Daily Behavior Modeling and Health State Prediction Through Smartphone and Smartwatch Data

Health Data Privacy and Security Analysis with Smartphone and Smartwatch Devices

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Wei Wang,

Technical Project Team Members

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature Wei Wang Date 05/07/2021

Approved _____ Date _____
Afsaneh Doryab, Department of Systems and Information Engineering

Approved _____ Date _____
Sharon Tsai-hsuan Ku, Department of Engineering and Society

Introduction

As smartphone and smart wearable technologies are becoming more prevalent, more data is collected from users and used for various purposes. One of the numerous applications of such data is in the health field. Most wearable devices and some smartphones are equipped with sensors that can collect specific health information, including heart rate and oxygen levels. The health data obtained by these devices are often simple; however, with advanced data analysis techniques such as machine learning, it is possible to determine whether there exists a connection between this data and other related information. In this research, studies will be conducted to determine the potential of inferring physical and mental health from this data.

Additionally, one issue introduced by these data collection methods is privacy. In many cases, the data collected will be transferred to business organizations, such as device manufacturers and software developers. The privacy and security of this data is ensured solely by these organizations' privacy policies, and users may lose the ownership of their data. Therefore, another important focus of this research is to study and understand the privacy concerns with data collected from these devices.

Technical Topic

This technical research's main topic is to extract information relating to users' physical and mental health from the data collected from smartphones, smartwatches, and other types of wearable devices.

The primary method employed to infer the physical and mental health of users of various devices is machine learning. Several different types of machine learning models will be constructed and trained with the data collected from these devices. These models may include regression, classification, and clustering depending on the data available, such as whether the data is labeled or not. Additionally, certain machine learning models may be more suitable than others, depending on the types of labels. For example, if the labels are health index scores or of similar kind, regression models will be more optimal than others. If the labels are categories such as unhealthy, healthy, and healthiest, classification models should be applied. If the data is unlabeled, clustering models will be used.

The data required for this research will be collected from smartphone and smartwatch users. Health data will be exported from the participating users' devices, and surveys, questionnaires, or interviews will be conducted with users to generate scores regarding their health status. These scores may be used as the labels for the data collected or will be categorized into different health groups for classification. Depending on the data label type, different machine learning models will be utilized for data analysis. Another source of data is public databases. Various public databases will be searched to find relevant datasets, preferably labeled.

Due to time and resource constraints, the sample size of the data collected may be small, which will result in biases if unrepresentative of the overall population. However, it will be possible to avoid such an issue if large datasets are found in public databases that can be used as

supplements to the primary data collected from participating users. Another possible source of bias in this research is the survey and interview questions as inappropriate questions will lead to inaccurate results.

Data from users and public databases are the most crucial component in this research since they will determine the type of machine learning models used. Thus, the first step in this research will be data collection. Approximately two to three weeks will be required to complete the data collection process; however, an additional two weeks will be allocated to this process if the quality or size of the data collected is unsatisfactory. Various machine learning models will be trained and tested for their accuracy based on the data obtained, which will require one or two weeks. Once outcomes are obtained from the models, the remaining time will be spent on synthesizing the report.

STS Prospectus

Introduction

Current smartwatch and smartphone technologies enable people to track specific health data, such as heart rate, and sleep quality. This data can be used to help people build healthy daily routines. However, despite the convenience provided by these new technologies, they also create new security and privacy issues, such as possible improper usages of users' health data. According to the U.S. Consumer Privacy Index 2016 provided by the TRUSTe/National Cyber Security Alliance (NCSA), 68% of the population were concerned about their data privacy, whereas only 57% were worried about losing their primary source of income (*TRUSTe National Cyber Security Alliance U.S. Consumer Privacy Index 2016 Infographic* 2016).

Health data is crucial when it comes to public health management. For example, this data can be analyzed to reveal insights regarding prevention of potential health crises. With the advancement of technology, health data collection has become more convenient with smart devices such as smartphones and smartwatches. Health professionals may be utilizing these technologies to closely monitor their patients' health status in real-time and detect issues earlier. However, there are risks involved with using these devices as a source of health data. Traditionally, health data is collected by healthcare service providers and other related organizations, including hospitals and health insurance companies. These organizations and their business associates are regulated by the HIPAA and HITECH act, which list the privacy and security practices that must be strictly followed. However, by using smart devices to collect health data, organizations like technology companies may also store and access this data, and are not required to comply with the HIPAA and HITECH act. Thus, these organizations may use the

data according to their own interests, and users may lose ownership of their health data. For example, in 2017, Facebook created an algorithm to analyze users' mental health from the posts made. The data collected by the company was not held at the same level of security and privacy standard as healthcare service providers and was collected without users' consent (Goggin, 2019). Lastly, user data collected are often shared between business partners, but it is often difficult for users to determine the flow of their data between these partners and understand the privacy policy of each.

Research Question

Privacy in a digitized world is essentially the right of users to determine "who can access what data about me, where, when, and for what purpose" (Newman, 2019). Since information is stored and collected virtually, it is more difficult to prevent and track unauthorized accesses. Additionally, some data can be used to derive health information, but may not be considered health data, such as the social media posts used by Facebook to analyze users' mental health. With more health data collected and created by organizations outside of HIPAA regulations, it is difficult to determine the legal ownership of this data. However, more health data can lead to better health management and risk detection. To better understand the advantages and disadvantages of health data privacy in the digital age, it is essential to answer the following questions: what types of data are shared and categorized as health data, what technologies are used to ensure security and privacy, how is this data used by different organizations and their business partners, and what challenges are created by privacy regulations.

Literature Review

Within the past ten years, numerous studies have been done to learn the various forms of privacy risks regarding health data in a digitized world. This data is usually collected from a variety of platforms, including social media and wearable devices. Most of these studies are focused on privacy issues with current technologies; however, very few are considering the possible new problems introduced by the evolution of these technologies in the future.

Throughout these studies, several common themes appeared consistently. For example, several research projects analyzed the users' awareness of the privacy risks involved with using wearable health devices or other health software and applications. These studies often look at whether users understand the privacy conditions that they approved and whether companies have purposely included specific terms that would enable them to use data in their interests (Al Ameen, Liu and Kwak, 2010; Raij, Ghosh, Kumar and Srivastava, 2011; Paul and Irvine, 2014; Obar, 2015; Ostherr et al., 2017; Gostin, Halabi and Wilson, 2018). In some cases, the technologies or methodologies have been developed recently, and the users and service providers may not fully realize their privacy risks (Motti and Caine, 2015). Flaws within a technology may lead to serious privacy concerns. Many research projects have analyzed the technologies involved with wireless communication between different devices and evaluated their security (Zhou and Piramuthu, 2014; Ching and Singh, 2016).

Regulatory agencies and governments' roles in data privacy protection is another focus in many studies. Some groups have analyzed the current regulations and law systems to understand the authority’s efforts to ensure the confidentiality of consumer data (Van Dijck and Poell, 2016; McDermott, 2017). Another perspective looked at scenarios where the government has become a source of privacy issues (Vezyridis and Timmons, 2017). Lastly, modern data analysis methods allow organizations to infer users' mental health state based on their social media activities. One

research project has been done to study the ethical tensions and other related social problems involved with such practice (Chancellor et al., 2019).

Most of these studies and research projects have analyzed user-health data privacy issues from many different perspectives and focused on various stakeholders. One common conclusion that can be drawn from these analyses is that it would be impractical to rely on users to control the privacy of their data in most cases. One study stated that data privacy self-management is a fallacy and practically impossible as users would often accept service terms and conditions without a complete and well understanding (Obar, 2015). Consequently, it will be necessary for the government and regulatory agencies to step in and prevent certain actions against user data privacy and security. An example of the government's effort to protect consumer data privacy is the California Consumer Privacy Act (CCPA), which has become effective on January 1, 2020. This act enhances the consumer privacy rights of California residents by allowing them to request deletions of their data stored by businesses and their affiliates (Dean, 2020).

Research studies have also suggested that as technology improves over time, the privacy and security model will evolve. As a result, more advanced standards should be created to encounter these changes. These standards may be new regulations or more advanced technical procedures that will bring security to a new level.

These studies provide a useful framework for developing the proper privacy and security practices for current technologies and systems. However, these technologies will evolve quickly; new systems may be designed and used for data sharing and applications. It would be critical to consider these possibilities and method that can be used to encounter these changes.

STS Framework and Method

When studying the privacy issues involved with current mobile and wearable health technologies, the SCOT (social construction of technology) framework will be most suitable to conduct such an analysis. This STS framework allows the researcher to study the impacts of the relevant technologies on various social groups. Both positive and negative effects will be analyzed, which will provide helpful insights into the different interests of each social group. These differences may lead to conflicts among these groups; thus, it is crucial to study these differences and learn how they will be resolved.

The central technology involved with this research is mobile health devices, and some of the relevant social groups include the users and healthcare service providers. Healthcare professionals can utilize this technology to track their patients' health status and provide useful advice promptly based on the data collected. It can help the patients to avoid potential health risks better. The users of mobile health technologies can be categorized based on use cases and applications. For example, older people and those with higher health risks can use this technology to monitor their health information closely and receive timely warnings. Another group of health users including fitness professionals, may use this technology to record essential health data and track their exercise progress. Lastly, mobile health technology may not necessarily be used for health-related purposes. This technology is often integrated with smartphones and wearable devices. Users may own one of these devices for its non-health-related functionalities, such as entertainment, communication, and style. Finally, the devices are designed and manufactured by technology companies, such as Apple and Samsung; thus, they will inevitably become a relevant social group.

The various relevant social groups of mobile health technology have different ways of utilizing this technology; hence, they will likely have varying degrees of influence over this

technology's development. The manufacturers and software designers will determine the various functionalities of the devices. As a result, their influence will be the greatest. With other social groups, the level of influence can be correlated with the criticality of their applications. For example, healthcare providers and doctors can use mobile health devices to help patients detect and avoid health risks more efficiently. This application has a higher importance than others. Thus, doctors should be more influential than other groups.

Depending on the application, these social groups may have similar or different expectations and interests regarding the technology's functions. One shared expectation among health users is accuracy. Doctors and health users will use this technology for health risk detection or fitness progress tracking. As a result, the accuracy of the health data collected will be crucial. Most groups, including non-health users, will agree that their health data's privacy and security should be prioritized and ensured. However, groups may have different expectations regarding the level of privacy and security provided based on the types of data collected. For example, if more sensitive information is collected, more secure protocols should be used. Another point of disagreement will be the cost and functionalities. Doctors and users with higher health risks will be interested in advanced health diagnosis features, such as real-time analysis and warnings, for better risk prevention. These functionalities will be unnecessary for regular health users, such as fitness users and non-health users. These functions may result in higher costs of mobile health devices.

To settle the disagreements, different social groups can negotiate with each other to reach a common interest. However, since there are some very different needs and expectations, it will be difficult to agree on all topics through negotiations. Alternatively, these social groups can approach the technology companies that design and manufacture mobile health devices. They

can discuss with the companies and request them to create multiple designs that will address different social groups' needs. Multiple designs will likely resolve any considerable disagreements among the social groups, such as the cost and functionalities. Smaller disputes may be settled by negotiating and establishing standards. For example, social groups may agree upon a standardized protocol to ensure their health data's privacy and security.

Timeline

Data will be the most important part of this research, and will be collected with a few different methods. Surveys and interviews will be conducted with the various stakeholders involved with mobile health technologies. These stakeholders include manufacturers, users, healthcare professionals, and regulatory agencies. One potential bias that may be involved with this method of data collection is under coverage. Due to limited time and resources, a relatively small sample size will be used, which may be unrepresentative of the overall population. Furthermore, the interview or survey questions may not cover all perspectives of the problems involved in this research, resulting in another bias. Additionally, a different project will be conducted alongside with this research that focuses on modeling mental and physical health states from smartwatch and smartphone data. Data collected from this research project may also be referenced as quantitative data if found relevant. Lastly, secondary data found in public databases and past research projects will be analyzed and used.

Data collection will be the first step in this research. Within the first week, interview and survey questions should be created, and the potential candidates should be determined. Over the next two to three weeks, these interviews and surveys should be conducted with the potential candidates. Concurrently, various public databases and past research results will be searched for

relevant information. This time frame may be extended if the data collected is insufficient. Once the data collection process is completed, the rest of the time will be dedicated to creating the research report.

Conclusion

In short, the outcomes of this research will lead to a better understanding of the privacy issues involved with the data collection and sharing process in mobile health devices. They will also provide a clearer insight into how health data is defined in the new digitized era. Most importantly, the conclusions will provide a vision into the privacy issues that may be introduced with technological advancement in the future and provide suggestions regarding the privacy practices that can be applied to encounter these issues.

Bibliography

Al Ameen, M., Liu, J. and Kwak, K., 2010. Security and Privacy Issues in Wireless Sensor Networks for Healthcare Applications. Journal of Medical Systems, 36(1), pp.93-101.

Chancellor, S., Birnbaum, M., Caine, E., Silenzio, V. and De Choudhury, M., 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19,.

Ching, K. and Singh, M., 2016. Wearable Technology Devices Security and Privacy Vulnerability Analysis. International Journal of Network Security & Its Applications, 8(3), pp.19-30.

Dean, S., 2020. It's 2020 And You Have New Privacy Rights Online. But You Might Have To Show ID. [online] Los Angeles Times. Available at: <https://www.latimes.com/business/technology/story/2020-01-01/ccpa-california-internet-rights-what-you-need-know> [Accessed 8 November 2020].

Goggin, B. (2019, January 06). Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts. Retrieved December 03, 2020, from https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12?op=1

Gostin, L., Halabi, S. and Wilson, K., 2018. Health Data and Privacy in the Digital Era. JAMA, 320(3), p.233.

McDermott, Y., 2017. Conceptualizing the right to data protection in an era of Big Data. Big Data & Society, 4(1), p.205395171668699.

Motti, V. and Caine, K., 2015. Users' Privacy Concerns About Wearables. Financial Cryptography and Data Security, pp.231-244.

Newman, D. (2019, May 03). What Is Privacy In The Age Of Digital Transformation? Retrieved December 03, 2020, from https://www.forbes.com/sites/danielnewman/2019/05/02/what-is-privacy-in-the-age-of-digital-transformation/?sh=246b3e96628e

Obar, J., 2015. Big Data andThe Phantom Public: Walter Lippmann and the fallacy of data privacy self-management. Big Data & Society, 2(2), p.205395171560887.

Ostherr, K., Borodina, S., Bracken, R., Lotterman, C., Storer, E. and Williams, B., 2017. Trust and privacy in the context of user-generated health data. Big Data & Society, 4(1), p.205395171770467.

Paul, G. and Irvine, J., 2014. Privacy Implications of Wearable Health Devices. Proceedings of the 7th International Conference on Security of Information and Networks - SIN '14,.

Raij, A., Ghosh, A., Kumar, S. and Srivastava, M., 2011. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11,.

TRUSTe National Cyber Security Alliance U.S. Consumer Privacy Index 2016 Infographic [PDF]. (2016). TRUSTe, Inc.

Van Dijck, J. and Poell, T., 2016. Understanding the promises and premises of online health platforms. Big Data & Society, 3(1), p.205395171665417.

Vezyridis, P. and Timmons, S., 2017. Understanding the care.data conundrum: New information flows for economic growth. Big Data & Society, 4(1), p.205395171668849.

Zhou, W. and Piramuthu, S., 2014. Security/privacy of wearable fitness tracking IoT devices. 2014 9th Iberian Conference on Information Systems and Technologies (CISTI),.