MACHINE LEARNING METHODS FOR CLINICAL AND HEALTHCARE APPLICATIONS USING ELECTRONIC HEALTH RECORDS

a Thesis submitted by

SHRUTI KAUSHIK

(D15043)

for the award of the degree of Doctor of Philosophy

SCHOOL OF COMPUTING & ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

December 7, 2020

THESIS CERTIFICATE

This is to certify that the work contained in the thesis entitled “Machine learning methods for clinical and healthcare applications using electronic health records” being submitted by Ms. Shruti Kaushik (Enroll. No: D15043) has been carried out under my supervision. In my opinion, the thesis has reached the standard fulfilling the requirement of regulation of the Ph.D. degree. The results embodied in this thesis have not been submitted elsewhere for the award of any degree or diploma.

Varun Dutt

December 7, 2020

Dr. Varun Dutt (Supervisor)
Associate Professor
School of Computing and Electrical Engineering
School of Humanities and Social Sciences
Indian Institute of Technology Mandi
Kamand, Himachal Pradesh, India
Email: varun@iitmandi.ac.in

Declaration by the Research Scholar

I hereby declare that the entire work embodied in this Thesis is the result of investigations carried out by me in the ***School of Computing and Electrical Engineering***, Indian Institute of Technology Mandi, under the supervision of ***Dr. Varun Dutt***, and that it has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on findings of other investigators.

Place: IIT Mandi, Kamand

Date: December 7, 2020

Signature: Skaushik

Name: Shruti Kaushik

Declaration by the Research Advisor

I hereby certify that the entire work in this Thesis has been carried out by ***Shruti Kaushik***, under my supervision in the ***School of Computing and Electrical Engineering***, Indian Institute of Technology Mandi, and that no part of it has been submitted elsewhere for any Degree or Diploma.

Signature: Varun Dutt

Name of the Guide: Dr. Varun Dutt

Date: December 7, 2020

ABSTRACT

World health organization estimates an increasing global trend of healthcare costs, and it is anticipated that the machine learning (ML) models may help to predict and manage these costs. However, ML research for predicting patients' expenditures using EHRs is relatively new. Furthermore, for multivariate time-series predictions in the healthcare domain, the use of multi-headed neural network architectures has been less explored in the literature. Additionally, researchers have not explored generative adversarial networks (GANs) for predicting healthcare outcomes using multivariate time-series datasets. In this thesis, a number of experiments addressed these gaps in literature. In the first experiment, the potential of Apriori frequent item-set mining approach was evaluated to discover the frequently appearing diagnoses or procedure codes among several features in healthcare datasets. The selected features combined with demographic and clinical features were used to classify patients according to the medicine consumed by them. Classification algorithm results revealed that the performance of all ML algorithms improved when only frequent features selected from Apriori were used in classification compared to all the features in a US dataset. However, this finding was not robust across a second dataset collected in India. In the second experiment, state-of-the-art feature selection approaches (information gain, correlation coefficient score, LASSO, and ridge regression) and feature transformation approaches (principal component analysis and auto-encoders) were evaluated to find relevant features in healthcare datasets. Results revealed that feature engineering helped in improving the classification accuracy in certain healthcare datasets. In the third experiment, statistical models (persistence and autoregressive integrated moving average (ARIMA)), multi-layer perceptron (MLP), long short-term memory (LSTM), and a novel ensemble model combining predictions of the ARIMA, MLP, and LSTM models were developed and evaluated on their prediction of expenditures of certain prescription-based medications. The best performance on test data was obtained from the ensemble model, followed by MLP, LSTM, persistence, and ARIMA models. In the fourth experiment, multi-headed ML models (MLP, LSTM, convolutional neural network (CNN), ConvLSTM, and CNN-LSTM) were developed using multivariate time-series datasets for predicting patients' expenditures. The performance of these multi-headed models was compared against their single-headed counterparts and baseline vector autoregression (VAR) model. Results revealed that all the multi-headed models outperformed the corresponding single-headed architectures and the VAR model. In the last experiment, a novel generative adversarial network model (variance-based GAN or V-GAN) was developed that specifically minimized the difference in variance between model and actual data during model training to perform time-series predictions of medicine-related expenditures. The performance of V-GAN model was compared with other GAN-based variants and several ML models. Results revealed that the V-GAN model outperformed other models in correctly predicting medicine expenditures of patients. This thesis highlights the utility of various ML methods and feature engineering techniques for healthcare expenditure forecasting.

Keywords: *Machine learning, Electronic health records, Time-series prediction, Healthcare, Feature engineering, Frequent pattern mining, Medicine expenditures, Generative adversarial networks, Classification, Ensemble.*

Acknowledgments

First and foremost, I would like to thank God Almighty for bestowing upon me with this research opportunity. I am grateful for His provisions of joys, success, and challenges that manifested in me to endure hardships and succeed in this path to the dissertation. I am indebted to many people for their support throughout my graduate school career.

It is with immense gratitude and profound thanks that I acknowledge the support of my Ph.D. supervisor, Dr. Varun Dutt, in completing this thesis. His invaluable guidance, insightful discussions, priceless advice, constructive feedback, and constant encouragement are the bases for the success of this research. I want to express my sincere gratitude towards him for welcoming me into his research team, providing me an excellent research atmosphere, and supporting me to participate in national and international research venues. His mentorship has significantly influenced, for good, the way I understand and approach research work.

I have great pleasure in acknowledging my gratitude towards co-chairs of my dissertation committee: Dr. Aditya Nigam, Dr. Sriram Kailasam, and Dr. Deepak Swami for generously offering their time, guidance, and goodwill throughout my Ph.D. duration. I would also like to acknowledge Dr. Samar Agnihotri and Dr. Bharat Singh Rajpurohit for their feedback on my work. Without their constant push, I would not have been able to shape my thesis in its current form.

I would like to extend my deepest gratitude to Prof. Timothy Gonsalves, Ex-Director of IIT Mandi for inspiring students with his body of work.

I warmly thank my friend and colleague Abhinav Choudhury for creating a cordial working environment and supporting me in difficult times. I am thankful to my fellow lab members (Palvi Aggarwal, Zahid Maqbool, Medha Kumar, Akash k Rao, Ankush Pathania, Shashank Uttrani, Praveen, Priyanka, Harsh, Tushar, Gitanshu, and Shashank Bhargav) for their invaluable support in the lab. Special thanks to my friends, Merlin, Pavin, Pooja, Ayushi, Monika, Shweta, Sonika, Anita, Anindita, and Ramona. Thank you for your well-wishes and for being there whenever I needed a friend. I must also thank the Table-Tennis team, Hiking & Trekking club, and Gourmet club of IIT Mandi for providing the stress-free environment and lifetime memories.

I am extremely grateful to my grandfather Sh. Fateh Chand Shastri, my parents Mrs. Sneh Lata Kaushik and Mr. Surender Pal Kaushik, and my brother (Sourabh) for their unconditional love, support, and constant encouragement throughout my life. I am forever indebted to my parents for believing in me and cheering me up in bad times. Words fail me to express my appreciation to my beloved parents. The love, patience, and understanding shown by them during my research period are immeasurable. This journey would not have been possible if not for them, and I dedicate this milestone to them.

Lastly, a very special gratitude to Mr. Larry A. Pickett, Jr. and Mr. Sayee Natarajan from RxDataScience Inc. for providing me with the funding for my research work.

SHRUTI KAUSHIK

TABLE OF CONTENTS