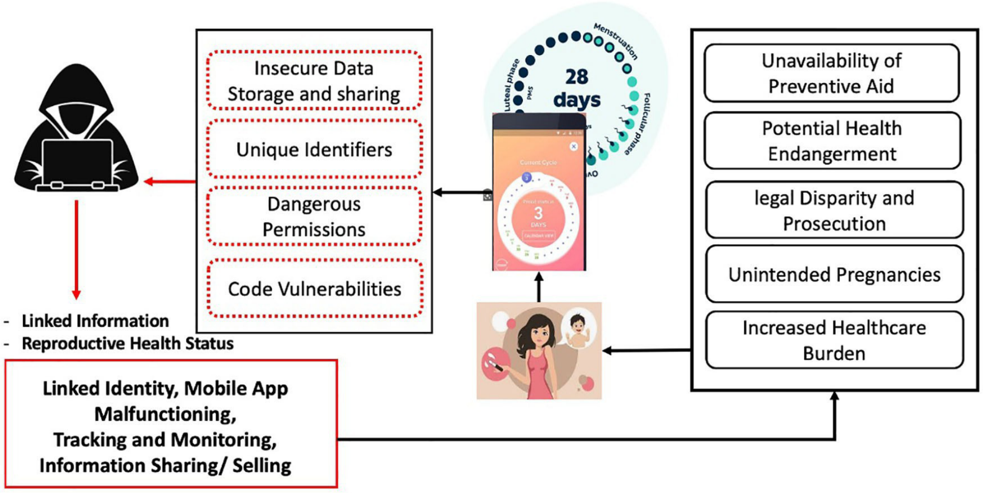

Insecure Data Storage and sharing
Unique Identifiers
Dangerous Permissions
Code Vulnerabilities
- Linked Information
- Reproductive Health Status
Linked Identity, Mobile App Malfunctioning, Tracking and Monitoring, Information Sharing/ Selling
28 days
Menstruation
Follicular phase
Luteal phase
PMS
3
Unavailability of Preventive Aid
Potential Health Endangerment
legal Disparity and Prosecution
Unintended Pregnancies
Increased Healthcare Burden

Figure 6.1: Privacy and Security Threats of Vulnerabilities of Reproductive Mobile Apps to Women's Health

Figure 6.1 illustrates the potential threats and negative impacts of exploiting vulnerabilities in reproductive health apps, highlighting the harm that can result from malicious manipulation of a useful tool. Our work demonstrates that studied period-tracking and fertility-monitoring mobile apps have many exploitable vulnerabilities posing a significant threat to users' privacy and security. Our contribution is manifold through this study as follows:

- Our work verifies that majority of studied reproductive health apps collect Personal Identification Information (PII), geographical data, and sensitive health data, which poses the potential threat of women being tracked and penalized for their reproductive health choices.

- Study results show that there are many scenarios where dangerous permissions and attack surface vulnerabilities can access the restricted app activities and data, and security assessment tools may not be detecting and reporting all. These hidden vulnerabilities pose a significant threat to potential privacy and security attacks.

  Post Roe v. Wade, vulnerabilities in period-tracking and fertility-monitoring apps put women at increased risk of privacy and security breach. These apps should be categorized separately with robust enforced rules and regulations to minimize the potential abuse to already disadvantaged and vulnerable population.

- The study findings encourage to initiate a collaborative effort among health professionals, developers, and policymakers to help developing more secure and trustworthy digital health management apps for women's health.

Our research highlights the increased risks that reproductive health apps can pose to women's health. The study results reveal that even the most popular period-tracking and fertility-monitoring apps with well-established market presence exhibit significant privacy and security vulnerabilities. These apps collect sensitive personally identifiable information (PII) and sensitive health data. These risks are particularly concerning for women who face disadvantages such as limited access to necessary healthcare and preventive interventions [196]. Therefore, it is crucial to prioritize the privacy and security of these apps to protect the reproductive health rights of vulnerable women.

## 6.2 Background

The 2020 CDC Abortion Surveillance Report shows that 615,911 abortions were reported, with an abortion rate of 11.2 per 1,000 women aged 15-44 years [208]. Teen unintended pregnancies are a social, medical, and financial burden, and the teen birth rate dropped 64% between 1991 and 2015, resulting in $4.4 billion in public savings in 2015 alone [209, 210]. However, 30% of women aged 13-44 in the United States are in need of publicly funded contraception, and 98% lack reasonable access to these resources, particularly vulnerable populations such as young women in foster care [210, 209].

### 6.2.1 Reproductive Health Legal Disparity and Health Endangerment

The right to abortion is an essential liberty tied to basic rights such as family matters and bodily autonomy [211]. Unsafe abortion practices can lead to serious risks and even death, and accessing safe and legal abortion services can be challenging, particularly for vulnerable populations. The COVID-19 pandemic has forced healthcare providers to adopt telemedicine, raising concerns about privacy and security. With the overturning of Roe v. Wade, there are fears that app users' menstrual cycle data could be used to prosecute them [201, 202]. Period-tracking mobile apps may falsely claim to protect privacy and data security, and unauthorized access to healthcare-related data can be exploited. More evidence-based studies are needed to build trust in their clinical utility, efficacy, and safety. There are known biases and trends that can result in certain populations being targeted for the purpose of monetizing these predictions, and to gain more behavioral control [199]

### 6.2.2 Privacy, Security, and Legal Risks of Women's Health Mobile Apps

Reproductive health management mobile apps can assist women to be aware on fertile cycle and making conscious decision about the pregnancy, but collecting and misusing the app data can pose legal threats. There is also a possibility of biased legal actions towards minorities and widening the inequality gap [201, 212]. While it is true that privacy and security vulnerabilities may exploit mobile apps in general, women's health apps can have significant consequences including prosecution. Women's health apps may leak sensitive data such as menstrual cycles, fertility-monitoring, pregnancy stage, and use of contraceptives [213]. This information leakage can target vulnerable individuals that may lead to discrimination in employment and health insurance, harassment, and even violence. Period-tracking apps have raised concerns about user data security and privacy, particularly after the overturning of Roe v. Wade, which could potentially criminalize users [214]. Leading apps like Flo and Clue have issued statements to protect user privacy. Flo also introduced an anonymous mode, but it is only available on iOS and may have technical limitations [215, 216]. App developers face challenges in assuring users while complying with legal limitations. Users should crosscheck app features, privacy, and security claims when using these apps.

Digital healthcare can lower the burden on lower status healthcare workers, including low-income women of color. Period-tracking apps can help women manage their healthcare and family planning [217]. However, changes in medicine practices due to the COVID-19 pandemic and restrictions on abortion clinics have led to a significant increase in medication abortions, potentially putting app users at risk of prosecution for non-compliance with abortion laws if their identity is linked to the purchase of FDA-approved medications through vulnerabilities in the app [218].

## 6.3 Related Work

While there are efforts to discuss the practical benefits of digital health management for reproductive health, Lupton highlights significant ethical and privacy implications of self-tracking through reproductive activity tracking apps and the data they produce [219]. There is a challenge to ensure that healthcare apps, which can potentially impact patients and curate sensitive health data, undergo rigorous evaluation, and meet accuracy standards [220]. A 2020 study by LaMalva and Schmeelk shows that MobSF analysis identified at least one OWASP-defined security violation in 43.6% of studied healthcare android apps [221]. Alfawzan et al. found that all women's health apps, including reproductive health apps, allowed behavioral tracking and had poor privacy practices [34]. In a study by Shipp and Blasco, none of the privacy policies of 30 period tracking apps were easy to read, and the majority lacked transparency about the collected data [222].

Iyawa et al. found that mobile apps for self-management during pregnancy show positive impacts. however, mobile apps can be further studied for the identification of sexually transmitted infections, early warning signs of complexities during pregnancy and miscarriage [223]. A 2019 study by Grundy et al. showed that 79% of Google Play medicine related health apps regularly shared user data and were short of maintaining transparency with the app users. Studied apps demonstrate privacy leaks inferred with sharing sensitive information to remote server. User's privacy is threatened by apps receiving sensitive user data, including IP addresses and GeoIP [224]. Bull et al. analyzed data from the Natural Cycles app and found that tracking physiological parameters such as basal body temperature, rather than just cycle length, is critical to clinically identifying the fertile period [205]. A study found that the only FDA-approved Natural Cycles app, a highly effective hormone-free contraceptive method compared to traditional methods, has a failure rate of 8.3% and can produce unintended outcomes [35].

It is vital to know how non-technical issues may impact the privacy and security vulnerabilities of menstruation and fertility apps. For example, Aljedaani et al. found that 63% of mobile health apps' lack security because of absent security guidelines and regulations for developing secure mobile health apps. Developers of 56% apps may lack the necessary knowledge and expertise for secure app development [36]. Moreover, as per Earle et al., with rare involvement of health professionals or users in the design, development or deployment of period-tracking and fertility apps may restrict the understanding of the developer about the sensitivity of reproductive health data, and the critical consequences of privacy leaks and how the leaked information can harm users' physically, mentally, and legally [225].

We found limited studies of context-specific threats analysis of period-tracking and fertility apps. Post Roe v. Wade, users are more aware of legal implications, but technical analysis of privacy and security threats is needed to demonstrate potential damage from app vulnerabilities. Our study of 20 popular period-tracking and fertility-monitoring apps revealed exploitable vulnerabilities that pose privacy, security, and legal risks to users.

## 6.4 Research Objectives

We focused on menstrual cycle and ovulation tracking apps for female reproductive health, examining their availability and popularity on iOS and Android platforms. Our analysis of these apps' privacy policies, dangerous permissions, exploitable attack surface, and code flaws shows that flawed and vulnerable apps pose a significant risk to users' privacy and security, potentially leading to legal ramifications in the wake of the Post Roe v. Wade abortion rule in the USA.

Our primary goal is to identify the privacy and security vulnerabilities of reproductive health tracking apps and assess the potential risks they pose to women's health and rights. Our research objectives are as following:

- Analyze the availability of privacy policies and the information transparency.
- Analyzing the vulnerabilities associated with identifiable information collection and sharing.
- Identifying the exploitable App permissions, dangerous to manipulate app behavior and mishandle sensitive health data.
- Identifying the app threat surface which can cause the unintended app behavior and undesirable outcomes.

## 6.5 Methodology

We study 20 period-tracking and fertility-monitoring mobile apps which are available for both iOS and Android phones. We searched online for the available period-tracking and fertility-monitoring health apps in the Android/Google Play Store. Web search and app stores details were verified to get a list of apps, which are available in English language, free to download, with top-most ratings and a reasonable number of reviewers (High number of reviewers with higher rating is considered as app's credibility). These popular apps show users' trust with millions of downloads. These apps are listed as top mobile apps to consider for managing women's health digitally at a number of online resources [226, 227]. Table 6.1 presents the details of the considered period- tracking and fertility-monitoring apps in this work.

Table 6.1: Fertility, Ovulation, and Period-Tracking Mobile Apps-Downloads (M- Million; K-Thousand; +- More than), Ratings (out of 5) and # of Reviewers

| App Name | Downloads | iOS | | Android | |
|---|---|---|---|---|---|
| | | Rating | Reviews | Rating | Reviews |
| Always You: Period Tracker | 100K+ | 4 | 153 | 3.4 | 341 |
| Birth Control - Natural Cycles | 1M+ | 4.8 | 14,771 | 4.7 | 19.8K |
| Clover - Safe Period Tracker | 1M+ | 4.7 | 6,597 | 4.5 | 157K |
| Clue Period & Cycle Tracker | 10M+ | 4.8 | 340,427 | 4.3 | 1.17M |
| Eve Period Tracker | 1M+ | 4.7 | 107,014 | 4.4 | 26.5K |
| Fertility Friend- FF App | 1M+ | 4.8 | 6,958 | 4.8 | 18K |
| Flo | 100M+ | 4.8 | 1,051,482 | 4.6 | 3M |
| Glow | 1M+ | 4.7 | 65,426 | 4.3 | 70.6K |
| Kindara: Fertility Tracker | 100K+ | 4.7 | 8,837 | 3.2 | 2.02K |
| Luna | 5K+ | 4.3 | 9 | 3.7 | 61 |
| MagicGirl/Teen Period Tracker | 500K+ | 4.6 | 1,083 | 4.6 | 7.36K |
| Maya | 5M+ | 4.8 | 2,437 | 4.7 | 241K |
| My Calendar | 10M+ | 4.8 | 33,234 | 4.8 | 420K |
| MyDays X | 1M+ | 3.6 | 515 | 4.1 | 46.3K |
| Ovia Fertility & Cycle Tracker | 1M+ | 4.8 | 66,500 | 4.6 | 75K |
| Period Diary | 500K+ | 4.7 | 66,984 | 2.8 | 9.29K |
| Period Tracker | 10M+ | 4.8 | 61,374 | 4.6 | 360K |
| Premom Ovulation Tracker | 1M+ | 4.7 | 18,869 | 4.1 | 10.4K |
| Spot On Period Tracker | 500K+ | 4.3 | 15,211 | 4.2 | 7.33K |
| Stardust Period Tracker | 50K+ | 4.2 | 12,645 | 4.4 | 2.31K |

### 6.5.1 Mobile App: Components and Vulnerabilities

A mobile app consists of a User Interface (UI), app components, functionalities, and permissions. The Activity is a UI component with multiple screens, while the Service is a non-UI component that executes operations in the background. The Content Provider shares data within the same app or with other apps, and the Broadcast Receiver verifies the authenticity of intents from authorized sources [192].

App permissions enable an app to access device features and data resources necessary for its intended functionality. Android defines different categories of permissions, including runtime permissions that give apps additional access to restricted data or actions. These dangerous permissions can exploit sensitive information without user consent, such as location, contacts, and microphone and camera access [32, 228].

### 6.5.2 Vulnerability Analysis Methods and Tools

#### *6.5.2.1 Static Analysis*

Static Analysis is performed to examine decompiled APK files using Android Studio 3.6.1. Static Analysis or code analysis analyzes the source code of an application to high- light possible vulnerabilities without running the code. We acquired the latest APK file for all studied apps through an Android mobile app *APK Extractor*. Extracted .APK files are decompiled with an online APK decompiler.

#### *6.5.2.2 Dynamic Analysis*

Android dynamic testing tool *Drozer 2.3.4* [229] is used for dynamic analysis for finding the attack surface and the app permissions. Drozer is a well-known, open-source secu- rity assessment tool for Android apps. Test mobile device has a Drozer client installed, and Drozer commands were executed on Windows 10 machine. Drozer can discover the potential runtime exploitable vulnerabilities in Android apps [229].

###### *6.5.2.3 MobSF Analysis*

Mobile Security Framework (MobSF) is an automated, all-in-one mobile appli- cation (Android/iOS/Windows) pen-testing, malware analysis and security assessment framework capable of performing static and dynamic analysis. MobSF can analyze an individual .apk file uploaded to a MobSF Web tool. Mobsf generates a report containing information regarding the app's security and privacy vulnerabilities. MobSF generated comprehensive report breaks down the vulnerabilities in different categories. For example, report provides data about network security, permissions, code analysis, etc. MobSF report provides information to ascertain the overall security status of the app with low, medium and high-risk ratings.

#### 6.5.3 Vulnerability Analysis Categories

On the privacy and security vulnerabilities, we summarize the findings in major categories of Privacy Policy Analysis, Permissions Analysis, Attack Surface Analysis, and Code Analysis.

###### *6.5.3.1 Privacy Policy Analysis*

All app policies are analyzed manually by reading the details and ex- tracting the information regarding information being collected, sensitivity of the information col- lected, data handling and sharing policies. We also collected the information on data retaining policies and data security measures stated in the privacy policies.

###### *6.5.3.2 Permission Analysis*

Dangerous permissions are those that allow access to the user's private data, sensitive device features, or other resources that can affect the user's privacy or security. Examples of dangerous permissions include access to the user's location, contacts, camera, microphone, or storage. When an app requests a dangerous permission, the Android operating system prompts the user to grant or deny the permission. If the user grants the permission, the app can access the requested resource or data. However, if the app misuses the permission or abuses the access it has been granted, it can compromise the user's privacy or security.

*APK decompiler*.

MobSF https://mobsf.live/

First, we observed app permissions manually through app installation on test mobile devices. App permissions were also analyzed by performing dynamic analysis through security assessment tool Drozer. MobSF security framework generates a detailed report showing app permissions, their categories (Normal, dangerous), and how these can be manipulated in privacy and security attacks.

#### *6.5.3.3 Attack Surface Analysis*

The Android application exports a component for use by other appli- cations. If the app does not properly restrict the use of these components by defined apps, any app can launch the component or access the data it contains. In the absence of defined permissions, these exported components are vulnerable to exploitation by any app, which allows malicious users and apps to access to sensitive interactions and data involved. A malicious app can send an explicit intent to the target app, assuming all received intents as valid, which may lead to unintended app behavior [192]. To identify the vulnerable attack surface, we executed Drozer commands, and then we performed MobSF analysis, which generates a summary of vulnerable App components.

## 6.6 Data Analysis and Results

Our analyses of reproductive health tracking apps identified exploitable vulnerabilities that could allow attackers to manipulate the app's behavior, exploit identifiable device and user information, track a user's location, and link sensitive reproductive health data to a user's identity. These threats pose significant privacy and security risks to users, including the potential for legal prosecution. The detailed study results in the following sections highlight critical privacy and security risks faced by users of reproductive health tracking apps.

### 6.6.1 Manual Analysis

Manual analysis consists of studying the privacy policies manually and identifying the app features or data collection points which may pose a concern and threat for privacy and security of app users. As per health and federal guidelines on privacy protection in healthcare and personal identification information (PII) is protected [230, 231].

However, studied apps are collecting a range of identifiable data and personal identification information (PII) including demographics, mobile device-ID, and individual's reproductive health data.

There are two major categories of information being collected through the app. First is, where app requests users to provide personal, account related, sensitive healthcare, communication, and data about others. *Health Sensitive Data* refers to weight, menstrual cycle dates, details on pregnancy (if applicable), body temperature, body measurements, symptoms, etc. *Communication/User Generated Content* refers to the information the user posted while on the app (such as text, pictures, videos, personal notes, etc.) or recordings of phone calls and private messaging with the app professionals. *Data About Others* refers to what the user chooses to share about other people (family members, unborn baby, friends, etc.) while on the app or if the user had granted access to another person. The second set of information can be collected by the app automatically, generally through the mobile device information. Device information includes device model, information about the operating system and the version, unique device identifiers, mobile operator and network information, device storage information, etc.

#### *6.6.1.1 Privacy Policy: Information requested from the users*

We analyze the apps privacy policies for vulnerabilities associated with identifiable information and sensitive data collection. There are only 3 out of 20 apps offer anonymous access to the app, and majority of these apps require an account with collecting PII and sensitive data. It is observed that 95% of studied apps request user *Email*. Other most requested identifiable data is Name (85%), Phone Number (50%), Address (45%), Location (55%), and some form of ID (15%). Other than these primary identifiable information, age/birth year (80%) and language (25%) can also indicate the user demographics, and password (45%) can be exploited for linked attacks as users have a tendency to use exact same or similar password for multiple accounts and services [232, 233].

Though the nature of data these apps handle is sensitive reproductive health information, 70% of the apps declare to collect the sensitive healthcare data. There are 60% of these apps also collect Communication/User generated content, and 20% request users to provide information about others. Other noticeable information, picture is being collected by 40% of these apps, and 50% apps request Payment information. The interesting observation is that 65% of these apps declare collecting *Other Information*, which may conceal the details of data being collected.

#### *6.6.1.2 Privacy Policy: Default Data Captured by the Apps*

Mobile apps are capable of collecting various types of information by default that do not require user input. Therefore, it is important for app privacy policies to maintain transparency, ensuring that app users are aware of the information being collected by the app. This transparency helps users make informed decisions about the personal information they share and the potential risks associated with data breaches or cyber-attacks. Table 6.2 presents the types of information to collect, as stated by privacy policies of studied reproductive health apps. We observe that 95% of studied apps privacy policies declare to collect IP address. According to a study conducted in 2016 by Sivakorn et al., 90% of studied apps use cookies, making them vulnerable to HTTP cookie hijack- ing attacks on mobile devices [234]. Additionally, 85% of apps collect device information, which can potentially be used to recover the user's identity and reproductive health activities, leaving them vulnerable to attackers and legal consequences. About 75% of apps collect time zone/location in- formation, and 70% of apps record which features are accessed by the user. This information, combined with network provider and browser information, usage frequency, and online activity, can create a detailed digital footprint of the user. Attackers can exploit this information to track user behavior and target them for personal gain.

#### *6.6.1.3 Privacy Policy: Data Handling and Data Security*

Users' option to opt out or withdrawing from the intended data collection, data retaining, account association/termination, data deletion, data encryption, and following data security practices by app developers are important to assess if an app can support and maintain the user's privacy and security. However, the unclear strategies or absent details in the privacy policy make it difficult for user to trust the app. In absence of trust, user, either may not be using the app, even if it is needed, or may not follow the app as intended which defeats the purpose of these apps as trusted

Table 6.2: Privacy Policy: Default Data Collected by the studied Reproductive Apps

| Default Data Collected by App | % of Apps |
|---|---|
| IP Address | 95 |
| Uses Cookies | 90 |
| Device Information | 85 |
| Time Zone/ Location | 75 |
| Accessed Features Within App | 70 |
| Internet Browser | 65 |
| Frequency of Use | 60 |
| Mobile Service Provider | 35 |
| Online Activity Data | 35 |
| Network Information | 35 |

allies of women in need when other forms of help are not accessible. For example, only 10% apps declare to have a security *Penetration Testing*, which can identify the vulnerabilities and developers can fix/update to maintain the app as secure. 55% of apps use data encryption for keeping the data anonymized, and only 15% of these apps are using any security vulnerability scanning and periodic data protection assessments to evaluate the data security threats.

Use may want to delete or deactivate the account, but 50% of apps do not say anything in their privacy policy about deleting or deactivating the account. However, 85% of these apps retain data but do not specify any period or relevant information regarding deleted or deactivated accounts. 20% of apps declare retaining data, even after account deletion. User account also may not be deleted right away after requesting. For example, app Clue requires a 30 day period, while FF App considers 7 days to delete the account.

Additional noteworthy observations from app privacy policies include the fact that if a user chooses to opt-in to sharing personal information, the app may receive payments from advertisers or sponsors for sharing that information. Another app privacy policy states that while users can submit a request to delete their data, the app has the option to refuse and may charge a fee if the request is excessive. Many users may not be aware of these less common terms and conditions, which can result in undesirable sharing, delays, and fees. As a result, users may be discouraged from taking action and may simply accept the default data sharing and retention policies of apps.

Table 6.3: Privacy Policy: Data Handling and Security

| Data Handling and Security | % of Apps |
|---|---|
| Encrypted | 55 |
| Systematic Vulnerability Scanning | 15 |
| Penetration Testing | 10 |
| Legal Measures | 15 |
| Periodic Data Protection Assessments | 15 |
| Delete / Deactivate Account | 50 |
| Erase Data | 75 |
| Data Retained (Account Terminated) | 20 |
| Data Retained (Period Not Specified) | 85 |
| Security (Not Specified/ Guaranteed) | 85 |

#### *6.6.1.4 Manual Analysis of App Permissions*

We observed the app permissions on the test mobile device, which are being requested from the users. We collected this data for both iOS and Android versions. The iOS apps request for Location, Photos, and Camera while Android apps request permissions for Contacts, Storage, Phone, Microphone, and Calendar in addition to Location, and Camera. Among iOS apps, the most requested permission by 50% of the apps is Camera, while for Android apps, it is the Storage permission, requested by 65% of the apps.

### 6.6.2 Period-Tracking Mobile Apps: Static Analysis

We study the security issues and vulnerabilities existing in code analysis that may be exploited in security and privacy attacks. Broadly, as categorized in Android Studio, we observed errors and warnings for four categories of *Accessibility*, *Correctness*, *Performance*, *Usability*, and *Security*. We further focused on identifying major security flaws. Though there are no errors for the categories of Accessibility, Performance and Security, but warnings are present in all four categories and these warnings pose privacy and security concerns. Security warnings vary from one to thirty nine, and these warnings present exploitable vulnerabilities for security and privacy attacks. For example, about 61% of studied apps in Table 6.4 shows that exported service does not require permission, which means that app does not follow the preferred behavior by Android to secure the activities with minimum and necessary access, which presents a vulnerability to exploit. An entity must have a defined permission in order to launch the service or bind to it. Without this, any application can use this service [228, 32]. 33% of apps lack permissions for Content Provider, and 28.57% of apps do not require permission for Receiver. Additionally, 52.38% of studied apps use the dangerous feature File.setReadable() to make files world-readable, potentially compromising security. Although formal mechanisms such as ContentProvider, BroadcastReceiver, and Service are recommended, they also present attack surfaces in studied period-tracking mobile apps due to absent permissions [228]. Furthermore, around 28% of apps allow for the vulnerability of dynamically loading code from unsafe locations, while 23.81% of apps carry the vulnerability of AllowBackup/FullBackupContent Problems, which can be exploited through Android Debug Bridge (ADB) backup and allow users to read all application data once backed up. [235, 228].

Table 6.6 shows the security vulnerabilities under the correctness warnings of Android Studio code analysis. It is interesting to observe that though these are not categorized under security vulnerabilities, they can be exploited for security attacks. For example, 71% of the studied apps show battery life issues, indicating that code may negatively affect battery life by consuming battery excessively, and that can cause inaccessibility to apps and the phone itself. All apps are using private APIs which may cause abrupt stopping or crash of app due to incompatibility with newer devices, which will cause user to lose data [236]. The unintended and unexpected app failure or inaccessibility may cause unintended pregnancy, and maybe a forced decision for abortion with potential legal and health risks.

### 6.6.3 Dynamic Analysis

In dynamic analysis of studied Period- tracking reproductive health mobile apps focus on identifying dangerous permissions, attack surface, and exploitable vulnerabilities on Android apps at runtime.

Table 6.4: Static Analysis- Security Vulnerabilities

| Security Vulnerabilities | # of Apps | % |
|---|---|---|
| Exported service does not require permission | 13 | 61.90 |
| The Network Security configuration allows the use of user certificate in the release version | 1 | 4.76 |
| Cipher.getInstance with ECB | 4 | 19.05 |
| File.setReadable() used to make file world-readable | 11 | 52.38 |
| Insecure Hostname Verifier | 4 | 19.05 |
| Receiver Does not require permission | 6 | 28.57 |
| addJavascriptInterface Called | 2 | 9.52 |
| Hardware ID Uage | 2 | 9.52 |
| Potential Multiple Cerficate Exploit | 2 | 9.52 |
| Using setJavaScriptEnabled | 2 | 9.52 |
| Using the result of check permission calls | 2 | 9.52 |
| load used to dynamically load code | 6 | 28.57 |
| Content provider does not require permisiion | 7 | 33.33 |
| AllowBackup/FullBackupContent Problems | 5 | 23.81 |
| Insecure TLS/SSL trust manager | 4 | 19.05 |

#### *6.6.3.1 Android Permissions: Dangerous*

App permissions analysis can identify the potential risk of violating users' privacy and security. Standard permissions allow access to data and actions which present minimal risk to the user's privacy. Runtime permissions allow additional access to restricted data and may significantly perform restricted actions that impact the system and other apps. For example, many runtime permissions access private user data, potentially sensitive information like user's location and contact information. Table 6.5 shows the dangerous permissions captured by the tools for the studied period-tracking and fertility apps. It provides tool's (Drozer and MobSF) ability to detect the vulnerability and a brief description of exploitable vulnerability caused by a dangerous permission. **Drozer-Dangerous Permissions** A per Table 6.7, there are 12 different dangerous permission identified in Drozer dynamic analysis [229]. Max 62% of apps have READ_EXTERNAL_STORAGE permission, while 43% of apps have WRITE_EXTERNAL_STORAGE. Coarse location permission can track the user's approximate location and 33% of apps have permission to do so, while 25%

Table 6.5: Dangerous Permissions: Drozer and MobSF Analysis; 1-Identified, 0-Not Identified

| Dangerous Permission | Drozer | MobSF | Exploitable Vulnerability |
|---|---|---|---|
| ACCESS_COARSE_LOCATION | 1 | 1 | Malicious applications can use this to determine approximately where user is. |
| ACCESS_FINE_LOCATION | 1 | 1 | Malicious applications can use this to determine where user is and may consume additional battery power. |
| AUTHENTICATE_ACCOUNTS | 0 | 1 | Allows an application to use the account authenticator for creating as well accounts as obtaining and setting their passwords. |
| CALL_PHONE | 0 | 1 | Malicious applications may cause unexpected calls on your phone bill. |
| CAMERA | 1 | 1 | Allows application to take pictures and videos with the camera, and collecting images that the camera is seeing at any time. |
| GET_ACCOUNTS | 1 | 1 | Allows access to the list of accounts in the Accounts Service. |
| MANAGE_ACCOUNTS | 0 | 1 | Allows an application to perform operations like adding and removing accounts and deleting their password. |
| POST_NOTIFICATION | 1 | 0 | Allows an app to post notifications |
| READ_CALENDAR | 1 | 1 | Malicious applications can use this to send your calendar events to other people. |
| READ_CONTACTS | 1 | 0 | Allows an application to read the user’s contacts data. |
| READ_EXTERNAL_STORAGE | 1 | 1 | Allows an application to read from external storage. |
| READ_PHONE_STATE | 1 | 1 | An application can determine the phone number and serial number of this phone, whether a call is active, the number that call is connected to and so on. |
| RECORD_AUDIO | 1 | 1 | Allows application to access the audio record path. |
| SYSTEM_ALERT_WINDOW | 0 | 1 | Malicious applications can take over the entire screen of the phone. |
| USE_CREDENTIALS | 0 | 1 | Allows an application to request authentication tokens. |
| WRITE_CALENDAR | 1 | 1 | Malicious applications can use this to erase or modify your calendar events or to send emails to guests. |
| WRITE_EXTERNAL_STORAGE | 1 | 1 | Allows an application to write to external storage. |

Table 6.6: Static Analysis- Major Correctness Vulnerabilities

| App | Battery Life Issues | Using Private APIs |
|---|---|---|
| Always You | √ | √ |
| Clover | √ | √ |
| Clue | √ | √ |
| Eve | χ | √ |
| FF App | χ | √ |
| Flo | √ | √ |
| Glow | χ | √ |
| Kindara | χ | √ |
| Luna | χ | √ |
| MagicGirl | √ | √ |
| Maya | √ | √ |
| My Calendar | √ | √ |
| My Tracker | √ | √ |
| MyDays | √ | √ |
| Natural Cycles | √ | √ |
| Ovia | √ | √ |
| P.Tracker | √ | √ |
| Period Diary | √ | √ |
| Premom | √ | √ |
| Spot On | √ | √ |
| Stardust | √ | √ |

of apps can track fine location of users which is more accurate location and thus, potentially more damaging to user's privacy and security. 29% of apps have Camera permissions to access sensitive audio/video content. 14% of Apps may have capability to access and manipulate the accounts. 10% of these apps can access audio record functionality and 5% of apps can read and write the calendar.

#### *6.6.3.2 Dynamic Analysis-Drozer: Attack Surface*

Dynamic analysis tool Drozer identifies exploitable attack surface of the android apps. Table 6.8 shows that Activities, the user interactions with UI, are most vulnerable with as high as 39 activities exported by a single app. We observed total of 239 components exported forming the attack surface for 20 period-tracking apps. Activities exported are forming 44.77% (107) of all exposed attack surface components identified for studied apps.

Table 6.7: Dangerous Permissions Carried by Studied Period-Tracking Apps: Drozer and MobSF Analysis

| Dangerous Permissions | Drozer<br>(% of Apps) | MobSF<br>(% of Apps) |
|---|---|---|
| ACCESS_COARSE_LOCATION | 33 | 30 |
| ACCESS_FINE_LOCATION | 24 | 20 |
| AUTHENTICATE_ACCOUNTS | 0 | 20 |
| CALL_PHONE | 0 | 5 |
| CAMERA | 29 | 25 |
| GET_ACCOUNTS | 14 | 15 |
| MANAGE_ACCOUNTS | 0 | 15 |
| POST_NOTIFICATIONS | 29 | 0 |
| READ_CALENDAR | 5 | 5 |
| READ_CONTACTS | 5 | 0 |
| READ_EXTERNAL_STORAGE | 62 | 50 |
| READ_PHONE_STATE | 10 | 10 |
| RECORD_AUDIO | 10 | 10 |
| SYSTEM_ALERT_WINDOW | 0 | 5 |
| USE_CREDENTIALS | 0 | 20 |
| WRITE_CALENDAR | 5 | 5 |
| WRITE_EXTERNAL_STORAGE | 43 | 60 |

Content Provide is least vulnerable with 4.18% (10), while unprotected Service (21% (51)) and Broadcast Receiver (29% (71)) components may allow attackers to access sensitive information and manipulate app's behavior.

#### 6.6.3.3 *Dynamic Analysis: Injection and Other Vulnerabilities*

SQL injection attack can expose private data, corrupt database contents, and even compromising of backend infrastructure. SQL can be vulnerable to injection via queries that are created dynamically by concatenating user input before execution [237].

Each of the studied apps was scanned for SQL injection in content providers using Drozer [229]. Though, 19 out of 20 studied apps did not show any injection vulnerabilities, one app is vulnerable to content provider injection and shows two instances each for the projection and selection vulnerabilities. Injection vulnerabilities can expose sensitive user or application data, overcome authentication and authorization restrictions, and leave databases vulnerable to corruption or deletion. Impacts can include dangerous and lasting implications for users who's personal

Table 6.8: Attack Surface: Drozer and MobSF Analysis, Activity Exported-AE; Broadcast Receiver Exported- BRE; Content Provider Exported-CPE; Service Exported-SE

| | Drozer | | | | MobSF | | | |
|---|---|---|---|---|---|---|---|---|
| App Name | AE | BRE | CPE | SE | AE | BRE | CPE | SE |
| Always You | 2 | 3 | 1 | 2 | 1 | 3 | 1 | 3 |
| Natural Cycles | 2 | 2 | 0 | 1 | 3 | 3 | 1 | 2 |
| Clover | 1 | 2 | 0 | 1 | 0 | 1 | 0 | 2 |
| Clue | 39 | 4 | 0 | 5 | 38 | 4 | 0 | 5 |
| Eve | 7 | 7 | 4 | 5 | 6 | 7 | 4 | 6 |
| FF App | 2 | 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| Flo | 7 | 4 | 0 | 3 | 6 | 4 | 0 | 4 |
| Glow | 8 | 7 | 3 | 5 | 7 | 7 | 3 | 6 |
| Kindara | 2 | 1 | 0 | 4 | 1 | 1 | 0 | 4 |
| Luna | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| MagicGirl | 1 | 1 | 0 | 2 | 0 | 1 | 0 | 2 |
| Maya | 5 | 5 | 0 | 2 | 4 | 5 | 0 | 2 |
| My Calendar | 1 | 6 | 0 | 4 | 0 | 6 | 0 | 4 |
| MyDays X | 1 | 2 | 1 | 3 | 0 | 2 | 1 | 3 |
| Ovia | 5 | 6 | 0 | 2 | 4 | 6 | 0 | 3 |
| Period Tracker | 2 | 2 | 1 | 4 | 1 | 2 | 1 | 4 |
| Period Diary | 4 | 1 | 0 | 2 | 3 | 1 | 0 | 3 |
| Premom | 8 | 5 | 0 | 2 | | | | |
| Spot On | 2 | 4 | 0 | 2 | 1 | 4 | 0 | 3 |
| Stardust | 7 | 7 | 0 | 2 | 10 | 4 | 1 | 2 |

data has been exposed. Providers of apps and services risk losing intellectual property or user trust. Injection vulnerabilities can also affect Content providers if permissions are not as intended or absent [237].

### 6.6.4 MobSF Analysis

We performed the MobSF analysis to compare the app privacy and security vulnerabilities compared with other analysis performed in this study. MobSF provides an overall security score (1-100) to show the security risk level of the app. 10% of apps show *Low Risk*, 75% of apps are with *Medium Risk* with 40-59 score, and other 15% of apps show *High Risk*.

#### *6.6.4.1 MobSF- Dangerous Permissions*

Table 6.7 shows 15 different dangerous permission identified in MobSF analysis. More than 50% of apps have READ_EXTERNAL_STORAGE and WRITE_EXTERNAL_STORAGE permissions.

Coarse location permission can track the user’s approximate location and 30% of apps have permission to do so, while 20% of apps can track fine location of users which is more accurate location and thus, potentially more damaging to user’s privacy and security. 25% of apps have Camera permissions to access sensitive audio/video content. 20% of apps show permission to authenticate account. 15% of Apps may have capability to access and manipulate the accounts. 10% of these apps can access audio record functionality and 5% of apps can read and write the calendar.

#### *6.6.4.2 MobSF- Network Security*

MobSF evaluates the *Network Security* by verifying the possibility of Man-In-The-Middle (MITM) attacks to intercept the information exploiting the insecure net- work configuration. As per MobSF analysis, app is secure if Base config is configured to disallow clear text traffic to all domains. An app’s network security is considered highly vulnerable when Domain config is insecurely configured to permit clear text traffic to domains in scope. It is ob- served that 20% of analyzed apps are secure, 15% of apps are highly vulnerable, and rest of apps do not show network security information in MobSF analysis.

#### *6.6.4.3 MobSF- Code Analysis*

Code analysis by MobSF identifies the vulnerabilities in app code .apk files with the security standards defined by OWASP and CWE [238, 192]. There are eighteen dif- ferent known and defined code vulnerabilities detected in code analysis of studied period-tracking mobile apps, and eleven out of eighteen issues fall into OWASP Top 10 vulnerabilities. Appendix Table A.1 presents these vulnerabilities and relevant OWASP/CWE references. 20% of apps have high risk vulnerabilities, including insecure communication, inadequate integrity checking, and improper platform usage. Vulnerabilities, such as MITM or Padding Oracle attacks, can lead to unauthorized access and decryption of sensitive reproductive health information. Other issues include Hardware ID usage, IP address disclosure, SQL injection, and hardcoding sensitive data.

#### *6.6.4.4 MobSF- Network Security and Certificate Analysis*

20% of studied apps found as *Secure* on network security, and 15% of apps present high risk. Network security severity is considered high.

if domain configuration is insecurely configured to permit clear text traffic to domains in scope. About 60% of studied apps do not show information about network security.

Android apps are signed with digital certificates to ensure authenticity and integrity, with multiple signatures (v1, v2, and v3) available to maintain validity and security. A false signature could indicate that the app's certificate has been tampered with or modified, potentially indicating malicious intent. MobSF analysis shows that 20% of apps have v1 signature as false, and 35% of studied period-tracking apps show v3 signature as false. A violation of an app's authenticity and integrity can call into question the validity of its behavior and guidance for managing monthly cycles or fertility monitoring.

#### *6.6.4.5 Threat of Hidden and Undetected Dangerous Permissions*

We observed seven different dangerous permissions manually with the installed apps. However, the drozer dynamic analysis identified twelve, and the MobSF analysis captured fifteen different dangerous permissions used by the studied period-tracking apps. Table 6.5 shows a summary of all dangerous permissions observed by Drozer and MobSF analysis for studied period-tracking apps. Different tools may detect and handle potentially exploitable app vulnerabilities reporting differently. Another observation on runtime permissions is that users may not know all the relevant information on dangerous permissions. For example, app may request user's permission for *Location*, but user wouldn't know if app is tracking absolute (FINE) or approximate (COARSE) location or both.

## 6.7 Challenges and Limitations

The app selection process for our study was limited to apps supported by both iOS and Android to keep the study practical. We focused on analyzing free apps that are more accessible to the intended population, which also restricted the scope of analysis for certain apps included in the study. For example, MobSF could not analyze apps with very large .apk files. We collected some data and then reshuffled the apps list to maintain consistency with the manual, dynamic, and static analysis tools utilized. In terms of vulnerabilities, we observed numerous privacy and security warnings in the code analysis, but it is beyond the scope of this study to describe every vulnerability that could be used to manipulate the app and user behavior, potentially leading to unintended outcomes. Additionally, this study is limited in examining both iOS and Android app counterparts, and Android apps were more in-depth analyzed due to the availability of .apk files. We were also not able to analyze the complete app functionality of unpaid versions, and vulnerability testing tools may not be exhaustive in identifying vulnerabilities.

### 6.8 Discussion

Period-tracking apps are rapidly growing to meet the reproductive healthcare needs of women, adapting to changing healthcare practices and social environments. However, legal disparities are exacerbating the vulnerabilities of these apps and posing a greater contextual threat to women's reproductive health. These privacy and security threats have a heightened impact, not only physically, mentally, and emotionally on women's reproductive health, but also in the criminalization of abortion, where women bear a disproportionate burden, while male counterparts face little or no accountability in the process. As such, there is a pressing need for greater attention to the privacy and security implications of period-tracking apps, particularly in the context of women's health.

Amid the rapid development of mobile apps to meet changing market needs, it is crucial to prioritize general vulnerabilities over the usability and accessibility of app features. Developers should only request dangerous permissions, when necessary, handle user denials gracefully, and implement safeguards to prevent misuse. Enforcing guidelines and rules can increase developer accountability, while adhering to best practices recommended by development frameworks and federal government agencies can reduce attack surfaces. Additionally, authentic guidance for selecting reproductive health management apps, backed by clinical trials and vetted by authorities such as the FDA, FTC, NIH, and HHS, can help users make secure and safe choices.

Post Roe v. Wade, there is an increased attention by app developers to provide anonymous access, but implementation may take longer with technical and operational challenges [215, 216]. We should not wait for major incidents to occur before taking action. Developing a framework for evaluating period-tracking apps pre-market can help minimize post-market privacy and security breaches. This approach can also reduce the burden on the legal system and the cost of unintended pregnancies, which continues to pose a significant healthcare burden in the United States [212, 210, 208]. A longitudinal study can observe patterns in developing sensitive health data handling apps with improved security by design, providing valuable insights for future policies to protect women's privacy and security in the context of health apps.

### 6.9 Conclusion

We analyzed the privacy and security vulnerabilities of period-tracking mobile apps for women's reproductive health management. While these apps offer promise in reaching disadvantaged populations in need of medical consultation and assistance, they are vulnerable to exploitation due to technical loopholes. App development faces challenges in complying with secure development practices that must adapt to changing market needs. While period-tracking and fertility apps are not regulated by HIPAA, they are accountable under FTC rules to protect sensitive health data. Transparent privacy policies and peer evaluations can help users make informed decisions.

An app collecting identifiable information, sensitive reproductive health data and with unclear or inappropriate data-sharing policies can jeopardize user's privacy and security. A manipulated period-tracking app can provide incorrect guidance and results which may result in undesirable situations like unintended pregnancy, and pregnancy complexities may force someone to opt for the abortion. Post Roe-V-Wade, identifying the users and tracking their reproductive patterns pose a serious threat of using this information against already burdened population. Illegal abortion is more challenging and damaging with inaccessible right medical resources, which can endanger the patient health including losing lives [212].

Our study revealed exploitable privacy and security vulnerabilities in period-tracking and fertility-monitoring apps, which could also put users at legal risk in the post Roe v. Wade era. These vulnerabilities, risks, and threats impose a burden on social, financial, and legal resources. However, investing in defining standards and enforcing rules can hold app developers accountable for maintaining security by design principles. Collaborative efforts among medical professionals, users, app developers, and legal experts can help develop secure digital health management tools for women's health, reducing the overall healthcare burden.

# 7. THE VULNERABILITY OF VOICE CONVERSION IN PRESERVING SPEAKER ANONYMITY: AN ANALYSIS WITH SPEAKER RECOGNITION SYSTEMS*

Anonymized voice data is essential for maintaining privacy in voice-based exchanges of sensitive healthcare information. Voice conversion is a commonly used method to obscure the original speaker's identity. Voice conversion methods aim to transform the voice characteristics of a source speaker to make them sound like a target speaker, but often prioritize altering the voice identity to target speaker's over completely obscuring the source speaker's voice features. However, speaker recognition systems are advancing in their ability to identify speakers with greater precision and accuracy and may be able to detect residual subtle voice features of the source speaker in voice-converted samples. Our work addresses the research problem of undermining the anonymity of voice-converted samples by identifying the source speakers, rather than verifying the effectiveness of voice conversion for voice-biometric hacking, a problem that has been studied well [239, 240]. In voice-based applications, both preserving voice anonymity and correctly identifying the speaker are contextually critical but open to attacks. This creates a persistent challenge of balancing between maintaining speaker anonymity and improving speaker recognition accuracy. Violating voice anonymity can result in subsequent privacy and security threats. Our contribution lies in demonstrating the threat of identifying source speakers from converted voice samples. We found that the possibility of breaking voice anonymity is more than random guessing with multiple voice conversion tools. One-to-one voice conversion is more vulnerable to voice anonymity-breaking threats than many-to-many and any-to-any voice conversion. The target voice features, voice conversion techniques, and speaker recognition methods affect the likelihood of identifying the original speaker from the voice-converted anonymized speech data.

*Submitted work, Under Review-2023

## 7.1 Introduction

Voice anonymization refers to the process of obscuring or removing personally identifiable information (PII) from a speaker's voice to protect their privacy. Voice anonymity can conceal the speaker's identity, including timbre, pitch, speaking rate, and speaking style, while preserving the naturalness and intelligibility of the spoken content [241, 242]. The desire to communicate without revealing one's identity is often driven by specific scenarios, such as voice transcripts in a medical patient-provider setting [243]. Integrating voice technologies into clinical practice can enhance patient-provider interaction, especially with the increasing adaptation of telemedicine involving voice inputs and voice notes [16]. Remote mental health assistance has helped a large, distressed population during the COVID-19 crisis. However, it has also raised privacy and security concerns. Voice anonymization can play a critical role in protecting the privacy of individuals seeking mental health assistance. Therefore, it is important to ensure that voice-based technologies used in such settings are secure and maintain the confidentiality of patients' personal information.

While voice identity can be crucial in identifying suspects and serving as forensic evidence, protecting the privacy of voice inputs from victims or whistleblowers is of great importance. Gender recognition and emotion recognition can also help improve workforce balance and healthcare needs [244, 245]. However, exploitation of speaker-dependent voice features can introduce biases and can be used against re-identified individuals or groups. Preserving privacy in voice data is a challenging task, especially against attackers who may gain legal or illegal access to voice data and attempt to infer the speaker's identity and other personally identifiable information [246, 243]. Effective privacy preservation techniques are essential for balancing the benefits of voice-based technologies with the need to protect the privacy of individuals and groups.

A speaker's identity can be re-established through linkage attacks, linking anonymized voice data and other sources of PII through data mining, cross-referencing of voice data, or voice recordings of digital voice assistants with social media or public records [247]. Another attack can employ machine learning techniques to build models to predict an individual's identity from anonymized voice data. Privacy leaks through unintended disclosure of sensitive personal information can be used to re-identify an individual from anonymized voice data.

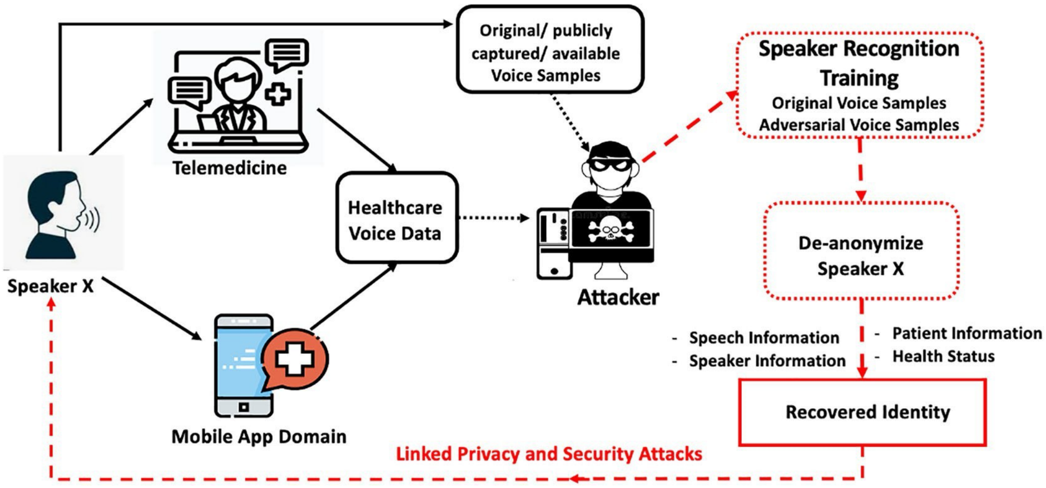


Figure 7.1: High-Level Overview of Threats to VC-based Speaker Anonymization through Speaker Recognition Systems

Voice Conversion (VC) can achieve voice anonymization by transforming the original speaker's voice into a different voice (target speaker) so that the speaker's identity is obscured. However, there may be subtle clues of speaker-dependent voice characteristics in the transformed voice, which could potentially be used to re-identify the source speaker. The effectiveness of voice conversion may also depend on the quality and specificity of the training data used to develop the conversion model, as well as the accuracy of the conversion algorithm. While voice anonymization can protect the privacy of patients when their voice data is collected as part of their medical records, it may not be able to completely preserve the speaker's privacy. In fact, weak anonymization techniques can be easily breached with advanced speaker recognition methods, leading to a potential privacy breach. It is important to continue improving voice anonymization techniques to ensure that speaker identities are protected.

The breaching of voice anonymity can be dependent on applied methods for voice conversion and speaker recognition. Therefore, we design multiple attack scenarios to find the probability of identifying source speakers from the anonymized voice samples. A detailed analysis of multiple

VC methods and SR systems is performed to evaluate the relative threat of attacking scenarios in the field of voice manipulation and speaker recognition advancements. Figure 7.1 depicts a high-level overview of threat of breaching the voice anonymity. Due to the increased availability of tools and resources, an attacker may be able to obtain a sufficient quantity of voice samples to train a speaker recognition (SR) model. This model would be capable of identifying the source speaker from newly acquired voice samples.

Our contribution lies in showing the practical feasibility of attacking the voice anonymity achieved through a range of voice conversion methods and employing off-the-shelf speaker recognition systems. Our findings demonstrate that percentage of source speaker identification is more than random guessing in majority of the studied attacking scenarios. Our results show that an attack is attainable to re-identify the source speaker from the anonymized voice samples. We believe that these attacks can be more damaging with the ongoing advancement in the field of voice recognition technology. Therefore, it is vital to measure the present vulnerabilities to minimize the potential threat of breaching the speakers' privacy and security.

## 7.2 Background

Voice inputs are getting more mainstream as multitasking users can talk to humans and machines hands-free. Voice-based machine-human interactions can assist speech-impaired, disabled, and aging populations [15]. Though extracting voice features to represent each voice uniquely is significant to maintain the low probability of duplication [248]. These unique speaker-dependent voice features are also key to generate efficient synthesis voices to improve human lives with applications like producing audiobooks or helping speech-disabled people [249]. Personalized digital assistants or virtual coaches can be more comforting for living alone seniors in a convincing or familiar voice [250]. It can also enhance the audience's grasp with natural-sounding speech translation services.

### 7.2.1 Threat of Breaching Voice-Privacy

In the privacy domain, attacks targeting data privacy include, for example, an attacker aiming to determine if the voice of a certain individual was used for training a speaker identification system. In response to these potential attacks, privacy-preserving defenses have been designed to prevent privacy leakage of the raw data. These defenses fall between anonymization and cryptography [251]. For example, anonymization aims to make the speech input unlinkable, i.e., ensure that no utterance can be linked to its original speaker by altering a raw signal and mapping the identifiable personal characteristics of a given speaker to another identity [40]. Various studies have proposed anonymization methods based on noise addition [252], voice conversion [39], speech synthesis [29], and adversarial learning [253], considering the speaker identity [251] or emotion [254] as a sensitive attribute. Another study by Yoo et al. proposed to adopt many-to-many voice conversion techniques based on variational autoencoders (VAEs) to anonymize the speech data shows that certain VC techniques can be used to protect the personal information hidden in the voice data samples, which are used to train the SR models [255]. There is a need to mitigate inherent biases in speaker recognition systems as machine learning process may be biased because of imbalanced data inputs. It is observed that female speakers are more vulnerable, and separate models should be developed for male and female speakers to reduce the bias in voice-based solutions [256, 257]. Our work also analyzes if the studied SR methods behave any differently towards identifying source speakers from the VC samples converted using male or female target voices. We also observe if source speakers of one gender are more vulnerable to be identified than the other.

Voice data can be vulnerable to insecure communication channels, possibly attacked and hijacked by malicious apps or users. If the voice data is stored in the cloud, this presumably protected communication can be accessed by exploiting cloud data security threats. Natural language processing techniques can extract information from a user's voice search history, voice command history, and voice-based text messages and emails. Voice data can have details of users' demographic, habits, schedules, and preferences. Once a user is identified in real-life, inferred details can be used in more serious attacks, such as stalking or robbery [29]. In Illinois, US, courts are increasingly inspecting cases involving people's voice data. McDonald's, Amazon, and Google are all facing judicial scrutiny over how they use people's voice data [258]. The decisions in these cases could lay down new rules for the protection of people's voices. User awareness is critical to communicate responsibly, but it is interesting how technology can be a savior and destructive as well if not handled appropriately.

Research shows that worldwide, 45% of smart speaker users are concerned about voice data privacy, and 42% worry about voice data hacking. In another survey, 59% of respondents said privacy is an important factor when using voice control devices [37]. Information, such as classified company data or health and medical details recorded by a doctor's note-taking voice assistant, is considered sensitive. Using cloud services in voice and language applications has significant disadvantages related to security, safety, and privacy concerns [39].

In the following parts of this section, we present the base knowledge of the voice characteristics, voice anonymization, voice conversion, and voice de-anonymization techniques in practice, which can be utilized in designing voice anonymity attacks.

### 7.2.2 Voice Identity Features

Each voice sample has speaker-dependent voice characteristics, such as fundamental frequency (F0), formants, intonation, intensity and duration can be critical for speaker recognition systems to identify the speakers [259, 260]. The uniqueness of a voice pitch may vary depending individual's physiology, age, gender, and speaking style, and can be used as a relatively stable speaker characteristic for speaker recognition purposes. However, a unique voiceprint is a combination of other voice features including Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and formant frequencies. Additionally, pitch can also be affected by factors such as stress, fatigue, medical condition, or speaking non-native language [248]. Formant frequency is the resonance frequency in a speech sound produced by the acoustic resonance of the vocal tract, contributing to the perception of the unique vowel sound quality.

### 7.2.3 Voice Anonymization Techniques

Voice anonymization has basic goal to hide the speaker's identity and the technique depends on the desired level of privacy, output quality, and the computational resource available. Voice transformation modifies the spectral and prosodic characteristics of the voice signal, such as pitch, formants, and spectral envelope, to alter the speaker's voice and make it more difficult to recognize. Adding noise to the voice signal can obscure the speaker's voice, and voice-masking with white noise or a frequency-masking filter can obscure the speaker's voice. There can be multiple target speakers used to hide the source speaker's identity to make it more difficult for adversary to identify the source speaker to follow the k-anonymity concept for hiding sensitive information by introducing k-1 dummies [261, 262].

Voice de-identification identifies and removes personal information from the voice signal, such as speaker identity, gender, and age. The anonymization method can also be dependent if it is for a human listener or a machine listener. True anonymization is not possible without completely changing the voice and when you completely change the voice, then it's not the same voice. Despite this, it is still worth developing voice-privacy technology, as no privacy or security system is totally secure [263]. Voice conversion techniques can be used to convert the speaker's voice to a different target speaker and obscure the speaker's identity. We focus on voice conversion methods to achieve the speaker anonymization, and then evaluate the possibility of breaking the speaker's anonymity using speaker recognition techniques.

### 7.2.4 Voice Conversion Techniques

There has been a noticeable progression in the field of voice conversion techniques. Statistical parametric VC is based on the spectral and prosodic characteristics of the source and target speaker's voices using a Gaussian Mixture Model (GMM), a Deep Neural Network (DNN), or a Generative Adversarial Network (GAN)[264]. Non-parametric VC do not rely on explicit statistical models, but instead use techniques such as formant shifting, pitch shifting, or formant preservation to convert the source speaker's voice to the target speaker's voice [265]. Cycle-Consistent

Adversarial Network (CycleGAN) is a type of GAN with the unsupervised learning of model to map between the source and target speaker's voice characteristics. CycleGAN-VC allows the voice conversion with a sufficient number of training examples and learning one-to-one-mappings [266]. However, another variation of GAN, StarGAN-VC allows non-parallel many-to-many mappings [267].

Phonetic Posteriorgram (PPG) based VC uses phonetic information to convert the source speaker's voice to the target speaker's voice [268]. X-vector representation is a deep neural network-based technique for speaker representation as a fixed-dimensional embedding that captures speaker-discriminative information, and voice conversion model maps the x-vector of the source speaker's speech to the target speaker [269, 270].

There are voice conversion categories defined based on the number of source and target speakers, conversion method, and involved voice data samples. *Parallel VC* needs same voice utterances from source and target speakers but *Non-Parallel VC* can work with different sets of utterances. *One-to-One VC* is a parallel VC where voice feature mapping is one-to-one between a source speaker and a target speaker. *Many-to-Many VC* can convert an individual's voice from one speaker to multiple target speakers, for all speakers seen in the model training, and *Any-to-Any VC* can convert an individual's voice from one speaker to any other speaker, regardless of the number of speakers involved in the model training. There are multiple methods using a combination of deep learning models and statistical techniques to learn the mapping between source and target speaker voices.

### 7.2.5 Voice De-anonymization Techniques

An informed attacker is more equipped to breach the voice anonymity with having the information of available methods and which method is more useful to re-identify in a specific scenario. *Voice Recognition* algorithms can learn the speaker classification features to match the unknown voice to any of the speakers in a database. The *Speaker Diarization* process can identify individual speakers from a mixture of audio sources, which can then be compared against a database of known voices to identify the speaker. An *Acoustic Analysis* can identify the speakers based on the closest similarity of the physical properties of a speaker's voice like pitch, rhythm, and formants. The speaker's identity can be analyzed based on the *Side Channel Information*, like background noise or accelerometer-sensed reverberations [271]. *Humans* as voice experts can listen to the unknown sample and may identify the speaker. Though these are different ways to identify the speakers from the voice samples, focus of our work is to re-identify the anonymized voice-converted samples through speaker recognition tools. We analyze the anonymized voice samples through deep learning solutions for speaker recognition. Speech and voice are used interchangeably in this work.

## 7.3 Related Work

There is a potential threat of revealing not only the speaker's identity but other sensitive information from the utterances such as emotions, and health state by speaker recognition methods [272, 273, 274]. Voice conversion using bilinear functions and the warping functions are invertible and can reveal the speaker's identity but using random pitch and distortion strength parameters and adding noise can prevent such attacks [29].

Jin et al. achieved high quality speaker de-identified voice output by modifying the spectral and prosodic features using GMM-based and Phonetic approach-based voice transformations [241]. In 2014, speaker de-identification was tried to convert source speakers with a target synthetic voice with a de-identification rate of 91% [275]. Justin et al. proposed speaker de-identification using diphone recognition and speech synthesis which is limited by degraded naturalness of de-identified samples. The conclusion is contradictory saying that the proposed approach lacks the ability to produce (de-identified) speech [276]. Another study was done by Bahmaninezhad et al. by mapping MCPE and Log-F0 to average of all target speakers (gender dependent and gender independent) and aperiodicity (AP) features were directly mapped from source to de-identified speaker [277]. Qian et al. presented VoiceMask to disguise the speaker's voiceprint by randomly modifying the speaker's voice via robust voice conversion satisfying differential privacy to protect the voices shared in the cloud [29].

Fang et al. proposed anonymizing speakers through separating speaker identity and the linguistic content in the form of phoneme posteriorogram (PPG). PPG and pre-trained x-vector system to encode the anonymous speaker's identity with some degraded speech quality [270]. Speaker anonymization through Vocal Tract Length Normalization (VTLN) and sanitization by substituting sensitive keywords of voice input before sharing from mobile devices to cloud was proposed by Qian et al. [278]. It shows that speaker recognition reduced from 100% to 16% with only decreasing speech recognition accuracy by 14.2%.

Voice Conversion Challenge (VCC) 2016 analysis talks about the similarity of converted samples to the target and source speakers but it presents more data on similarity towards target speakers and the analysis is subjective based on human listeners which shows more than 80% similarity for female-to-female conversion more than random guessing of 50% [279]. Another work by Cai et al. explores source speaker identification with one-to-three VC using any-to-any VC tools. Their work demonstrates some feasibility of identifying source speakers with low overall prediction [280]. The relevant problems of voice-conversion and speaker-identification may continue with emerging technologies improving both processes. A more efficient VC may help to anonymize source speaker, but an advanced speaker recognition system may identify the source speaker by learning the minute details of unique voice-identity.

Our work is first, best to our knowledge, elaborated analysis to study the vulnerabilities of identifying source speakers from the various voice conversion and speaker recognition systems to identify the source speakers with focusing on in-depth analysis on probability of individual source speaker, and studying the impact of target speaker gender, accent, and training with the adversarial samples. We analyze parallel, non-parallel datasets in one-to-one, many-to-many, and any-to-any voice conversion methods using statistical, Generative Adversarial Network, and Zero-Shot deep learning neural network models.

### 7.4 Threat Model

In the threat scenario, victim speakers can be spk 1 to spk n interacting with voice inputs in sensitive healthcare scenarios, especially in mental health therapy sessions. Speakers can be sharing sensitive information using direct speaking, using phone or through voice assistants. In most common scenarios, original voice data can be stored on local devices, local databases, and

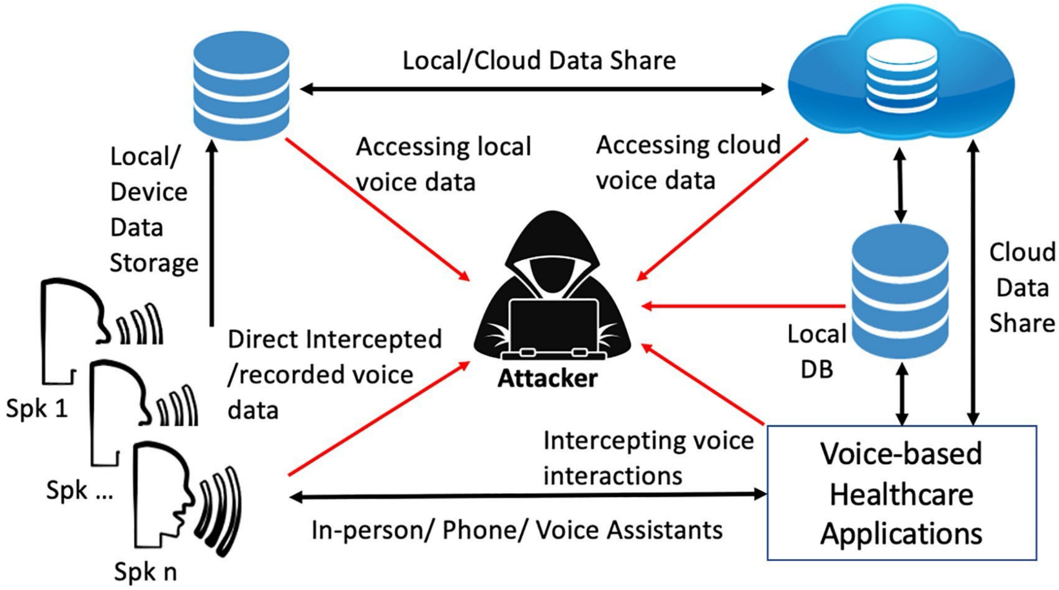

Local/Cloud Data Share
Accessing local voice data
Accessing cloud voice data
Local/ Device Data Storage
Cloud Data Share
Local DB
Attacker
Direct Intercepted /recorded voice data
Spk 1
Spk ...
Spk n
Intercepting voice interactions
In-person/ Phone/ Voice Assistants
Voice-based Healthcare Applications

Figure 7.2: Attacker's Access to Original Voice Data

in the cloud. It is assumed that an attacker can get access to original voice samples through the local and cloud storage and also from intercepting and recording the voice interactions. Figure 7.2 shows that an attackers may have the ability to collect original voice samples either through public access or through exploiting communication channels and data storage. In an enhanced security environment, speakers' voice samples can be anonymized before sharing with the other party and storing locally or in the cloud. Figure 7.3 shows the attacker's accessibility to sensitive healthcare voice samples from local or cloud voice-records for a large number of speakers. In this scenario, the attacker has access to only anonymized voice samples. An attacker can be informed on used anonymization process for implementing apt de-anonymization process with greater success rate. Once the attacker has access to the original/anonymized voice samples, attacker can label them with a speaker's real or made-up identity. The attacker can also generate relevant VC anonymized samples for these original samples to train SR model. Figure 7.4 shows the high-level threat of breaking the voice anonymity of source speakers using off-the-shelf speaker recognition systems. The threat of identifying the source speaker from the voice-converted samples allows the adversary to design and execute targeted privacy and security attacks [281, 280].

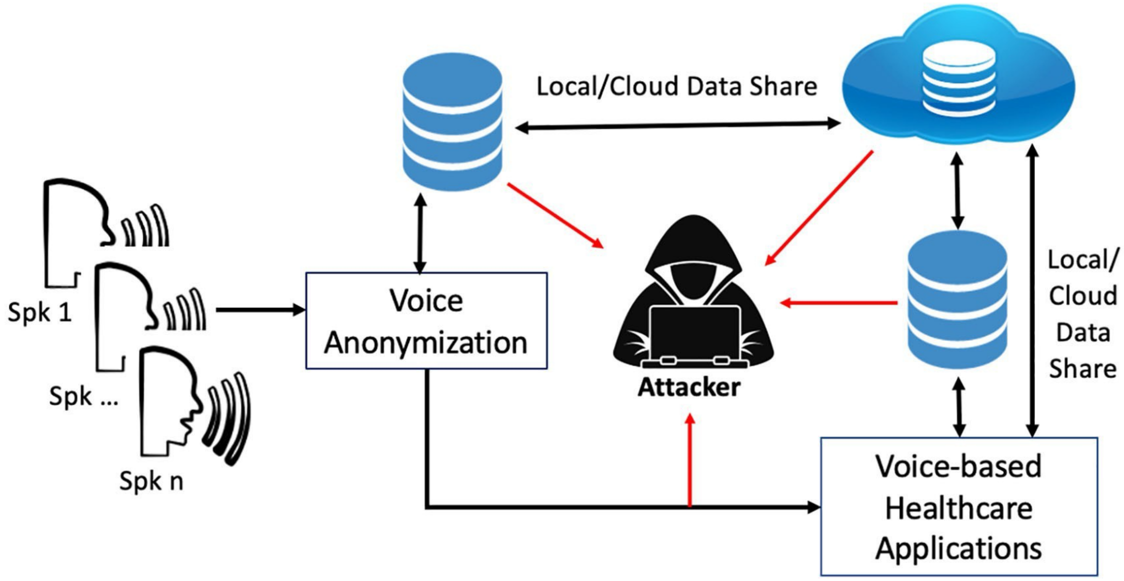


Figure 7.3: Attacker's Access to Anonymized Voice Data

We assume that adversary may or may not know the original voice conversion method to anonymize the voice samples. However, adversaries have unlimited resources otherwise to attain the unlimited amounts of victim's original voice samples. The adversaries have a knowledge of voice conversion methods to generate additional voice samples. State-of-the-art speaker recognition systems are available to the adversaries to identify the source speakers from the converted voice samples.

## 7.5 Study Design and Objectives

To demonstrate the broad applicability of proposed attacks with different fundamental VC approaches, we consider one-to-one, many-to-many, and any-to-any VC methods with parallel and non-parallel voice data. We propose that if converted samples do not preserve speaker anonymity perfectly, and if some source speakers can be identified (at least better than random guessing), a possibility of breaking voice anonymity exists. In this work, we focus on three attacks: (i) Voice de-anonymization attacks against probable set of speakers with the availability of original voice samples (ii) Voice de-anonymization attacks without availability of the original voice samples,

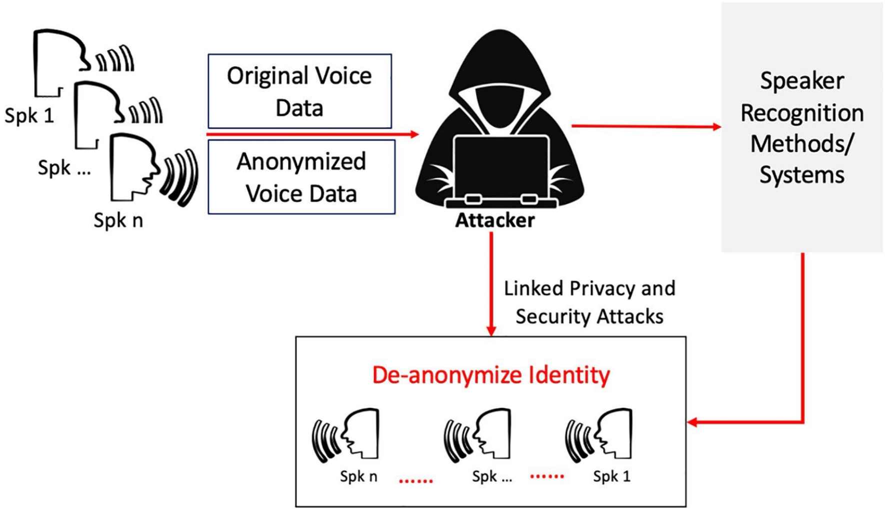


Figure 7.4: Breaching the Voice Anonymity Using Speaker Recognition

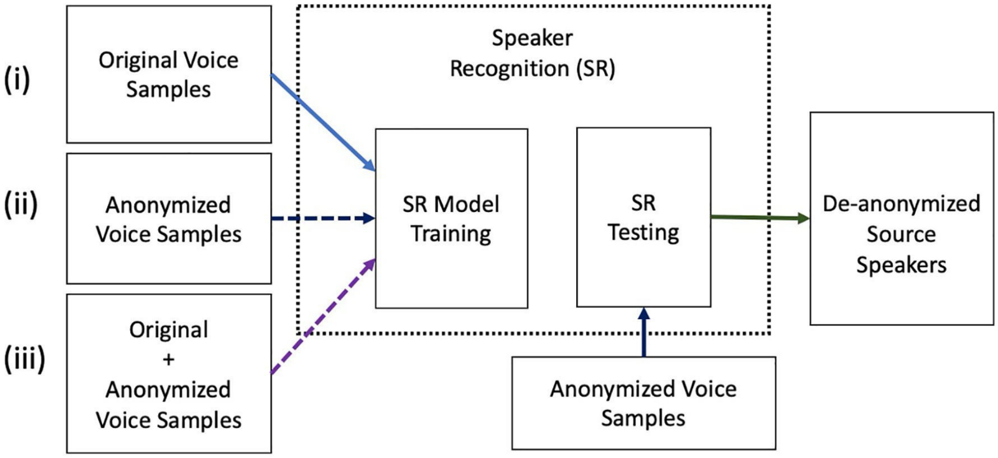


Figure 7.5: Attack Scenarios to Breach the Voice Anonymity: Case(i)- SR Model Training with the original voice samples; Case (ii)- SR Model Training with VC anonymized voice samples; and Case (iii)- SR Model Training with the original and VC anonymized voice samples

and (iii) Voice de-anonymization attacks with adversarial SR training. Figure 7.5 depicts these attacks to breach the speaker's anonymity. All three scenarios are independent of each other, and are shown as separate cases for SR model training. SR testing is done with the VC anonymized voice samples in all cases. However, the test voice samples are relevant to the scenario regarding the male/female target voice, same/different accent or seen/unseen converted voice samples. The proposed attacking scenarios are designed and intended to identify source speakers from the voice conversion-based anonymized voice samples.

A seen speaker or the utterance is the one which is used in the VC model training and then used to create the voice-converted samples. An unseen speaker or the utterance is one which is not used in the VC model training but used in generating voice converted samples. In the context of speaker recognition, speakers are seen in SR model training and testing, but utterances are seen and unseen for SR testing to verify the vulnerability of identifying the source speakers from the voice converted samples. The variation with number of speakers, training data for VC and SR, and seen and unseen speakers and utterances cover potential practical real-life scenarios. The adversarial SR training may model learn better to classify and predict. We study multiple voice-conversion methods to alter the speaker's identity and then multiple speaker recognition methods are applied to uncover the identity. We also address the relative threat of changing scenarios in the field of voice manipulation and speaker recognition advancements by considering multiple systems.

The following sections provide details about the utilized voice conversion methods, speaker recognition methods, and the voice datasets for this work.

### 7.5.1 Voice Conversion Methods

#### *7.5.1.1 sprocket VC*

Open-source VC sprocket applies a trajectory-based conversion method using a Gaussian mixture model (GMM). The one-to-one parallel voice conversion (VC) system known as sprocket applies statistical voice conversion techniques for acoustic feature extraction, time alignment between the source and target voice features, Gaussian mixture model (GMM) training, feature conversion, and waveform generation. When performing voice conversion, speaker-dependent statistics such as the mean and standard deviation of fundamental frequency (F0) and the global variance (GV) of Mel-frequency cepstral coefficients (MFCCs) are typically calculated. Joint feature vectors are constructed for the GMM modeling, the iterative time-alignment estimation is performed using DTW, GMM modeling, and conversion of the source feature vectors. The GMM for the mel-cepstrum is trained and the GV of the converted feature vector is calculated using the trained GMM to convert source speaker to the target speaker [264].

#### *7.5.1.2 StarGAN-VC*

StarGAN-VC is a non-parallel many-to-many VC, a variant of a generative adversarial network (GAN) called StarGAN, which does not require parallel utterances, transcriptions, or time alignment procedures for speech generator training. It can learn many-to-many mappings across different attribute domains using a single generator network. It can generate converted speech signals quickly enough to allow real-time implementations, and it can work with only several minutes of training voice samples to generate reasonably realistic sounding speech [267].

A generator takes an acoustic feature sequence instead of a single-frame acoustic feature as an input and outputs an acoustic feature sequence of the same length obtaining conversion rules that capture time dependencies using convolutional neural network (CNN)-based architecture. For each utterance, a spectral envelope, a logarithmic fundamental frequency (log F0), and aperiodicities (APs) are extracted, and 36 mel-cepstral coefficients (MCCs) are extracted from each spectral envelope. The F0 contours were converted using the logarithm Gaussian normalized transformation. Network architectures of generator G, real/fake discriminator D and domain classifier C. All the networks are fully convolutional with no fully connected layers, thus allowing inputs to have arbitrary sizes [267].

#### *7.5.1.3 VQMIVC: Vector Quantization and Mutual Information-Based VC*

VQMIVC is unsupervised speech representation disentanglement for One-Shot voice conversion. During inference, one-shot VC is achieved by only replacing the source speaker representation with the target speaker representation derived from a single target utterance. As demonstrated in the work, it mitigates information leakage issues by learning accurate content representation to preserve source linguistic content, speaker representation to capture desired speaker characteristics, and pitch representation to retain source intonation variations, producing high-quality converted voice [282].

### 7.5.2 Speaker Recognition Methods

#### *7.5.2.1 Keras Speaker Recognition*

A generalized Convolutional Neural Network (CNN) used in Keras SR tools to identify the source speakers. Input samples get analyzed by the convolutional layers to extract the features and pooling does the feature mapping. The last layer does the classification to predict the output class.

#### *7.5.2.2 Resemblyzer*

Resemblyzer can provide a high-level representation of a voice through a deep learning model, a voice encoder. It creates a summary vector of 256 values of a given audio clip that summarizes the characteristics of the voice spoken [283, 284]. Resemblyzer can generate voice similarity metric to compare different voices and get a value on how similar they sound which can assist in Speaker verification, Speaker diarization, and Fake speech detection. High-level feature extraction can be done as well through Resemblyzer by using embeddings generated as feature vectors for machine learning or data analysis [283].

### 7.5.3 Voice Datasets

#### *7.5.3.1 VCTK Corpus*

This CSTR VCTK Corpus includes speech data uttered by 109 English speakers with various accents. Each speaker reads out about 400 sentences [285, 286]. The VCTK corpus consists of 44 hours of speech data uttered by 109 native speakers of English with various accents. This corpus was originally aimed for HMM-based text-to-speech synthesis systems, especially for speaker-adaptive HMM-based speech synthesis that uses average voice models trained on multiple speakers and speaker adaptation technologies. This corpus is also suitable for DNN-based multi-speaker text-to-speech synthesis systems and neural waveform modeling.

#### *7.5.3.2 Voice Conversion Challenge (VCC) Dataset*

The Voice Conversion Challenge (VCC) was launched at Interspeech 2016. The voice samples were recorded by professional US English speakers in a professional recording studio without significant noise effects. There were five source speakers (3 F and 2 M) and five target speakers (2 F and 3 M), manually segmented into 216 utterances in each speaker, down-sampled to 16 kHz [279]. VCC 2016 dataset was developed from DAPS (Data And Production Speech) dataset which also has been used in many research implementations [287].

## 7.6 Preliminaries: Resources and Setup

The study environment was built using MacOS and Linux OS. VC Model training was carried out on High Performance Research Cluster (HPRC) and Ubuntu 20.02.4 LTS machine with Intel Xeon 3.6 GHz, 32 GB of RAM, and Nvidia RTX A500 GPU. Publicly available Github repositories, demo scripts, and web-based tools are utilized in studied voice conversion methods. Keras and Resemblyzer, deep learning methods, are applied for Speaker Recognition (SR) experiments. Most of Keras speaker recognition and Resemblyzer experiments were carried out on MacOS. Audacity 3.0.5 and FlicFlac 1.10 audio converters are used to prepare voice samples for modifying sampling rate and the audio file format. The sampling frequency was set to 16000 Hz for .wav format voice samples.

### 7.6.1 Voice Conversion Experiments

We considered multiple voice conversion methods as described in section 3.1 using multiple databases. One-to-One VC is adopted from sprocket Github implementation [288], and Any-to-Any voice conversion was done using their VQMIVC online tool [289]. A PyTorch variant of StarGAN-Voice-Conversion [267] is adopted from the Github [290], is primary Many-to-Many VC method used in majority of the experiments. VC model was trained for 70 VCTK speakers

VQMIVC (`https://replicate.com/wendison/vqmivc`)

Table 7.1: Voice Conversion Scenarios

| VC Scenario # | Source Speaker Type | Target Speaker Type |
|---|---|---|
| M/F-to-F | Seen | Seen |
| M/F-to-M | Seen | Seen |
| M/F-to-F | Seen | Unseen |
| M/F-to-M | Seen | Unseen |
| M/F-to-F | Unseen | Seen |
| M/F-to-M | Unseen | Seen |
| MF-to-F | Seen-Similar Accent | Seen-Different Accent |

using approximately 28000 voice samples (avg. 400 samples for each speaker). VC model training was extended with merging last two versions of VCTK Corpus for generating enough converted seen and unseen samples for SR training and testing. Extended dataset provided approximately 63000 voice samples (avg. 900 samples for each speaker). For sprocket VC, the model was trained on the VCC dataset with 864 samples and VCTK with 120 samples. We used around 100,000 voice samples for around 80 speakers for VC model training and generated around 10000 voice samples for SR model training and testing in multiple scenarios.

Speakers and utterances used in testing are considered as *seen* if used in training as well. Most of the target speakers are not part of SR training or testing but sprocket VC has seen target male and female speakers in SR training and testing for VCTK samples. Table 7.1 summarizes multiple scenarios to generated converted voice samples. In each scenario, we assess the breaking source speakers' anonymity through attackers that leverage state-of-the-art speaker verification techniques.

### 7.6.2 Source Speaker Identification through Speaker Recognition (SR)

To test the anonymity-preserving characteristics of voice conversion, we applied two different speaker recognition methods: a Keras convolutional neural network (CNN) based method and Resemblyzer. We found that both speaker recognition systems were able to identify the source speakers from the converted voice samples, even when tested with both seen and unseen utterances. We conducted our testing by first establishing a baseline observation with original voice

Table 7.2: Speaker Recognition Evaluation Scenarios

| SR Scenario # | Training | Testing | Utterances in VC |
|---|---|---|---|
| Benign | Original | Original | Seen |
| Conv | Original | Conv | Seen |
| Conv | Original | Conv | Unseen |
| Conv-Conv | Conv | Conv | Seen |
| Conv-Conv | Conv | Conv | Unseen |
| Conv-Conv | Conv | Conv | Both |
| Conv-Acc | Original | Conv | Seen |
| Conv-Adv | Original+Conv | Conv | Seen |

samples in a benign setting, and then repeating the testing with converted samples generated for each voice conversion scenario. We also trained the speaker recognition model with both seen and unseen converted samples, and then tested the model with both seen and unseen voice samples. Additionally, we performed adversarial training by training the speaker recognition model with a combination of original and converted samples, and then tested the model with converted samples. We also trained the Keras speaker recognition model with ten and twenty unseen VCTK speaker utterances, both original and converted, as part of an adversarial setting. However, due to the low accuracy of the model (only 25% to 35%), we did not test these scenarios for source speaker identification. All other SR testing scenarios were executed with the model with an accuracy of 75% or above.

Table 7.2 shows the training and testing scenarios in primary SR method for source speaker identification. All converted and tested samples were labeled as source speaker to get the identification probability and confusion matrix for the scenario. The majority of SR experiments are executed with 100 test samples and with 10 iterations to get the average predictions. With a larger set of speakers (10, 20, 30), test sample was 1000 and average was considered from 5 iterations.

## 7.7 Results

We present the findings of voice conversion and speaker recognition evaluations for VCTK and VCC datasets in the following sections, as observed by different scenarios outlined in Table 7.1

and Table 7.2. Though all scenarios are not replicated for each of the datasets, VC type and SR type, conducted experiments confirm the possibility of breaching the voice anonymity of source speakers in almost all verified scenarios. To evaluate the efficacy of utilized SR methods, we run original voice samples from both datasets in benign settings.

### 7.7.1 Baseline

We first observed the baseline case in benign settings for VCTK and VCC datasets. Table 7.3 shows that, for the VCTK dataset, speaker recognition tool, Keras SR as SR1, shows model accuracy of above 91% and validation accuracy of 97%. An average prediction accuracy is 100% with seen samples and 89% with unseen samples. For VCC dataset, Keras SR has a model accuracy of 98% and a validation accuracy of 96%. VCC dataset shows an average prediction of 96% with seen and 88% with unseen utterances. Resemblyzer as second Speaker Recognition tool evaluates the voice samples by comparing voice embeddings. Appendix Figure B.1 shows clear identity of speakers' embeddings in benign setting for both the voice datasets. Resemblyzer compares different voices and gets a value on how similar they sound. Appendix Figure B.1 shows that for VCTK and VCC datasets benign setting, similarity median for same speaker is around 99% for same speaker in comparison to 70% for different speakers. Utterances similarity median is observed around 85% for similar speakers and 55% for different speakers.

Table 7.3: Voice Datasets- Keras SR- Benign Settings

| | VCTK Seen | VTCK Unseen | VCC Seen | VCC Unseen |
|---|---|---|---|---|
| Sample Size | 2800 | 2560 | 1296 | 696 |
| Training Sample Size | 2520 | 2304 | 1167 | 627 |
| Validation Sample size | 280 | 256 | 129 | 69 |
| Model Accuracy (%) | 91 | 93 | 98 | 98 |
| Validation Accuracy (%) | 97 | 98 | 97 | 96 |
| Testing Sample Size | 100 | 100 | 100 | 100 |
| Avg. Prediction Accuracy (%) | 97 | 80 | 96 | 88 |

Table 7.4: Keras SR Model Accuracy for Scenarios evaluated in Table 7.5

| Scenario | Total Samples | SR Model accuracy % | Val Acc % |
|---|---|---|---|
| Seen M/F converted with unseen M target spk | 2011 | 90.55 | 87.06 |
| Seen M/F converted with unseen F target spk | 2011 | 93.48 | 91.54 |
| Unseen M/F converted with seen M target spk | 2084 | 94.08 | 93.27 |
| Unseen M/F converted with seen F target spk | 2084 | 94.3 | 88.46 |
| VC Seen spks, seen utterances-M/F converted to seen M spk | 3029 | 85.44 | 80.46 |
| VC Seen spks, seen utterances-M/F converted to seen F spk | 3029 | 76.42 | 77.48 |
| Converted with F target spk of similar accent | 2097 | 90.36 | 89.00 |
| Converted with F target spk of different accent | 2097 | 90.89 | 86.60 |

### 7.7.2 SR Evaluation of StarGAN-VC Samples

With StarGAN VC, VCTK, and SR testing shows better source speaker identification with seen utterances and seen target speakers in comparison to unseen utterances and unseen target speakers. Average overall source speaker identification is higher than random guessing in most of the tested scenarios. We can see that random guessing is 20% for 5 speakers, 10% for 10 speakers and 20% for 5 speakers. All testing scenarios show that at least one source speaker gets identified. We analyze the source speaker identification vulnerability for each individual speaker, we utilized the confusion matrix and calculated the identification percentage from the samples identified divided by the samples tested for that individual speaker. Table 7.5 shows that identification percentage of an individual source speaker is as high as around 93%.

VC samples with male target voices have higher probability of identifying source speakers in multiple SR scenarios. We observed that Keras SR identified all 5 of 5 source speakers, using male target voice and seen utterances with adversarial SR training more than random guessing (max of 77.77%) and thus pose potential threat of breaching the voice anonymity. Adversarial testing was extended with 10 and 20 VCTK speakers with SR training of original and VC samples. The results show higher number of identifying the source speakers from a pool of speakers tested.

Table 7.5: StarGAN VC with VCTK Speakers- Keras SR Source Speaker Identification Vulnerability

| SR Training | SR Testing | SR testing (Utterances-seen/unseen) | VC Target Spk | Identifiable source Spks | Identifiable source Spks (Gender) | Most vulnerable Source Spk gender, % |
|---|---|---|---|---|---|---|
| Original | Conv | Unseen | M | 5 out of 10 | 2 F, 3 M | M, 55.939 |
| Original | Conv | Unseen | F | 5 out of 10 | 3 F, 2 M | M, 70.11 |
| Original | Conv | Unseen | M | 7 out of 20 | 2 F, 5 M | M, 68.1 |
| Original | Conv | Unseen | F | 4 out of 20 | 2 F, 2 M | M, 38.78 |
| Original | Conv | Seen | M | 3 out of 5 | 1 F, 2 M | F, 51.76 |
| Original | Conv | Seen | F | 1 out of 5 | 1 M | M, 80.67 |
| Conv | Conv | Unseen | M | 5 out of 10 | 2 F, 3 M | M, 75 |
| Conv | Conv | Seen | M | 4 out of 5 | 2 F, 2 M | F, 82.28 |
| Conv | Conv | Unseen | F | 4 out of 10 | 2 F, 2 M | F, 45.83 |
| Conv | Conv | Seen | F | 2 out of 5 | 2 F | F, 77.97 |
| Original + Conv | Conv | Seen | M | 5 out of 5 | 3 F, 2 M | F, 77.77 |
| Original + Conv | Conv | Seen | F | 3 out of 5 | 3 F | F, 78.65 |
| Original + Conv | Conv | Unseen | M | 10 out of 10 | 5 F, 5 M | M, 77.91 |
| Original + Conv | Conv | Unseen | F | 10 out of 10 | 5 F, 5 M | F, 93.36 |
| Original + Conv | Conv | Unseen | M | 19 out of 20 | 9 F, 10 M | M, 68.57 |
| Original + Conv | Conv | Unseen | F | 19 out of 20 | 9 F, 10 M | F, 79.41 |

Another attacking scenario of SR training with converted samples identified 4 out of 5 source speakers with a max of 82.28%. Test scenarios with male target voice could identify 54% of source speakers, while average source spk identification in scenarios using female target voice is around 41%. We also extended SR training with only converted samples for 10 and 20 VCTK speakers

Table 7.6: Keras SR Source Spk Identification with multiple sets of VCTK speakers with StarGAN VC

| **Source spk#** | **Target spk** | **Keras SR Training- Testing- Utterances** | **Identifiable Source spk #out of total** | **Identifiable Source spk** | **Most Vulnerable Source spk, Gender, %** |
|---|---|---|---|---|---|
| 5 F, 5 M | F | Orig-Conv-Seen | 3 out of 10 | 2 F, 1 M | p236, F, 42.86 |
| 10 F, 10 M | F | Orig-Conv-Seen | 6 out of 20 | 3 F, 3 M | p236, F, 71.31 |
| 15 F, 15 M | F | Orig-Conv-Seen | 8 out of 30 | 4 F, 4 M | p236, F, 69.04 |
| 5 F, 5 M | M | Orig-Conv-Seen | 3 out of 10 | 1 F, 2 M | p284, M, 77.77 |
| 10 F, 10 M | M | Orig-Conv-Seen | 7 out of 20 | 1 F, 6 M | p298, M, 47.82 |
| 15 F, 15 M | M | Orig-Conv-Seen | 9 out of 30 | 8 F, 1 M | p236, F, 72.34 |

Table 7.7: Keras SR with StarGAN VC using Unseen/Seen Source and Target Spks

| VC Target Spk Gender | VC Source Spk | VC Target Spk | Identifiable Source Spks | Most vulnerable Source Spk gender, % |
|---|---|---|---|---|
| M | Seen | Unseen | 2 of 5 | M, 95.72 |
| F | Seen | Unseen | 2 of 5 | M, 92.07 |
| M | Unseen | Seen | 1 of 5 | M, 72.72 |
| F | Unseen | Seen | 2 of 5 | F, 87.22 |

but the model accuracy was found comparatively much lower than rest of the other test scenarios. For example, model accuracy was from 25% to 37% in all four scenarios to train with M/F target voice converted sample for 10 and 20 speakers.

We also measured Keras SR performance with varied number of speakers in sets of 10, 20, and 30 speakers to train with original samples and test with converted samples. Table 7.6 shows that male target voice VC samples could identify more source speakers in comparison to female target voice in similar settings. A male source speaker was most vulnerable with 77.77% using male target voice in a set of 10 (5 M, 5 F) VCTK speakers.

We also evaluated another scenario where VC source speakers or target speakers were unseen during VC model training. Table 7.7 shows vulnerability to identifying at least one source speaker from a set of five speakers. Keras SR was trained with the original samples of these seen/unseen source speakers in the VC process, and SR testing was performed with the converted samples using seen/unseen target speakers during VC. In this case, we can see that male speakers are more vulnerable with the identifying probability of as high as 95.72%. It is also observed that unseen target speaker performs better than unseen source speakers in identifying the number of source speakers and the identifying percentage.

#### *7.7.2.1 Resemblyzer- Speakers' Embeddings and Cross-Similarity*

VC samples are analyzed through Resemblyzer to observe the patterns in speakers' embeddings projection and the cross-similarity between same and different speakers. We compared converted voice samples using female and male target voice. Figure 7.7 shows the converted samples with target male voice, labeled as source speakers and Figure 7.6 shows the speakers' embeddings of converted samples with the female target voice. It is clearly shown that conversion with target male voice is more vulnerable of breaking the source speaker's anonymity as forming more identifiable clusters. Appendix Figure B.2 shows that for female target voice converted samples, utterance cross-similarity is around 80%, and speaker cross-similarity is around 98% for different speakers which makes them less distinguishing from each other. As per Appendix Figure B.3, using male target voice is more distinguishable with utterance cross-similarity of 70% and speaker cross-similarity of around 90% for different speakers.

### 7.7.3 SR Evaluation of sprocket VC Samples

sprocket VC was tested with VCTK and VCC speakers in Keras SR system. The sample size was kept at 100 as consistent and took an average of 5 iterations. For VCC dataset, with male target speakers and seen utterances, Keras SR could identify 3 out of 4 source speakers above random guessing of 25% and identified 1 of 2 male source speakers with 96%. For female target voice, SR identified 2 out of 4 source speakers (1 F, 1 M) with the max of 66% identification of a female source speaker. With VCTK speakers, 1 out of 4 source speakers was identified with 78.66%, using female target speakers. Two out of four source speakers (1 M, 1 F) were identified with male target speaker with a max of with 55%.

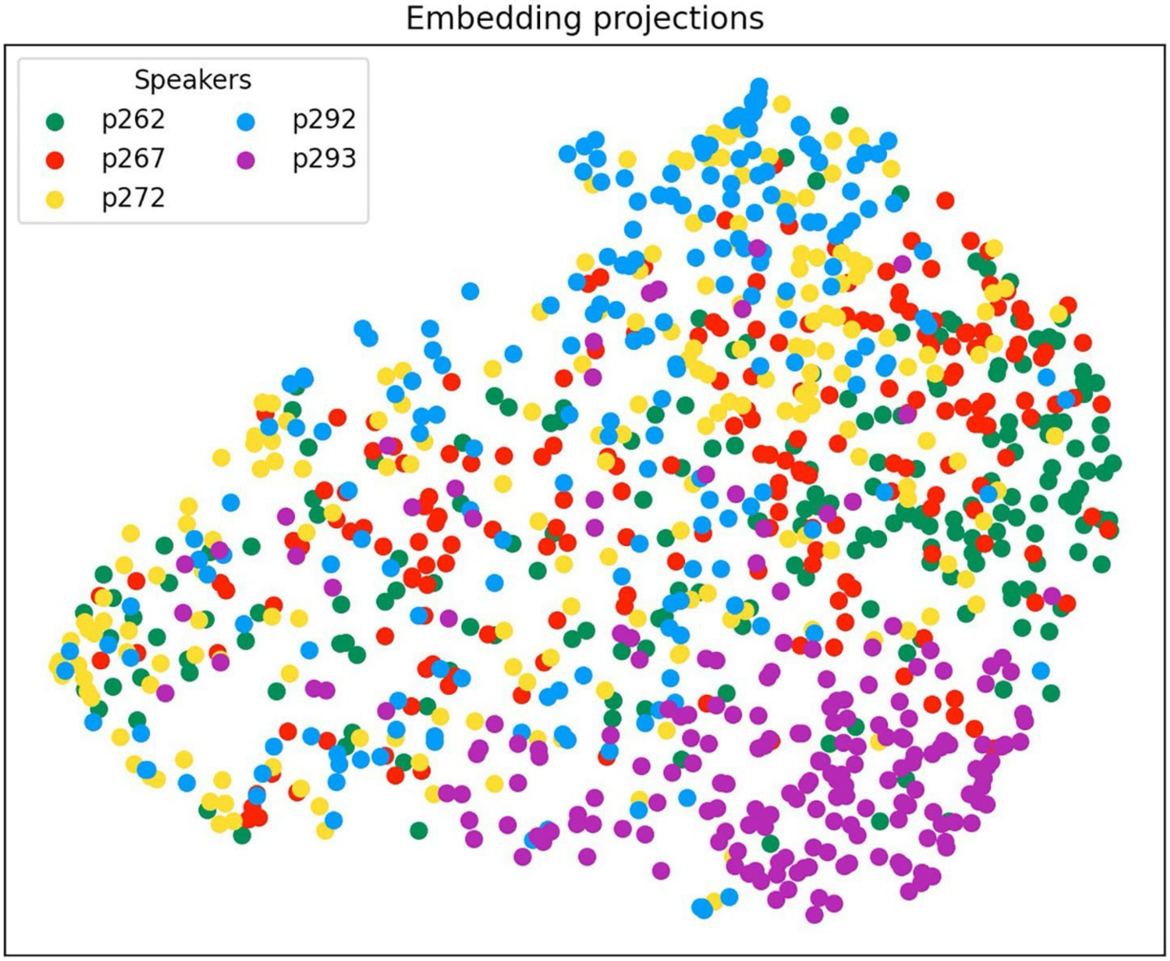


Figure 7.6: Resemblyzer: StarGAN-VC- Target spk-F

Table 7.8 presents different scenarios with male/female target voice converted samples and seen/unseen utterances in SR testing. In these experiments, with same gender for source and target speakers in VC, showed higher percentage for the source speaker identification.

#### *7.7.3.1 Resemblyzer- Speakers' Embeddings and Cross-Similarity*

For Resemblyzer evaluation, Appendix Figure B.4 shows the speaker's embeddings of converted voice samples with female target speaker, which shows overlapping between both male source speakers. However, both female source speakers present more dense and isolated clusters. Speakers' embeddings generated for converted samples with male target voice, in Appendix Figure B.5, show loosely clustered but identifiable separation of source speakers.

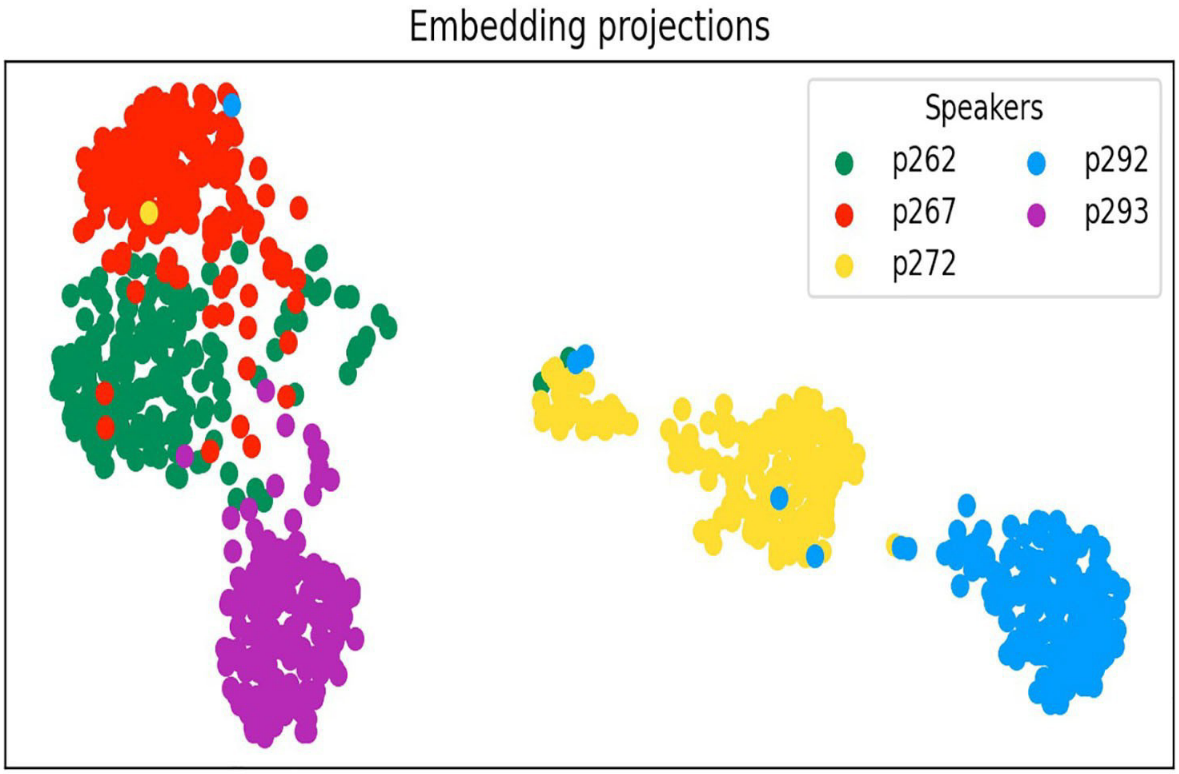


Figure 7.7: Resemblyzer: StarGAN-VC- Target spk-M

With male target voice, sprocket VC samples show utterance cross-similarity of 86% for same speaker and 75% for different speakers. Speaker cross-similarity is 99% for same speaker and 93% for different speakers. For female target voice, Resemblyzer shows lower percentage in comparison of male target voice which means that speaker identification is better with female target voice.

We see that finding of both, Resemblyzer and Keras, SR systems, support each other, and sprocket VC seems more vulnerable to speaker anonymization with female target voice. However, testing with male target voice identified more source speakers with a lower percentage in comparison to female source speaker identification percentage.

### 7.7.4 SR Evaluation for VQMIVC Samples

VQMIVC is more recent VC with any-to-any conversion and has the lowest probability of identifying source speakers from the converted voice samples. Four Male and Female source speakers were converted with a male/female target voice. There were 25 utterances, randomly picked and

Table 7.8: Keras SR Evaluation for VC sprocket with VCTK and VCC Datasets

| Voice Dataset | SR Training | SR Testing | SR Testing Utterances (seen/ unseen) | VC Target Speaker (M/F) | Identifiable Source Speakers | Identifiable Source Speaker Type (M/F) | Most Vulnerable Source Speaker, Gender, % |
|---|---|---|---|---|---|---|---|
| VCTK | Original | Converted | Seen | F | 1 out of 4 | 1F | P250, F, 78.66 |
| VCTK | Original | Converted | Seen | M | 2 out of 4 | 1 F, 1 M | P246, M, 55.39 |
| VCC | Original | Converted | Seen | F | 2 out of 4 | 1 F, 1 M | SF2, F, 66.15 |
| VCC | Original | Converted | Seen | M | 3 out of 4 | 1 F, 2 M | SM2, M, 96 |
| VCC | Original | Converted | Unseen | F | 2 out of 4 | 2 F | SF1, F, 83.48 |
| VCC | Original | Converted | Unseen | M | 1 out of 4 | 1 M | SM1, M, 78.39 |

converted from the sentence pool for VCTK speakers. Results show that using male target voice, male source speaker identification is 24.62%, which indicates some possibility of anonymity attack. Another observation with using female target voice, a female source speaker identification is 34.87% higher than the male source speaker and above the random guessing. VQMIVC converted samples in Keras SR testing showed better probability of identifying source speaker when target voice gender was same as the source speaker. In studied scenarios, VQMIVC converted samples using female target voice are more vulnerable to breach the speaker's anonymity. However, identification percentage is comparatively low with all other observed results with other studied VC methods and SR results.

For Resemblyzer evaluation of VQMIVC converted samples shows that with both male and female target voice, there is no clear identifiable clustering for each speaker's samples. However, there is a bit more possibility of identifying source speakers with female target voice. Appendix Figure B.6 and Figure B.7 show that female target voice conversion has somewhat higher probability of breaking source speaker's anonymity in comparison to male target voice VC samples.

### 7.7.5 Summary of Results

Our evaluation of multiple voice datasets, VC, and SR methods revealed that male target speaker converted samples are more vulnerable to breaking source speaker anonymity. Adversarial SR training with male target voice converted samples produced the best results in identifying source speakers. Voice conversion with unseen speakers and utterances was less vulnerable compared to that with seen speakers and utterances. Additionally, voice conversion with similar or different accents did not show any significant difference in source speaker identification.

From all different tested scenarios, we see that converted voice samples may not hold the source speaker's anonymity completely and the probability of breaking the speaker's anonymity is dependent on the VC method, SR method, and the training data. Most of the results validate the source speaker identification above random guessing, which presents a vulnerability to breach the speaker's anonymity. It is also observed that different voice conversion methods and speaker recognition methods may vary from low to high probability to identify the individual source speaker but overall, there is a consistent threat of identifying the source speakers from the voice-converted samples. Results demonstrate the possibility of breaching the source speaker's anonymity from the converted voice samples. Once the identity is predicted, there is a higher probability of launching linked privacy and security attacks. Results also demonstrate that a source speaker can be converted into multiple target speakers in inter-gender or intra-gender settings with different VC techniques to maintain k-anonymity, but even off-the-shelf available speaker recognition systems have the reasonable ability to identify the source speaker.

## 7.8 Challenges and Discussion

Voice conversion and speaker recognition, both are evolving areas, and an empirical study to observe the changes in the vulnerabilities of breaking the source speaker's anonymity can be more useful in developing future defense mechanisms to minimize the threat. To broaden the analysis, we evaluated a couple of existing VC systems with a multiple voice datasets and SR approaches. Therefore, we have a reasonable confidence in our findings. Training and retraining the huge speaker voice dataset can be time consuming, even with utilizing the HPRC. Pre-trained models can help to improve the time requirements in some cases but using a different dataset need the retraining. Data set samples may not be consistent for each speaker and parallel data requirements for some tools may be challenging. Though more recent VC methods are progressing towards zero-shot learning and may work well with limited set of voice samples, it is vital to maintain the model training accuracy and balanced data classes of voice samples.

From the diverse set of VC and SR testing scenarios, we observed that studied SR model accuracy may get impacted negatively with a much larger pool of speakers. These off-the-shelf SR tools may be more efficient with certain parameters like number of classes to train with, the quality of the voice samples, and the VC method used to obtain or create voice converted samples. However, it is still feasible to train these ready-to-use SR models with a fewer speakers separately, and achieve a higher confidence in identifying the source speakers.

Voice-based healthcare solutions are still seeking broader practical implementation, especially because of data sensitivity involved. For sensitive voice records, protecting the speaker's anonymity is critical to maintain the trust. However, it is expected that emerging advanced methods may cause even greater threat to identifying the source speakers from anonymized voice samples. Future work to reproduce the vulnerabilities with more advanced voice conversion and speaker recognition tools can further verify the threat and its extension with the emerging technologies in the voice domain.

### 7.9 Conclusion

In a broader perspective, our work concludes the multiple existing voice conversion methods cannot protect the source speaker's anonymity. A failure of protecting the voice-privacy can identify the speaker which can lead to many real-life inconveniences and even physical or mental harassment. Voice conversion and privacy-protecting solutions also suffer from maintaining the naturalness and intelligibility of the converted voice samples. Our study shows the vulnerability of breaking the source speaker's anonymity in studied voice conversion through multiple speaker recognition systems. This can be a bigger problem in sensitive information sharing in healthcare settings, especially in mental health cases or recovering addicts. It can be upsetting to breach the speaker's privacy itself but it can trigger targeted attacks by linking other physical, financial information, and health information. These privacy and security attacks can aggravate the mental stress further and these can lead to even more dangerous outcomes for physically disabled speakers.

The study shows that the vulnerability of breaking the source speaker's anonymity in voice conversion methods can lead to privacy and security attacks, which can be especially dangerous for vulnerable populations. Therefore, it is important to identify the threats and take appropriate measures to protect the privacy and security of vulnerable populations who may be at higher risk. It is vital to pay more attention to ensure availability of private and secure communication channels for voice data exchanges, especially for healthcare sensitive data. Healthcare provider's voice-notes or patients' voice-inputs also need greater attention for secure voice-data storage solutions where data can stay unlinked and cannot be exploited for linked privacy and security attacks.

# 8. KEY CONTRIBUTIONS AND TAKEAWAYS

Identifying vulnerabilities of emerging healthcare technologies is challenging in rapidly evolving environment. Frequent monitoring and re-evaluation of the potential threats and defenses is necessary to minimize the potential harm caused by privacy and security vulnerabilities present in each of interconnected and interdependent components of modern medicine practices. For example, on diagnoses and treatments, medical AI integrity is crucial, and the treatment cycle is involving mobile healthcare apps and the voice-based interactions and voice healthcare data. As mobile health apps and voice-anonymity can be subject to threat of establishing the user's identity, it can present an extended threat to connect with user's health conditions, criticality of the diseases or symptoms, mental illness and/or reproductive choices. This dissertation work focuses on the vulnerabilities and potential defense strategies of three domains (Research Literature-Based Medical AI Domain, Mobile Health Apps Domain, and Voice Domain). We present the contributions and key takeaways for each domain and then the overall.

## 8.1 Medical AI Domain: Research Literature Dependent MedAI

- Presented the critical threat of predatory research impacting real-world Med AI solutions.
- Verified the traversal of predatory research from predatory sources to MedAI output.
- Proposed the defense by classifying predatory research.
- Without efficient defense mechanism, targeted attacks may do greater harm.

## 8.2 Mobile Health App Domain: Mental Health and Reproductive Health Apps

- Mental health apps show the threats of being intrusive and worsening the mental burden.
- Women's health apps pose serious threats of being identified, monitoring, tracking, and prosecution in changing legal landscape.
- Adhering to secure app development with rigorous pre/ post-market evaluations.

- Without efficient defense mechanism, targeted Heightened risk to vulnerable populations using context-specific sensitive healthcare apps. may do greater harm.

## 8.3 Voice Domain: Breaching Voice Anonymity

- Voice anonymity is crucial to maintain in sensitive voice-based healthcare scenarios.
- Voice-conversion-based anonymity can be potentially breached by using off-the-shelf speaker recognition methods.
- Other data security practices in combination to voice-conversion can improve the anonymity protection.

## 8.4 The overall main contributions and key takeaways

- Emerging healthcare technologies, including Medical AI, mobile healthcare apps, and voice-based solutions, are vulnerable to privacy and security threats that can compromise patient care and trust in their clinical use.
- Exploitation of these vulnerabilities can lead to unauthorized access to sensitive information, privacy and security attacks, and adversarial attacks that manipulate Medical AI performance and outcomes.
- Prioritizing privacy and security is crucial to ensure safe and effective healthcare delivery, especially for vulnerable populations.
- Developing feasible defense strategies are crucial to build the trust in emerging technologies in healthcare and improving overall patient care.

# 9. FUTURE RESEARCH DIRECTIONS

The threat of predatory research impacting MedAI is a pressing concern that demands immediate attention for developing defense strategies. While initial defense strategies have focused on identifying predatory venues, it is vital to expand the scope of predatory research identification via efficient data collection and content-level verification. Collaboration between biomedical and technical experts is key to achieving this. Moreover, exploring the feasibility of simulated targeted attacks will be crucial to understand the potential for harmful research publications to infiltrate reputable or predatory venues and subsequently become part of MedAI solutions.

Another crucial area of research is studying the usability, privacy, and security concerns, as well as functionality preferences of mobile healthcare apps among healthcare providers and patients. Developing a secure prototype app that follows best practices may reveal practical limitations and challenges. Research studies can validate the effectiveness of the app in overcoming practical challenges. Furthermore, we need to explore the risks associated with voice anonymity in healthcare settings. Advanced voice conversion methods and speaker recognition techniques can further validate the threat and assist in developing strategies to address anonymity breaches. Additionally, human studies can provide valuable insights into the concerns and challenges faced by healthcare providers in keeping voice inputs confidential.

In conclusion, a comprehensive user study can illuminate the interconnected nature of the various components in modern healthcare and raise awareness of the benefits and potential threats associated with emerging healthcare technologies.

# 10. SUMMARY AND CONCLUSIONS

This dissertation work analyzes data integrity, privacy, and security threats to emerging healthcare technologies, including research literature-dependent MedAI solutions, mobile healthcare apps, and voice-based solutions. Manipulation of these components can compromise the integrity of technological tools, erode trust in their clinical use, and cause direct and indirect harm to patients, especially vulnerable populations. We present interconnections among these components to deliver comprehensive patient care and emphasize the need for prioritizing privacy and security to ensure the safe and effective delivery of healthcare.

While Medical AI solutions have the potential to extract valuable knowledge from the research literature, they are vulnerable to untargeted and targeted attacks. The *Publish or Perish* phenomenon can lead to unreliable and predatory publications that can be part of MedAI inputs and potentially harm patients. Manipulated input data can result in incorrect diagnoses, misdirected treatment options, and negative aftereffects, ultimately burdening the healthcare system. It is crucial to address the risks associated with predatory publications to ensure the safety and effectiveness of Medical AI solutions.

Digital health management, such as mobile apps and voice-based solutions, can improve healthcare accessibility, but emerging technologies introduce new vulnerabilities that can endanger already vulnerable populations. Our research found significant privacy and security vulnerabilities in mobile healthcare apps, particularly in sensitive scenarios such as mental health and women's health. Additionally, voice-based solutions used in private settings like counseling sessions can be the target of anonymity-breaching attacks. To make sure that digital health management benefits everyone, especially vulnerable populations, we need to prioritize privacy and security defenses.

Our research identified the challenges in adapting trustworthy technologies in real-world scenarios and quantified the risks that arise from the interdependency of these technologies in healthcare decision-making. We found that exploitation of vulnerabilities in emerging healthcare technologies can grant unauthorized access to sensitive information, leading to various privacy and security attacks. These attacks can include adversarial attacks that manipulate medical AI performance and outcomes, ultimately impacting patient care. We also highlight the heightened risk to women's health solutions due to changing legal landscapes, which can subject them to increased scrutiny and penalties for their reproductive healthcare choices.

Our work provides new perspectives on the threat landscape of healthcare, where research and emerging technologies are essential components of modern-day patient care. However, these components also attract malicious attention that can disrupt healthcare processes and resources, and that can have serious implications for disadvantaged, underserved, and vulnerable populations. Identifying and measuring these threats in rapidly evolving technological advancements is challenging. Nevertheless, our findings demonstrate the importance of developing more secure and trustworthy technologies with well-defined standards, rigorous evaluation, and accountability. Doing so can mitigate the risks associated with emerging healthcare technologies and ensure safe and effective healthcare delivery for all.

# REFERENCES

[1] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[2] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, "Systematic poisoning attacks on and defenses for machine learning in healthcare," *IEEE journal of biomedical and health informatics*, vol. 19, no. 6, pp. 1893–1905, 2014.

[3] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pp. 531–548, 2018.

[4] FDA, "Artificial intelligence and machine learning (ai/ml)-enabled medical devices," Oct 2022. `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices`, Accessed online on 03/21/2023.

[5] F. permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems, April 2018. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye.

[6] FDA, "Fda allows marketing of first direct-to-consumer app for contraceptive use to prevent pregnancy," 2018. `https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-app-contraceptive-use-prevent-pregnancy`.

[7] Shetty, "A promising step forward for predicting lung cancer," May 2019. `https://www.blog.google/technology/health/lung-cancer-prediction/`.

[8] A. Ramesh, C. Kambhampati, J. R. Monson, and P. Drew, "Artificial intelligence in medicine.," *Annals of the Royal College of Surgeons of England*, vol. 86, no. 5, p. 334, 2004.

[9] Y. Zoabi, O. Kehat, D. Lahav, A. Weiss-Meilik, A. Adler, and N. Shomron, "Predicting bloodstream infection outcome using machine learning," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[10] S. Smith, "230 startups using artificial intelligence in drug discovery," 2017. `https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery`.

[11] Y.-Z. Tu, Y.-T. Chang, H.-Y. Chiou, K. Lai, *et al.*, "The effects of continuous usage of a diabetes management app on glycemic control in real-world clinical practice: retrospective analysis," *Journal of Medical Internet Research*, vol. 23, no. 7, p. e23227, 2021.

[12] R. Rowntree and L. Feeney, "Smartphone and video game use and perceived effects in a community mental health service," *Irish Journal of Medical Science (1971-)*, vol. 188, no. 4, pp. 1337–1341, 2019.

[13] "NIH- mental health information- statistics- mental illness," 2021. `https://www.nimh.nih.gov/health/statistics/mental-illness.shtml`.

[14] Sensley, "Molly: The virtual nurse," 2022. `https://sensely.com/`, Accessed online on 12/07/2022.

[15] V. Delić, M. Sečujski, N. V. Sedlar, D. Mišković, R. Mak, and M. Bojanić, "How speech technologies can help people with disabilities," in *International Conference on Speech and Computer*, pp. 243–250, Springer, 2014.

[16] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digital biomarkers*, vol. 5, no. 1, pp. 78–88, 2021.

[17] CDC, "Health and economic costs of chronic diseases," 2022. https://www.cdc.gov/chronicdisease/about/index.htm, Accessed online on 12/07/2022.

[18] F. Faurisson, "Survey of the delay in diagnosis for 8 rare diseases in europe: Eurordiscare2," *European Organisation for Rare Diseases Web site*, 2004.

[19] NIH, "Nih-budget," 2022. https://www.nih.gov/about-nih/what-we-do/budget, Accessed online on 12/06/2022.

[20] NIH-Research, 2021. https://www.nih.gov/about-nih/what-we-do/impact-nih-research, Accessed online on 05/07/2021.

[21] "Pubmed: Nih-ncbi research literature repository," 2021. https://pubmed.ncbi.nlm.nih.gov/.

[22] S. G. Finlayson, H. W. Chung, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *arXiv preprint arXiv:1804.05296*, 2018.

[23] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," *arXiv preprint arXiv:1812.05271*, 2018.

[24] J. Kotia, A. Kotwal, and R. Bharti, "Risk susceptibility of brain tumor classification to adversarial attacks," in *International Conference on Man–Machine Interactions*, pp. 181–187, Springer, 2019.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[26] S. Gaglio, A. Giammanco, G. Lo Re, and M. Morana, "Adversarial machine learning in e-health: Attacking a smart prescription system," in *International Conference of the Italian Association for Artificial Intelligence*, pp. 490–502, Springer, 2022.

[27] A. Newaz, N. I. Haque, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "Adversarial attacks to machine learning-based smart healthcare systems," *arXiv preprint arXiv:2010.03671*, 2020.

[28] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks on deep-learning models in natural language processing: A survey,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[29] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, “Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity,” in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pp. 82–94, 2018.

[30] K. O’Loughlin, M. Neary, E. C. Adkins, and S. M. Schueller, “Reviewing the data security and privacy policies of mobile apps for depression,” *Internet interventions*, vol. 15, pp. 110–115, 2019.

[31] T. Dehling, F. Gao, S. Schneider, and A. Sunyaev, “Exploring the far side of mobile health: information security and privacy of mobile health apps on ios and android,” *JMIR mHealth and uHealth*, vol. 3, no. 1, p. e8, 2015.

[32] J. Reardon, Á. Feal, P. Wijesekera, A. E. B. On, N. Vallina-Rodriguez, and S. Egelman, “50 ways to leak your data: An exploration of apps’ circumvention of the android permissions system,” in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 603–620, 2019.

[33] T. Riley, “An ftc order bars app maker easy healthcare from sharing additional personal health data with third parties for advertising,” 2023. https://cyberscoop.com/ftc-fertility-app-pregnancy-data/ .

[34] N. Alfawzan, M. Christen, G. Spitale, and N. Biller-Andorno, “Privacy, data sharing, and data security policies of women’s mhealth apps: Scoping review and content analysis,” *JMIR mHealth and uHealth*, vol. 10, no. 5, p. e33735, 2022.

[35] E. B. Scherwitzl, O. Lundberg, H. K. Kallner, K. G. Danielsson, J. Trussell, and R. Scherwitzl, “Perfect-use and typical-use pearl index of a contraceptive mobile app,” *Contraception*, vol. 96, no. 6, pp. 420–425, 2017.

[36] B. Aljedaani, M. A. Babar, *et al.*, “Challenges with developing secure mobile health applications: Systematic review,” *JMIR mHealth and uHealth*, vol. 9, no. 6, p. e15654, 2021.

[37] D. Cherkassky, “The voice privacy problem,” 2022. `https://www.kardome.com/blog-posts/voice-privacy-concerns`, Accessed online on 09/08/2022.

[38] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, “" hello, it’s me": Deep learning-based speech synthesis attacks in the real world,” in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 235–251, 2021.

[39] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, “Preech: A system for {Privacy-Preserving} speech transcription,” in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 2703–2720, 2020.

[40] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, “Evaluating voice conversion-based privacy protection against informed attackers,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2802–2806, IEEE, 2020.

[41] A. M. Williams, Y. Liu, K. R. Regner, F. Jotterand, P. Liu, and M. Liang, “Artificial intelligence, physiological genomics, and precision medicine,” *Physiological genomics*, vol. 50, no. 4, pp. 237–243, 2018.

[42] C. Krittanawong, “The rise of artificial intelligence and the uncertain future for physicians,” *European journal of internal medicine*, vol. 48, pp. e13–e14, 2018.

[43] “Nih-artificial intelligence- machine learning, and deep learning,” 2021. https://www.nibib.nih.gov/research-funding/machine-learning.

[44] “National human genome research institute,” 2023. https://www.genome.gov/dna-day/15-ways/rare-genetic-diseases, Accessed online on 03/02/2023.

[45] J. Lever, E. Y. Zhao, J. Grewal, M. R. Jones, and S. J. Jones, “Cancermine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer,” *Nature methods*, vol. 16, no. 6, pp. 505–507, 2019.

[46] W. E. Byrd, G. Rosenblatt, M. J. Patton, T. K. Tran-Nguyen, M. Zheng, A. Jain, M. Ballantyne, K. Zhang, M.-J. Chen, J. Whitlock, *et al.*, “medikanren: a system for bio-medical reasoning,” in *Proceedings of the 2020 ACM SIGPLAN international conference on functional programming*, 2020.

[47] C. Schardt, M. B. Adams, T. Owens, S. Keitz, and P. Fontelo, “Utilization of the pico framework to improve searching pubmed for clinical questions,” *BioMed Central medical informatics and decision making*, vol. 7, no. 1, p. 16, 2007.

[48] A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, and Z. Lu, “Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc,” *Nucleic acids research*, vol. 46, no. W1, pp. W530–W536, 2018.

[49] H. Poon, C. Quirk, C. DeZiel, and D. Heckerman, “Literome: Pubmed-scale genomic knowledge base in the cloud,” *Bioinformatics*, vol. 30, no. 19, pp. 2840–2842, 2014.

[50] C. Tao, Y. Zhang, G. Jiang, M.-M. Bouamrane, and C. G. Chute, “Optimizing semantic medline for translational science studies using semantic web technologies,” in *Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems*, pp. 53–58, ACM, 2012.

[51] IRIS.AI. https://iris.ai/, Accessed on 02.24.2021.

[52] E. Wood, A. K. Glen, L. G. Kvarfordt, F. Womack, L. Acevedo, T. S. Yoon, C. Ma, V. Flores, M. Sinha, Y. Chodpathumwan, *et al.*, “Rtx-kg2: a system for building a semantically standardized knowledge graph for translational biomedicine,” *BMC bioinformatics*, vol. 23, no. 1, pp. 1–33, 2022.

[53] K. D. Cobey, M. M. Lalu, B. Skidmore, N. Ahmadzai, A. Grudniewicz, and D. Moher, “What is a predatory journal? a scoping review,” *F1000Research*, vol. 7, 2018.

[54] S. Rawat and S. Meena, “Publish or perish: Where are we heading?,” *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, vol. 19, no. 2, p. 87, 2014.

[55] A. Grudniewicz, D. Moher, K. D. Cobey, G. L. Bryson, S. Cukier, K. Allen, C. Ardern, L. Balcom, T. Barros, M. Berger, *et al.*, “Predatory journals: no definition, no defence,” 2019. https://www.nature.com/articles/d41586-019-03759-y?sf225811500=1.

[56] T. F. Frandsen, “Are predatory journals undermining the credibility of science? a bibliometric analysis of citers,” *Scientometrics*, vol. 113, no. 3, pp. 1513–1528, 2017.

[57] R. E. Bartholomew, “Science for sale: the rise of predatory journals,” *Journal of the Royal Society of Medicine*, vol. 107, no. 10, p. 384, 2014.

[58] G. Richtig, M. Berger, B. Lange-Asschenfeldt, W. Aberer, and E. Richtig, “Problems and challenges of predatory journals,” *Journal of the European Academy of Dermatology and Venereology*, vol. 32, no. 9, pp. 1441–1449, 2018.

[59] B. Borrell, “A medical madoff: Anesthesiologist faked data in 21 studies,” March 2009. https://www.scientificamerican.com/article/a-medical-madoff-anesthestesiologist-faked-data/.

[60] F. Godlee, J. Smith, and H. Marcovitch, “Wakefield’s article linking mmr vaccine and autism was fraudulent,” January 2011. https://www.bmj.com/content/342/bmj.c7452/.

[61] T. Winey, “Garbage in, garbage out: Avoiding the common pitfalls of ai in healthcare,” 2017. https://www.beckershospitalreview.com/healthcare-information-technology/garbage-in-garbage-out-avoiding-the-common-pitfalls-of-ai-in-healthcare.html.

[62] E. J. Topol, “High-performance medicine: the convergence of human and artificial intelligence,” *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[63] L. Shamseer, D. Moher, O. Maduekwe, L. Turner, V. Barbour, R. Burch, J. Clark, J. Galipeau, J. Roberts, and B. J. Shea, “Potential predatory and legitimate biomedical journals:

can you tell the difference? a cross-sectional comparison," *BMC medicine*, vol. 15, no. 1, pp. 1–14, 2017.

[64] A. Manca, D. Moher, L. Cugusi, Z. Dvir, and F. Deriu, "How predatory journals leak into pubmed," *CMAJ*, vol. 190, no. 35, pp. E1042–E1045, 2018.

[65] M. S. Perlin, T. Imasato, and D. Borenstein, "Is predatory publishing a real threat? evidence from a large database study," *Scientometrics*, vol. 116, no. 1, pp. 255–273, 2018.

[66] A. S. K. P. Sorokowski, Emanuel Kulczycki, "Stop this waste of people, animals and money," September 2017. https://www.nature.com/news/stop-this-waste-of-people-animals-and-money-1.22554.

[67] A. L. Caplan, "The problem of publication-pollution denialism," in *Mayo Clinic Proceedings*, vol. 90, pp. 565–566, Elsevier, 2015.

[68] R. G. Steen, "Retractions in the scientific literature: is the incidence of research fraud increasing?," *Journal of medical ethics*, vol. 37, no. 4, pp. 249–253, 2011.

[69] R. G. Steen, "Retractions in the medical literature: how many patients are put at risk by flawed research?," *Journal of medical ethics*, vol. 37, no. 11, pp. 688–692, 2011.

[70] S. L. George and M. Buyse, "Data fraud in clinical trials," *Clinical investigation*, vol. 5, no. 2, p. 161, 2015.

[71] S. Lock, "Fraud in medicine," *British medical journal (Clinical research ed.)*, vol. 296, no. 6619, p. 376, 1988.

[72] R. Smith, "Research misconduct: the poisoning of the well," *Journal of the Royal Society of Medicine*, vol. 99, no. 5, pp. 232–237, 2006.

[73] R. J. Dinis-Oliveira, "Predatory journals and meetings in forensic sciences: what every expert needs to know about this "parasitic" publishing model," *Forensic sciences research*, vol. 6, no. 4, pp. 303–309, 2021.

[74] D. Mills and K. Inouye, "Problematizing 'predatory publishing': A systematic review of factors shaping publishing motives, decisions, and experiences," *Learned Publishing*, vol. 34, no. 2, pp. 89–104, 2021.

[75] S. Mertkan, G. Onurkan Aliusta, and N. Suphi, "Knowledge production on predatory publishing: A systematic review," *Learned Publishing*, vol. 34, no. 3, pp. 407–413, 2021.

[76] P. Thomas, J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser, "Geneview: a comprehensive semantic search engine for pubmed," *Nucleic acids research*, vol. 40, no. W1, pp. W585–W591, 2012.

[77] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu, "tmvar: a text mining approach for extracting sequence variants in biomedical literature," *Bioinformatics*, vol. 29, no. 11, pp. 1433–1439, 2013.

[78] C.-H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon, and Z. Lu, "tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine," *Bioinformatics*, vol. 34, no. 1, pp. 80–87, 2018.

[79] C.-H. Wei, H.-Y. Kao, and Z. Lu, "Pubtator: a web-based text mining tool for assisting biocuration," *Nucleic acids research*, vol. 41, no. W1, pp. W518–W522, 2013.

[80] J. Garcia-Pelaez, D. Rodriguez, R. Medina-Molina, G. Garcia-Rivas, C. Jerjes-Sánchez, and V. Trevino, "Pubterm: a web tool for organizing, annotating and curating genes, diseases, molecules and other concepts from pubmed records," *Database: The Journal of Biological Databases and Curation*, vol. 2019, 2019.

[81] A. Allot, K. Lee, Q. Chen, L. Luo, and Z. Lu, "Litsuggest: a web-based system for literature recommendation and curation using machine learning," *Nucleic acids research*, vol. 49, no. W1, pp. W352–W358, 2021.

[82] A. Sood, A. Ghosh, *et al.*, "Literature search using pubmed: an essential tool for practicing evidence-based medicine," *Journal-Association of Physicians of India*, vol. 54, no. R, p. 303, 2006.

[83] J. Dufour, J. Mancini, and M. Fieschi, “Searching for evidence-based data,” *Journal de chirurgie*, vol. 146, no. 4, pp. 355–367, 2009.

[84] A. Hoogendam, A. F. Stalenhoef, P. F. de Vries Robbé, and A. J. P. Overbeke, “Analysis of queries sent to pubmed at the point of care: observation of search behaviour in a medical teaching hospital,” *BioMed Central medical informatics and decision making*, vol. 8, no. 1, pp. 1–10, 2008.

[85] G. Bakal, P. Talari, E. V. Kakani, and R. Kavuluru, “Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations,” *Journal of biomedical informatics*, vol. 82, pp. 189–199, 2018.

[86] A. S. K. P. Sorokowski, Emanuel Kulczycki, “Predatory journals recruit fake editor,” March 2017. https://www.nature.com/news/predatory-journals-recruit-fake-editor-1.21662, Accessed on 02.24.2021.

[87] T. S. Rao and C. Andrade, “The mmr vaccine and autism: Sensation, refutation, retraction, and fraud,” *Indian journal of psychiatry*, vol. 53, no. 2, p. 95, 2011.

[88] M. Sun, F. Tang, J. Yi, F. Wang, and J. Zhou, “Identify susceptible locations in medical records via adversarial attacks on deep predictive models,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 793–801, 2018.

[89] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “Sok: Security and privacy in machine learning,” in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399–414, IEEE, 2018.

[90] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[91] F. Pesapane, M. Codari, and F. Sardanelli, “Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine,” *European radiology experimental*, vol. 2, no. 1, pp. 1–10, 2018.

[92] V. Duddu, "A survey of adversarial machine learning in cyber warfare," *Defence Science Journal*, vol. 68, no. 4, p. 356, 2018.

[93] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *The Lancet Oncology*, vol. 20, no. 5, pp. e262–e273, 2019.

[94] W. J. Rudman, J. S. Eberhardt, W. Pierce, and S. Hart-Hester, "Healthcare fraud and abuse," *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, vol. 6, no. Fall, 2009.

[95] C. S. Garver R, "Fda lets drugs approved on fraudulent research stay on the market." https://www.scientificamerican.com/article/fda-let-drugs-approved-on-fraudulent-research-stay-on-market/, Accessed on 03.19.2021.

[96] U. Jaffer and A. E. Cameron, "Deceit and fraud in medical research," *International Journal of Surgery*, vol. 4, no. 2, pp. 122–126, 2006.

[97] C. Seife, "Research misconduct identified by the us food and drug administration: out of sight, out of mind, out of the peer-reviewed literature," *JAMA internal medicine*, vol. 175, no. 4, pp. 567–577, 2015.

[98] H. Maisonneuve and D. Floret, "Wakefield's affair: 12 years of uncertainty whereas no link between autism and mmr vaccine has been proved," *Presse medicale (Paris, France: 1983)*, vol. 41, no. 9 Pt 1, pp. 827–834, 2012.

[99] F. Alshehri and G. Muhammad, "A comprehensive survey of the internet of things (iot) and ai-based smart healthcare," *IEEE Access*, vol. 9, pp. 3660–3678, 2020.

[100] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, "From blackbox to explainable ai in healthcare: existing tools and case studies," *Mobile Information Systems*, vol. 2022, 2022.

[101] S. Secinaro, D. Calandra, A. Secinaro, V. Muthurangu, and P. Biancone, "The role of artificial intelligence in healthcare: a structured literature review," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, pp. 1–23, 2021.

[102] S. Kaviani, K. J. Han, and I. Sohn, “Adversarial attacks and defenses on ai in medical imaging informatics: A survey,” *Expert Systems with Applications*, p. 116815, 2022.

[103] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, “A survey on knowledge graphs: Representation, acquisition, and applications,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.

[104] D. R. Unni, S. A. Moxon, M. Bada, M. Brush, R. Bruskiewich, J. H. Caufield, P. A. Clemons, V. Dancik, M. Dumontier, K. Fecho, *et al.*, “Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science,” *Clinical and Translational Science*, 2022.

[105] R. P. Singh, G. L. Hom, M. D. Abramoff, J. P. Campbell, M. F. Chiang, *et al.*, “Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient,” *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 45–45, 2020.

[106] P. Hamet and J. Tremblay, “Artificial intelligence in medicine,” *Metabolism*, vol. 69, pp. S36–S40, 2017.

[107] M. Afzal, S. R. Islam, M. Hussain, and S. Lee, “Precision medicine informatics: principles, prospects, and challenges,” *IEEE Access*, vol. 8, pp. 13593–13612, 2020.

[108] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, “Artificial intelligence in healthcare: past, present and future,” *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[109] G. S. Ginsburg and K. A. Phillips, “Precision medicine: from science to value,” *Health Affairs*, vol. 37, no. 5, pp. 694–701, 2018.

[110] O. Bodenreider, “The unified medical language system (umls): integrating biomedical terminology,” *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[111] L. Ogiela and M. R. Ogiela, “Fundamentals of cognitive informatics,” in *Advances in cognitive information systems*, pp. 19–49, Springer, 2012.

[112] T. C. Rindflesch and M. Fiszman, “The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text,” *Journal of biomedical informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[113] S. Zhou, X. Dai, H. Chen, W. Zhang, K. Ren, R. Tang, X. He, and Y. Yu, “Interactive recommender system via knowledge graph-enhanced reinforcement learning,” in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 179–188, 2020.

[114] X. Chen, S. Jia, and Y. Xiang, “A review: Knowledge reasoning over knowledge graph,” *Expert Systems with Applications*, vol. 141, p. 112948, 2020.

[115] S. V. K. Kayyali, David Knott, “The big-data revolution in us health care: Accelerating value and innovation,” April 2013. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care.

[116] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[117] S. Kundu, “Ai in medicine must be explainable,” *Nature medicine*, vol. 27, no. 8, pp. 1328–1328, 2021.

[118] Y. B. Masten and A. S. Ashcraft, “The dark side of dissemination: traditional and open access versus predatory journals,” *Nursing Education Perspectives*, vol. 37, no. 5, p. 275, 2016.

[119] J. Beall, “Predatory publishers are corrupting open access,” *Nature*, vol. 489, no. 7415, pp. 179–179, 2012.

[120] “Directory of open access journals (doaj),” 2023. https://doaj.org/, Accessed on 03.02.2021.

[121] “Definition of research misconduct,” 2023. https://ori.hhs.gov/definition-misconduct, Accessed on 03.02.2023.

[122] M. R. Mehra, S. S. Desai, F. Ruschitzka, and A. N. Patel, "Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis," 2020.

[123] S. Akça and M. Akbulut, "Are predatory journals contaminating science? an analysis on the cabells' predatory report," *The Journal of Academic Librarianship*, vol. 47, no. 4, p. 102366, 2021.

[124] K. D. M. Kabulo, U. S. Kanmounye, S. M. Ntshindj, K. Yengayenga, B. D. Takoutsing, P. Ntenga, L. Jokonya, J. Ntalaja, I. Esene, A. Musara, *et al.*, "Vulnerability of african neurosurgery to predatory journals: An electronic survey of aspiring neurosurgeons, residents, and consultants," *World Neurosurgery*, vol. 161, pp. e508–e513, 2022.

[125] O. Ivan, "Science reporter spoofs hundreds of open access journals with fake papers," October 2013. https://retractionwatch.com/2013/10/03/science-reporter-spoofs-hundreds-of-journals-with-a-fake-paper/, Accessed on 02.24.2021.

[126] E. Academy, "Fake peer review leads to massive retractions," 2021. https://www.enago.com/academy/fake-peer-review-leads-to-massive-retractions/.

[127] P. Notification, "Retraction notice regarding several articles published in tumor biology," 2021. https://pubmed.ncbi.nlm.nih.gov/34957978/.

[128] Z. Lu, "Pubmed and beyond: a survey of web tools for searching biomedical literature," *Database*, vol. 2011, 2011.

[129] T. C. Rindflesch, H. Kilicoglu, M. Fiszman, G. Rosemblat, and D. Shin, "Semantic medline: An advanced information management application for biomedicine," *Information Services & Use*, vol. 31, no. 1-2, pp. 15–21, 2011.

[130] X. Huang, J. Lin, and D. Demner-Fushman, "Evaluation of pico as a knowledge representation for clinical questions," in *AMIA annual symposium proceedings*, vol. 2006, p. 359, American Medical Informatics Association, 2006.

[131] B. Shepard, "Diagnosis in 2.127 seconds: Solving a years-long vomiting mystery using ai, research and brain power," August 2019. https://www.uab.edu/news/health/item/10703-diagnosis-in-2-127-seconds-solving-a-years-long-vomiting-mystery-using-ai-research-and-brain-power.

[132] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[133] S. Ren, Y. Deng, K. He, and W. Che, "Generating natural language adversarial examples through probability weighted word saliency," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1085–1097, 2019.

[134] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards crisp-ml (q): a machine learning process model with quality assurance methodology," *Machine learning and knowledge extraction*, vol. 3, no. 2, pp. 392–413, 2021.

[135] GPT3, "Text-generating algorithm from openai." https://www.digitaltrends.com/features/openai-gpt-3-text-generation-ai/.

[136] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[137] J. Stribling, M. Krohn, and D. Aguayo, "Scigen-an automatic cs paper generator," 2005. https://pdos.csail.mit.edu/archive/scigen/, Accessed on 03.19.2021.

[138] S. S. Gill, M. Xu, C. Ottaviani, P. Patros, R. Bahsoon, A. Shaghaghi, M. Golec, V. Stankovski, H. Wu, A. Abraham, *et al.*, "Ai for next generation computing: Emerging trends and future directions," *Internet of Things*, vol. 19, p. 100514, 2022.

[139] H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson, "The global burden of diagnostic errors in primary care," *British Medical Journal quality & safety*, vol. 26, no. 6, pp. 484–494, 2017.

[140] A. S. S. Tehrani, H. Lee, S. C. Mathews, A. Shore, M. A. Makary, P. J. Pronovost, and D. E. Newman-Toker, "25-year summary of us malpractice claims for diagnostic errors 1986–2010: an analysis from the national practitioner data bank," *BMJ quality & safety*, vol. 22, no. 8, pp. 672–680, 2013.

[141] C. Thoene, "Report of the national commission on orphan diseases- february 1989," 2021. `https://rarediseases.info.nih.gov/files/report_of_the_national_commission_on_orphan_diseases_february_1989.pdf`.

[142] T. Wang, Q.-R. Xing, H. Wang, and W. Chen, "Retracted publications in the biomedical literature from open access journals," *Science and engineering ethics*, vol. 25, no. 3, pp. 855–868, 2019.

[143] B. Borrell, "A medical madoff: Anesthesiologist faked data in 21 studies," 2009. `https://www.scientificamerican.com/article/a-medical-madoff-anesthestesiologist-faked-data/`.

[144] C. Seife, "Research misconduct identified by the us food and drug administration- out of sight, out of mind, out of the peer-reviewed literature." https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2109855.

[145] A. J. Wakefield, S. H. Murch, A. Anthony, J. Linnell, D. M. Casson, M. Malik, M. Berelowitz, A. P. Dhillon, M. A. Thomson, P. Harvey, *et al.*, "Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children," 1998.

[146] NIH-RareDiseases, "Genetic and rare diseases information center (gard)–an ncats program," 2021. `https://rarediseases.info.nih.gov/`.

[147] B. Shepard, "Diagnosis in 2.127 seconds: Solving a years-long vomiting mystery using ai, research and brain power - news," 2021. `https://www.uab.edu/news/health/item/10703-diagnosis-in-2-127-seconds-solving-a-years-long-vomiting-mystery-using-ai-research-and-brain-power`.

[148] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, “Deep learning for identifying metastatic breast cancer,” *arXiv preprint arXiv:1606.05718*, 2016.

[149] M. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, “Machine learning in genomic medicine: a review of computational problems and data sets,” *Proceedings of the IEEE*, vol. 104, no. 1, pp. 176–197, 2015.

[150] J. Wang, J. C. Ku, N. M. Alotaibi, and J. T. Rutka, “Retraction of neurosurgical publications: a systematic review,” *World neurosurgery*, vol. 103, pp. 809–814, 2017.

[151] J. M. Budd, M. Sievert, T. R. Schultz, and C. Scoville, “Effects of article retraction on citation and practice in medicine.,” *Bulletin of the Medical Library Association*, vol. 87, no. 4, p. 437, 1999.

[152] “Predatory journals.” https://predatoryjournals.com/journals/, Last Accessed on 12/11/2021.

[153] C. Shen and B.-C. Björk, “‘predatory’open access: a longitudinal study of article volumes and market characteristics,” *BMC medicine*, vol. 13, no. 1, pp. 1–15, 2015.

[154] M. Windsor, “A ‘high-speed dr. house’ for medical breakthroughs,” 2021. `https://www.uab.edu/news/research/item/10382-a-high-speed-dr-house-for-medical-breakthroughs`.

[155] GitHub-MediKanren. https://github.com/webyrd/MediKanren.

[156] A. Singhal, M. Simmons, and Z. Lu, “Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine,” *PLoS computational biology*, vol. 12, no. 11, p. e1005017, 2016.

[157] O. B. in AI Adoption in Healthcare. https://newsroom.intel.com/wp-content/uploads/sites/11/2018/07/healthcare-iot-infographic.pdf.

[158] E. Vayena, A. Blasimme, and I. G. Cohen, “Machine learning in medicine: addressing ethical challenges,” *PLoS medicine*, vol. 15, no. 11, 2018.

[159] B. Gastel, “Choosing a journal for submission: Don’t fall prey,” *Methodist Debakey Cardiovascular Journal*, vol. 17, no. 4, p. 90, 2021.

[160] “Predatory publishing: Author resources,” 2022. https://guides.library.uab.edu/c.php?g=826341&p=5922609.

[161] J. D. Olivarez, S. Bales, L. Sare, *et al.*, “Format aside: Applying beall’s criteria to assess the predatory nature of both oa and non-oa library and information science journals,” *College and research libraries*, vol. 79, no. 1, 2018.

[162] S. A. Elmore and E. H. Weston, “Predatory journals: what they are and how to avoid them,” *Toxicologic pathology*, vol. 48, no. 4, pp. 607–610, 2020.

[163] “Quality of doaj listed journals,” 2019. https://blog.doaj.org/2019/02/25/quality-of-doaj-listed-journals/.

[164] A. Mccallum and K. Nigam, “A comparison of event models for naive bayes text classification,” in *AAAI-98 Workshop on ’Learning for Text Categorization’*, 1998.

[165] A. Vahratian, S. J. Blumberg, E. P. Terlizzi, and J. S. Schiller, “Symptoms of anxiety or depressive disorder and use of mental health care among adults during the covid-19 pandemic—united states, august 2020–february 2021,” *Morbidity and Mortality Weekly Report*, vol. 70, no. 13, p. 490, 2021.

[166] CDC, 2021. https://www.cdc.gov/nchs/data/hus/2019/006-508.pdf.

[167] P. Winkler, T. Formanek, K. Mlada, A. Kagstrom, Z. Mohrova, P. Mohr, and L. Csemy, “Increase in prevalence of current mental disorders in the context of covid-19: analysis of repeated nationwide cross-sectional surveys,” *Epidemiology and psychiatric sciences*, vol. 29, 2020.

[168] K. C. Thomas, A. R. Ellis, T. R. Konrad, C. E. Holzer, and J. P. Morrissey, “County-level estimates of mental health professional shortage in the united states,” *Psychiatric Services*, vol. 60, no. 10, pp. 1323–1328, 2009.

[169] J. Proudfoot, “The future is in our hands: the role of mobile phones in the prevention and management of mental disorders,” *Australian & New Zealand Journal of Psychiatry*, vol. 47, no. 2, pp. 111–113, 2013.

[170] K. Wang, D. S. Varma, and M. Prosperi, “A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders,” *Journal of psychiatric research*, vol. 107, pp. 73–78, 2018.

[171] M. L. East and B. C. Havard, “Mental health mobile apps: from infusion to diffusion in the mental health social system,” *JMIR mental health*, vol. 2, no. 1, p. e10, 2015.

[172] M. Neary and S. M. Schueller, “State of the field of mental health apps,” *Cognitive and Behavioral Practice*, vol. 25, no. 4, pp. 531–537, 2018.

[173] L. Zhou, J. Bao, V. Watzlaf, and B. Parmanto, “Barriers to and facilitators of the use of mobile health apps from a security perspective: mixed-methods study,” *JMIR mHealth and uHealth*, vol. 7, no. 4, p. e11223, 2019.

[174] S. Chan, J. Torous, L. Hinton, and P. Yellowlees, “Towards a framework for evaluating mobile mental health apps,” *Telemedicine and e-Health*, vol. 21, no. 12, pp. 1038–1041, 2015.

[175] “Adaa- anxiety and depression association of america,” 2021. `https://adaa.org/about-adaa/press-room/facts-statistics`.

[176] Q. Liu, H. He, J. Yang, X. Feng, F. Zhao, and J. Lyu, “Changes in the global burden of depression from 1990 to 2017: Findings from the global burden of disease study,” *Journal of psychiatric research*, vol. 126, pp. 134–140, 2020.

[177] J. S. Khushalani, J. Qin, J. Cyrus, N. Buchanan Lunsford, S. H. Rim, X. Han, K. R. Yabroff, and D. U. Ekwueme, “Systematic review of healthcare costs related to mental health conditions among cancer survivors,” *Expert review of pharmacoeconomics & outcomes research*, vol. 18, no. 5, pp. 505–517, 2018.

[178] J. R. Cummings, “Rates of psychiatrists’ participation in health insurance networks,” *Jama*, vol. 313, no. 2, pp. 190–191, 2015.

[179] D. D. Luxton, R. A. McCann, N. E. Bush, M. C. Mishkind, and G. M. Reger, “mhealth for mental health: Integrating smartphone technology in behavioral healthcare.,” *Professional Psychology: Research and Practice*, vol. 42, no. 6, p. 505, 2011.

[180] K. G. Giota and G. Kleftaras, “Mental health apps: innovations, risks and ethical considerations,” *E-Health Telecommunication Systems and Networks*, vol. 2014, 2014.

[181] M. Olff, “Mobile mental health: a challenging research agenda,” *European journal of psychotraumatology*, vol. 6, no. 1, p. 27882, 2015.

[182] E. Chiauzzi and A. Newell, “Mental health apps in psychiatric treatment: a patient perspective on real world technology usage,” *JMIR mental health*, vol. 6, no. 4, p. e12292, 2019.

[183] L. Parker, V. Halter, T. Karliychuk, and Q. Grundy, “How private is your mental health app data? an empirical study of mental health app privacy policies and practices,” *International Journal of Law and Psychiatry*, vol. 64, pp. 198–204, 2019.

[184] J. Armontrout, J. Torous, M. Fisher, E. Drogin, and T. Gutheil, “Mobile mental health: navigating new rules and regulations for digital tools,” *Current psychiatry reports*, vol. 18, no. 10, p. 91, 2016.

[185] A. Powell, P. Singh, and J. Torous, “The complexity of mental health app privacy policies: a potential barrier to privacy,” *JMIR mHealth and uHealth*, vol. 6, no. 7, p. e158, 2018.

[186] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie, “Pscout: analyzing the android permission specification,” in *Proceedings of the 2012 ACM conference on Computer and communications security*, pp. 217–228, 2012.

[187] J. M. Robillard, T. L. Feng, A. B. Sporn, J.-A. Lai, C. Lo, M. Ta, and R. Nadler, “Availability, readability, and content of privacy policies and terms of agreements of mental health apps,” *Internet interventions*, vol. 17, p. 100243, 2019.

[188] J. Torous, G. Andersson, A. Bertagnoli, H. Christensen, P. Cuijpers, J. Firth, A. Haim, H. Hsin, C. Hollis, S. Lewis, *et al.*, "Towards a consensus around standards for smartphone apps and digital mental health," *World Psychiatry*, vol. 18, no. 1, p. 97, 2019.

[189] L. Parker, L. Bero, D. Gillies, M. Raven, and Q. Grundy, "The" hot potato" of mental health app regulation: A critical case study of the australian policy arena," *International journal of health policy and management*, vol. 8, no. 3, p. 168, 2019.

[190] A. Papageorgiou, M. Strigkos, E. Politou, E. Alepis, A. Solanas, and C. Patsakis, "Security and privacy analysis of mobile health applications: the alarming state of practice," *IEEE Access*, vol. 6, pp. 9390–9403, 2018.

[191] "Android developer guidelines," 2021. https://developer.android.com/.

[192] CWE, "Common weakness enumeration (cwe™)," 2021. https://cwe.mitre.org/data/definitions/926.html.

[193] D. Analysts, "Femtech industry landscape overview q4 2021," 2021. https://analytics.dkv.global/FemTech/Report-Q4.pdf.

[194] C. Stewart, "Global vc investments in femtech from 2012 to 2021," 2022. https://www.statista.com/statistics/1126913/femtech-vc-investment-worldwide/?locale=en.

[195] "Cisco annual internet report (2018–2023) white paper," 2020. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.

[196] R. Chandler, D. Guillaume, A. G. Parker, S. Carter, and N. D. Hernandez, "Promoting optimal sexual and reproductive health with mobile health tools for black women: Combining technology, culture and context," *Perspectives on sexual and reproductive health*, vol. 52, no. 4, p. 205, 2020.

[197] K. Zvarikova, V. Machova, and A. Pera, "Menstrual cycle tracking apps, fertility and reproductive data, and mobile health care management," *Journal of Research in Gender Studies*, vol. 12, no. 1, pp. 84–98, 2022.

[198] E. M. Greene, E. C. O'Brien, M. A. Kennelly, O. A. O'Brien, K. L. Lindsay, and F. M. McAuliffe, "Acceptability of the pregnancy, exercise, and nutrition research study with smartphone app support (pears) and the use of mobile health in a mixed lifestyle intervention by pregnant obese and overweight women: Secondary analysis of a randomized controlled trial," *JMIR mHealth and uHealth*, vol. 9, no. 5, p. e17189, 2021.

[199] S. Zuboff, N. Möllers, D. M. Wood, and D. Lyon, "Surveillance capitalism: An interview with shoshana zuboff," *Surveillance & Society*, vol. 17, no. 1/2, pp. 257–266, 2019.

[200] A. Ford, G. De Togni, and L. Miller, "Hormonal health: period tracking apps, wellness, and self-management in the era of surveillance capitalism," *Engaging Science, Technology, and Society*, vol. 7, no. 1, pp. 48–66, 2021.

[201] L. H. Harris, "Navigating loss of abortion services—a large academic medical center prepares for the overturn of roe v. wade," *New England Journal of Medicine*, vol. 386, no. 22, pp. 2061–2064, 2022.

[202] J. Weiss-Wolf, "Hhs issued guidance to protect private medical information. here are some best practices for users of period-tracking apps," 2022. https://msmagazine.com/2022/06/30/period-apps-women-health-data-information/ .

[203] "Hhs issues guidance to protect patient privacy in wake of supreme court decision on roe," 2022. https://www.hhs.gov/about/news/2022/06/29/hhs-issues-guidance-to-protect-patient-privacy-in-wake-of-supreme-court-decision-on-roe.html Accessed on 05/14/2023.

[204] FTC, "Complying with ftc's health breach notification rule," 2021. https://www.ftc.gov/business-guidance/resources/complying-ftcs-health-breach-notification-rule-0.

[205] J. R. Bull, S. P. Rowland, E. B. Scherwitzl, R. Scherwitzl, K. G. Danielsson, and J. Harper, "Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles," *NPJ digital medicine*, vol. 2, no. 1, p. 83, 2019.

[206] B. Reed, "Common weakness enumeration (cwe™)," 2018. https://www.theguardian.com/technology/2018/jan/17/birth-control-app-natural-cycle-pregnancies.

[207] D. Rosato, 2020. https://www.consumerreports.org/health-privacy/what-your-period-tracker-app-knows-about-you-a8701683935/.

[208] "Cdc 2020 abortion surveillance report," 2020. https://www.cdc.gov/reproductivehealth/data_stats/abortion.htm.

[209] "Progress pays off," 2018. https://powertodecide.org/sites/default/files/media/savings-fact-sheet-national.pdf.

[210] "Cdc 2019- about teen pregnancy," 2019. https://www.cdc.gov/teenpregnancy/about/index.htm.

[211] "Roe v. wade," 2022. https://reproductiverights.org/roe-v-wade/.

[212] A. J. Stevenson, "The pregnancy-related mortality impact of a total abortion ban in the united states: a research note on increased deaths due to remaining pregnant," *Demography*, vol. 58, no. 6, pp. 2019–2028, 2021.

[213] R. Bonta, "Attorney general becerra announces landmark settlement against glow, inc. – fertility app risked exposing millions of women's personal and medical information," 2020. https://oag.ca.gov/news/press-releases/attorney-general-becerra-announces-landmark-settlement-against-glow-inc-%E2%80%93.

[214] S. T. Campanella, "Menstrual and fertility tracking apps and the post roe v. wade era," 2022.

[215] "Flo- period tracking app: Anonymous mode faq," 2023. https://flo.health/privacy-portal/anonymous-mode-faq#.

[216] "Flo- period tracking app: Anonymous mode white paper," 2022. https://flo.health/flo-health-inc/news/anonymous-mode-whitepaper.

[217] K. Scales, A. Altman, S. Campbell, *et al.*, "It's time to care: A detailed profile of america's direct care workforce," *PHI: Quality Care Through Quality Jobs*, 2020.

[218] J. E. Allsworth, "Telemedicine, medication abortion, and access after roe v. wade," 2022.

[219] D. Lupton, "Quantified sex: a critical analysis of sexual and reproductive self-tracking using apps," *Culture, health & sexuality*, vol. 17, no. 4, pp. 440–453, 2015.

[220] C. L. Curchoe, "Smartphone applications for reproduction: from rigorously validated and clinically relevant to potentially harmful," *REPRODUCTIVE HEALTH*, 2020.

[221] G. LaMalva and S. Schmeelk, "Mobsf: Mobile health care android applications through the lens of open source static analysis," in *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–4, IEEE, 2020.

[222] L. Shipp and J. Blasco, "How private is your period?: A systematic analysis of menstrual app privacy policies.," *Proc. Priv. Enhancing Technol.*, vol. 2020, no. 4, pp. 491–510, 2020.

[223] G. E. Iyawa, A. R. Dansharif, and A. Khan, "Mobile apps for self-management in pregnancy: a systematic review," *Health and Technology*, vol. 11, pp. 283–294, 2021.

[224] Q. Grundy, K. Chiu, F. Held, A. Continella, L. Bero, and R. Holz, "Data sharing practices of medicines related apps and the mobile ecosystem: traffic, content, and network analysis," *BMJ*, vol. 364, 2019.

[225] S. Earle, H. R. Marston, R. Hadley, and D. Banks, "Use of menstruation and fertility app trackers: a scoping review of the evidence," *BMJ Sexual & Reproductive Health*, vol. 47, no. 2, pp. 90–101, 2021.

[226] A. M. Sarah Bradley, Elizabeth Bacharach and J. Spanfeller, "The 11 best period tracker apps to get to know your cycle, according to ob-gyns," 2022.

https://www.womenshealthmag.com/health/g26787041/best-period-tracking-apps/.

[227] D. Sullivan, "The 10 best period tracking apps," 2022. https://www.medicalnewstoday.com/articles/320758.

[228] Android, "Permissions on android," 2023. https://developer.android.com/guide/topics/permissions/overview.

[229] "Drozer- security testing framework for android," 2020. https://labs.f-secure.com/tools/drozer/.

[230] "Hipaa privacy: Personal identification information (pii)," 2023. https://www.govinfo.gov/content/pkg/CFR-2002-title45-vol1/pdf/CFR-2002-title45-vol1-sec164-514.pdf.

[231] "Hipaa privacy: Research on women's health," 2023. https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/HIPAA-privacy-rule.

[232] S. Riley, "Password security: What users know and what they actually do," *Usability News*, vol. 8, no. 1, pp. 2833–2836, 2006.

[233] R. Alomari and J. Thorpe, "On password behaviours and attitudes in different populations," *Journal of information security and applications*, vol. 45, pp. 79–89, 2019.

[234] S. Sivakorn, I. Polakis, and A. D. Keromytis, "The cracked cookie jar: Http cookie hijacking and the exposure of private information," in *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 724–742, IEEE, 2016.

[235] Android, "Static analysis using lint," 2023. https://developer.android.com/studio/write/lint.

[236] "Android: Restrictions on non-sdk interfaces," 2023. https://developer.android.com/guide/app-compatibility/restrictions-non-sdk-interfaces.

[237] "Sql injection," 2023. `https://developer.android.com/topic/security/risks/sql-injection`.

[238] "Owasp masvs (mobile application security verification standard)," 2023. https://mas.owasp.org/MASVS/.

[239] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *Computer Security–ESORICS 2015: 20th European Symposium on Research in Computer Security, Vienna, Austria, September 21-25, 2015, Proceedings, Part II 20*, pp. 599–621, Springer, 2015.

[240] M. Shirvanian, S. Vo, and N. Saxena, "Quantifying the breakability of voice assistants," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*, pp. 1–11, IEEE, 2019.

[241] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 529–533, IEEE, 2009.

[242] P. Champion, D. Jouvet, and A. Larcher, *A study of F0 modification for x-vector based speech pseudo-anonymization across gender*. PhD thesis, INRIA Nancy, équipe Multispeech, 2020.

[243] D. I. Adelani, A. Davody, T. Kleinbauer, and D. Klakow, "Privacy guarantees for de-identifying text transformations," *arXiv preprint arXiv:2008.03101*, 2020.

[244] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 338–343, IEEE, 2021.

[245] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE signal processing magazine*, vol. 11, no. 4, pp. 18–32, 1994.

[246] N. Tomashenko, B. M. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, *et al.*, "The voiceprivacy 2020 challenge," *VoicePrivacy, Feb*, 2020.

[247] S. Wolfson, "Amazon's alexa recorded private conversation and sent it to random contact," 2018. `https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation`, Accessed online on 04/16/2023.

[248] D. Prabakaran and R. Shyamala, "A review on performance of voice feature extraction techniques," in *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, pp. 221–231, IEEE, 2019.

[249] B. Dalvin, "Ai gave val kilmer his voice back. but critics worry the technology could be misused," 2022. `https://www.washingtonpost.com/technology/2021/08/18/val-kilmer-ai-voice-cloning`, Accessed online on 09/09/2022.

[250] L. Brinkschulte, N. Mariacher, S. Schlögl, M. I. Torres, R. Justo, J. M. Olaso, A. Esposito, G. Cordasco, G. Chollet, C. Glackin, *et al.*, "The empathic project: building an expressive, advanced virtual coach to improve independent healthy-life-years of the elderly," *arXiv preprint arXiv:2104.13836*, 2021.

[251] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, *et al.*, "Introducing the voiceprivacy initiative," *arXiv preprint arXiv:2005.01387*, 2020.

[252] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5500–5504, IEEE, 2016.

[253] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in asr: Reality or illusion?," *arXiv preprint arXiv:1911.04913*, 2019.

[254] R. Aloufi, H. Haddadi, and D. Boyle, "Emotion filtering at the edge," in *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*, pp. 1–6, 2019.

[255] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198637–198645, 2020.

[256] W. T. Hutiri and A. Y. Ding, "Bias in automated speaker recognition," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 230–247, 2022.

[257] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[258] L. Stiffler, "Lawsuit alleges amazon uses alexa interactions for ad targeting without users' knowledge or consent," 2022. `https://www.kardome.com/blog-posts/voice-privacy-concerns`, Accessed online on 09/09/2022.

[259] S. Furui, "Speaker-dependent-feature extraction, recognition and processing techniques," *Speech Communication*, vol. 10, no. 5-6, pp. 505–520, 1991.

[260] M. Marini, N. Vanello, and L. Fanucci, "Optimising speaker-dependent feature extraction parameters to improve automatic speech recognition performance for people with dysarthria," *Sensors*, vol. 21, no. 19, p. 6460, 2021.

[261] N. N. Etezady, "A survey of privacy metrics for smart homes," *CYBERSECURITY PEDAGOGY & PRACTICE JOURNAL*, 2023.

[262] L. Sweeney, "k-anonymity: A model for protecting privacy," *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.

[263] M. Burgess, "The race to hide your voice," 2022. `https://www.wired.com/story/voice-recognition-privacy-speech-changer/`, Accessed online on 09/09/2022.

[264] K. Kobayashi and T. Toda, "sprocket: Open-source voice conversion software.," in *Odyssey*, pp. 203–210, 2018.

[265] B. Sisman, J. Yamagishi, S. King, and H. Li, “An overview of voice conversion and its challenges: From statistical modeling to deep learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[266] T. Kaneko and H. Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1711.11293*, 2017.

[267] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273, IEEE, 2018.

[268] S. Zhao, H. Wang, T. H. Nguyen, and B. Ma, “Towards natural and controllable cross-lingual voice conversion based on neural tts model and phonetic posteriorgram,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5969–5973, IEEE, 2021.

[269] H. Turner, G. Lovisotto, and I. Martinovic, “Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020,” *arXiv preprint arXiv:2010.13457*, 2020.

[270] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker anonymization using x-vector and neural waveform models,” *arXiv preprint arXiv:1905.13561*, 2019.

[271] S. A. Anand, C. Wang, J. Liu, N. Saxena, and Y. Chen, “Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers,” *arXiv preprint arXiv:1907.05972*, 2019.

[272] K. R. Scherer, J. Koivumaki, and R. Rosenthal, “Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech,” *Journal of Psycholinguistic Research*, vol. 1, no. 3, pp. 269–285, 1972.

[273] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[274] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, “The interspeech 2011 speaker state challenge,” in *Proc. INTERSPEECH 2011, Florence, Italy*, 2011.

[275] M. Pobar and I. Ipšić, “Online speaker de-identification using voice transformation,” in *2014 37th International convention on information and communication technology, electronics and microelectronics (mipro)*, pp. 1264–1267, IEEE, 2014.

[276] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič, “Speaker de-identification using diphone recognition and speech synthesis,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 4, pp. 1–7, IEEE, 2015.

[277] F. Bahmaninezhad, C. Zhang, and J. H. Hansen, “Convolutional neural network based speaker de-identification.,” in *Odyssey*, pp. 255–260, 2018.

[278] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, X.-Y. Li, Y. Wang, and Y. Deng, “Voicemask: Anonymize and sanitize voice input on mobile devices,” *arXiv preprint arXiv:1711.11460*, 2017.

[279] M. Wester, Z. Wu, and J. Yamagishi, “Analysis of the voice conversion challenge 2016 evaluation results.,” in *Interspeech*, pp. 1637–1641, 2016.

[280] D. Cai, Z. Cai, and M. Li, “Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems,” *arXiv preprint arXiv:2206.09103*, 2022.

[281] Z. Wu and H. Li, “Voice conversion and spoofing attack on speaker verification systems,” in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9, IEEE, 2013.

[282] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. Interspeech 2021*, pp. 1344–1348, 2021.

[283] CorentinJ, "resemble-ai/ resemblyzer," 2020. `https://github.com/resemble-ai/Resemblyzer`, Accessed online on 09/09/2022.

[284] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883, IEEE, 2018.

[285] C. M. K. amagishi, JunichiVeaux, "Cstr vctk corpus)," 2019. `https://datashare.ed.ac.uk/handle/10283/3443`, Accessed online on 09/09/2022.

[286] J. Yamagishi, C. Veaux, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[287] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2014.

[288] K. Kobayashi, "sprocket," 2022. `https://github.com/k2kobayashi/sprocket`, Accessed online on 11/09/2022.

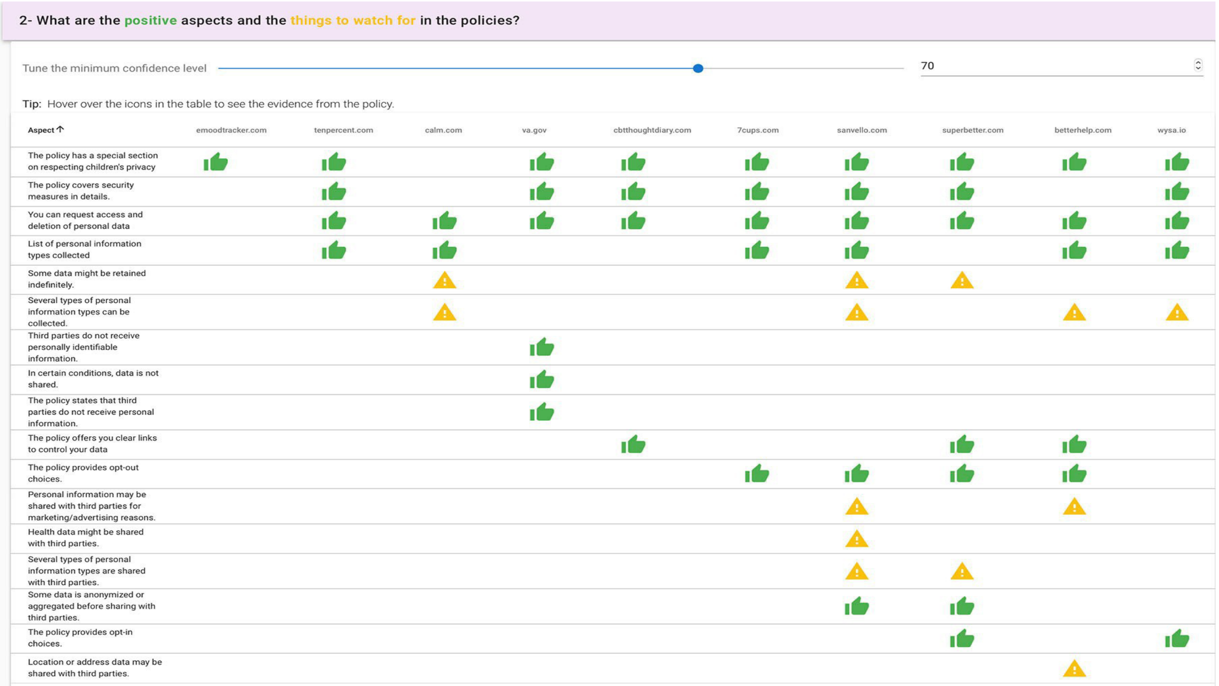
[289] Wendison, "Vqmivc," 2022. `https://github.com/Wendison/VQMIVC`, Accessed online on 11/09/2022.

[290] X. Liusong, "Stargan-voice-conversion," 2022. `https://github.com/liusongxiang/StarGAN-Voice-Conversion`, Accessed online on 11/09/2022.

APPENDIX A

# PRIVACY AND SECURITY VULNERABILITIES IN MOBILE HEALTHCARE APPS

## A.1 Mobile Mental Health (MMH) Apps



2- What are the positive aspects and the things to watch for in the policies?

Tune the minimum confidence level 70

Tip: Hover over the icons in the table to see the evidence from the policy.

| Aspect ↑ | emoodtracker.com | tenpercent.com | calm.com | va.gov | cbtthoughtdiary.com | 7cups.com | sanvello.com | superbetter.com | betterhelp.com | wysa.io |
|---|---|---|---|---|---|---|---|---|---|---|
| The policy has a special section on respecting children's privacy | 👍 | 👍 |  | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 |
| The policy covers security measures in details. |  | 👍 |  | 👍 | 👍 | 👍 | 👍 | 👍 |  | 👍 |
| You can request access and deletion of personal data |  | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 | 👍 |
| List of personal information types collected |  | 👍 | 👍 |  |  | 👍 | 👍 |  | 👍 | 👍 |
| Some data might be retained indefinitely. |  |  | ⚠ |  |  |  | ⚠ | ⚠ |  |  |
| Several types of personal information types can be collected. |  |  | ⚠ |  |  |  | ⚠ |  | ⚠ | ⚠ |
| Third parties do not receive personally identifiable information. |  |  |  | 👍 |  |  |  |  |  |  |
| In certain conditions, data is not shared. |  |  |  | 👍 |  |  |  |  |  |  |
| The policy states that third parties do not receive personal information. |  |  |  | 👍 |  |  |  |  |  |  |
| The policy offers you clear links to control your data |  |  |  |  | 👍 |  |  | 👍 | 👍 |  |
| The policy provides opt-out choices. |  |  |  |  |  | 👍 | 👍 | 👍 | 👍 |  |
| Personal information may be shared with third parties for marketing/advertising reasons. |  |  |  |  |  |  | ⚠ |  | ⚠ |  |
| Health data might be shared with third parties. |  |  |  |  |  |  | ⚠ |  |  |  |
| Several types of personal information types are shared with third parties. |  |  |  |  |  |  | ⚠ | ⚠ |  |  |
| Some data is anonymized or aggregated before sharing with third parties. |  |  |  |  |  |  | 👍 | 👍 |  |  |
| The policy provides opt-in choices. |  |  |  |  |  |  |  | 👍 |  | 👍 |
| Location or address data may be shared with third parties. |  |  |  |  |  |  |  |  | ⚠ |  |

Figure A.1: Polisis Analysis - Positive Aspects and Things to Watch for MMH Apps Privacy Policies

## A.2 Women’s Health Apps: Period-tracking and Fertility-monitoring Mobile Apps

Table A.1: MobSF Evaluation- Privacy and Security Vulnerabilities with References to OWASP/CWE Standards

| Privacy and Security Code Vulnerabilties | OWASP/CWE Reference |
|---|---|
| The App uses an insecure Random Number Generator. | CWE: CWE-330: Use of Insufficiently Random Values<br>OWASP Top 10: M5: Insufficient Cryptography |
| Files may contain hardcoded sensitive information like usernames, passwords, keys etc. | CWE: CWE-312: Cleartext Storage of Sensitive Information<br>OWASP Top 10: M9: Reverse Engineering |
| The App logs information. Sensitive information should never be logged. | CWE: CWE-532: Insertion of Sensitive Information into Log File<br>OWASP MASVS: MSTG-STORAGE-3 |
| IP Address disclosure | CWE: CWE-200: Information Exposure<br>OWASP MASVS: MSTG-CODE-2 |
| This App may request root (Super User) privileges. | CWE: CWE-250: Execution with Unnecessary Privileges<br>OWASP MASVS: MSTG- RESILIENCE-1 |
| This App may have root detection capabilities. | OWASP MASVS: MSTG- RESILIENCE-1 |
| App can write to App Directory. Sensitive Information should be encrypted. | CWE: CWE-276: Incorrect Default Permissions<br>OWASP MASVS: MSTG-STORAGE- 14 |
| App can read/write to External Storage. Any App can read data written to External Storage. | CWE: CWE-276: Incorrect Default Permissions<br>OWASP Top 10: M2: Insecure Data Storage |
| This App uses SSL certificate pinning to detect or prevent MITM attacks in secure communication channel. | OWASP MASVS: MSTG-NETWORK-4 |
| App creates temp file. Sensitive information should never be written into a temp file. | CWE: CWE-276: Incorrect Default Permissions<br>OWASP Top 10: M2: Insecure Data Storage |
| App uses SQLite Database and execute raw SQL query. Untrusted user input in raw SQL queries can cause SQL Injection. Also sensitive information should be encrypted and written to the database. | CWE: CWE-89: Improper Neutralization of Special Elements used in an SQL Command (’SQL Injection’)<br>OWASP Top 10: M7: Client Code Quality |
| MD5 is a weak hash known to have hash collisions. | CWE: CWE-327: Use of a Broken or Risky Cryptographic Algorithm<br>OWASP Top 10: M5: Insufficient Cryptography |
| SHA-1 is a weak hash known to have hash collisions. | CWE: CWE-327: Use of a Broken or Risky Cryptographic Algorithm<br>OWASP Top 10: M5: Insufficient Cryptography<br>OWASP MASVS: MSTG-CRYPTO-4 |
| Insecure Implementation of SSL. Trusting all the certificates or accepting self signed certificates is a critical Security Hole. This application is vulnerable to MITM attacks | CWE: CWE-295: Improper Certificate Validation<br>OWASP Top 10: M3: Insecure Communication<br>OWASP MASVS: MSTG-NETWORK-3 |
| The App uses the encryption mode CBC with PKCS5/PKCS7 padding. This configuration is vulnerable to padding oracle attacks. | CWE: CWE-649: Reliance on Obfuscation or Encryption of Security-Relevant Inputs without Integrity Checking<br>OWASP Top 10: M5: Insufficient Cryptography |
| Insecure WebView Implementation. Execution of user controlled code in WebView is a critical Security Hole. | CWE: CWE-749: Exposed Dangerous Method or Function<br>OWASP Top 10: M1: Improper Platform Usage<br>OWASP MASVS: MSTG-PLATFORM-7 |
| Remote WebView debugging is enabled. | CWE: CWE-919: Weaknesses in Mobile Applications<br>OWASP Top 10: M1: Improper Platform Usage |
| This App copies data to clipboard. Sensitive data should not be copied to clipboard as other applications can access it. | OWASP MASVS: MSTG-STORAGE-10 |

# APPENDIX B

# RESEMBLYZER- SPEAKER RECOGNITION

## B.1 Speaker Recognition by analyzing Cross-Similarity and Speaker Embedding

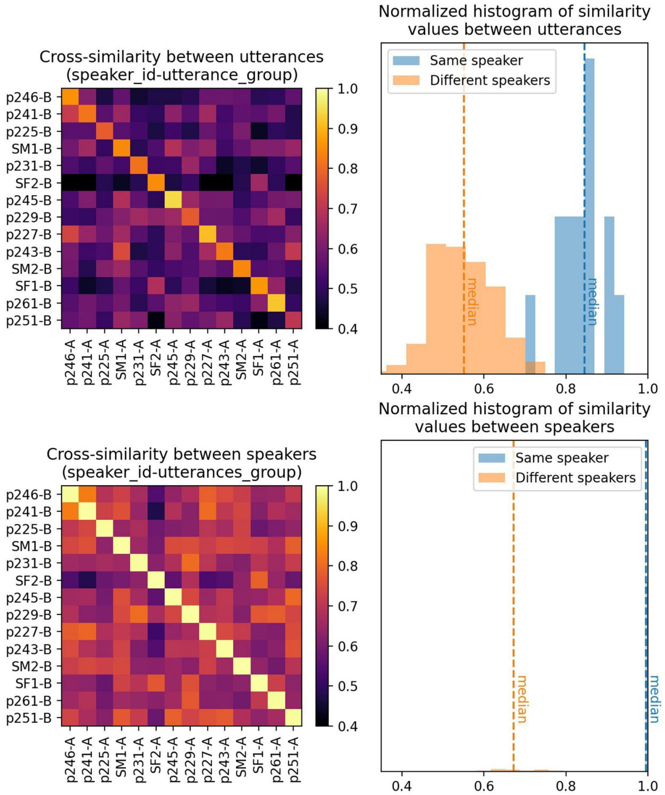


Figure B.1: Resemblyzer: Benign- VCTK and VCC Speakers and Utterances Cross-similarity

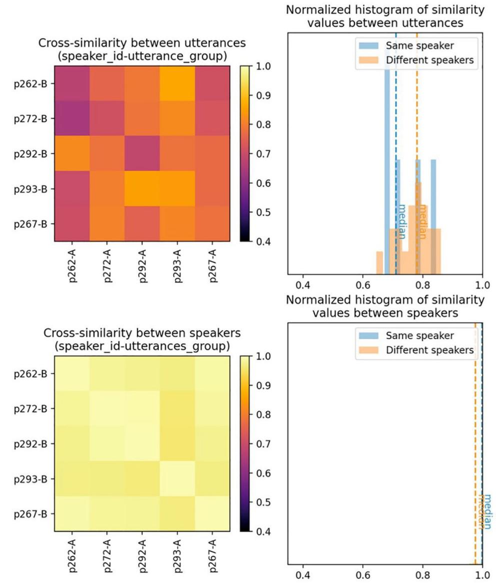


Figure B.2: Resemblyzer: VCTK Cross-Similarity: StarGAN-VC- Target spk-F

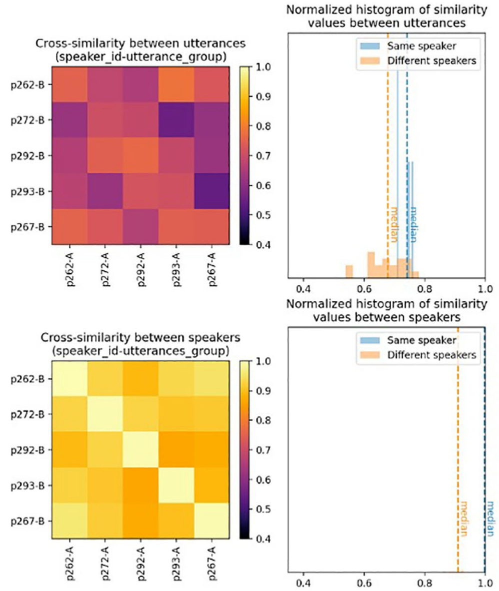


Figure B.3: Resemblyzer: Cross-Similarity: StarGAN-VC- Target spk-M

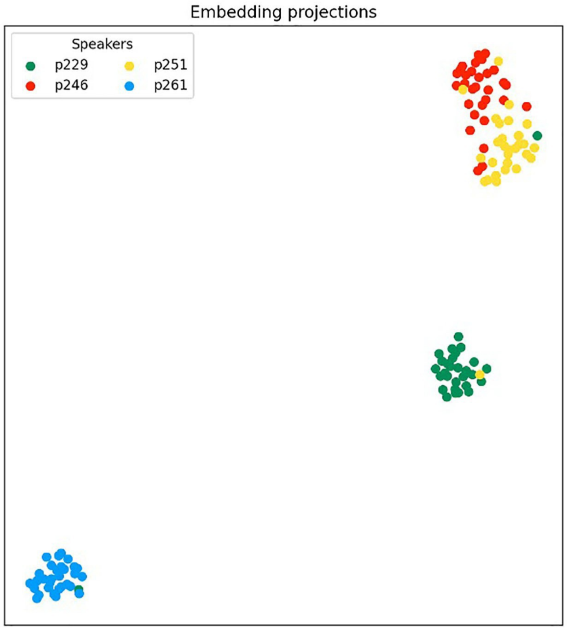


Figure B.4: Resemblyzer: Speaker Embedding for sprocket VC- Target spk-F

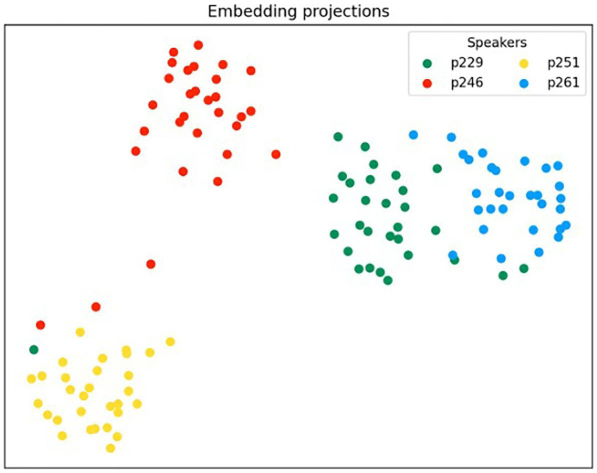


Figure B.5: Resemblyzer: Speaker Embedding for sprocket-VC- Target spk-M

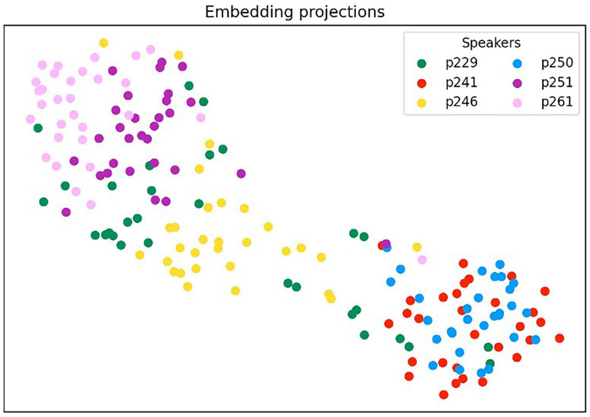


Figure B.6: Resemblyzer: Speaker Embedding for VQMIVC-Target Speaker F

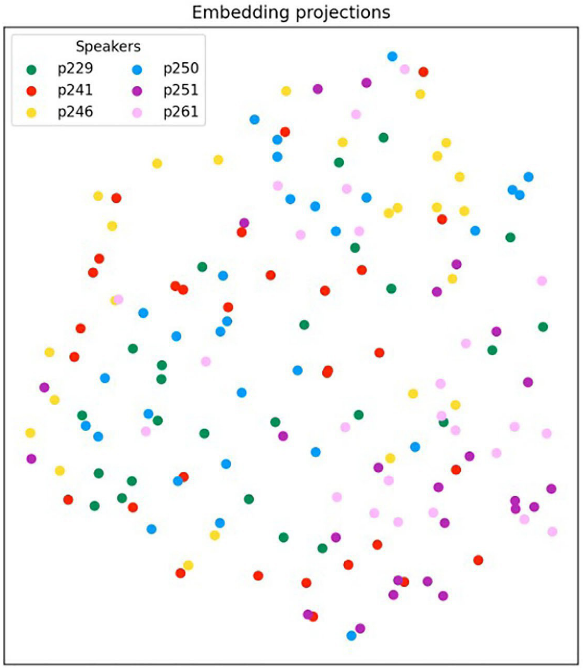


Figure B.7: Resemblyzer: Speaker Embedding for VQMIVC-Target Speaker M

# APPENDIX C

# LIST OF RESEARCH PUBLICATIONS

## C.1 Published Work

### As First Author

- Saini, Shalini, and Saxena, Nitesh. "A Survey of Threats to Research Literature Dependent Medical AI Solutions." ACM Computing Surveys (2023).

- Saini, S. and Saxena, N., 2022. Predatory Medicine: Exploring and Measuring the Vulnerability of Medical AI to Predatory Science. arXiv preprint arXiv:2203.06245. Conference on Health, Inference, and Learning (CHIL).

- Saini, S., Panjwani, D. and Saxena, N., 2022, August. Mobile Mental Health Apps: Alternative Intervention or Intrusion? In 2022 19th Annual International Conference on Privacy, Security & Trust (PST) (pp. 1-11). IEEE

### As Supporting Author

- Walker, P., Saini, S., Anand, S.A., Halevi, T. and Saxena, N., 2022, May. Hearing Check Failed: Using Laser Vibrometry to Analyze the Potential for Hard Disk Drives to Eavesdrop Speech Vibrations. In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (pp. 67-81). ASIACCS.

## C.2 Publications Under Review

- Saini, Shalini, and Saxena, Nitesh, 2023, “The Vulnerability of Voice Conversion in Preserving Speaker Anonymity: An Analysis with Speaker Recognition Systems"

- Saini, Shalini, and Saxena, Nitesh, 2023, “Privacy and Security of Femtech Apps: Navigating Threats to Women’s Health in a Changing Legal Landscape"