KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Technology

Building personalised machine learning models in health informatics with limited datasets

Chetanya Puri

Supervisors:
Prof. dr. ir. Bart Vanrumste
Prof. dr. Stijn Luca
(Ghent University, Belgium)
Dr. ir. Gerben Kooijman
(Philips Research, Netherlands)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

January 2023

Building personalised machine learning models in health informatics with limited datasets

Chetanya PURI

Examination committee:
Prof. dr. ir. David Moens, chair
Prof. dr. ir. Bart Vanrumste, supervisor
Prof. dr. Stijn Luca, supervisor
(Ghent University, Belgium)
Dr. ir. Gerben Kooijman, supervisor
(Philips Research, Netherlands)
Prof. dr. ir. Jean-Marie Aerts
Prof. dr. ir. Joost Vennekens
Prof. dr. ir. Peter Karsmakers
Prof. dr. Francesca Spigarelli
(Università di Macerata, Italy)
Prof. dr. Milan Petkovic
(Philips Research, Netherlands)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

January 2023

Acknowledgements

I cannot begin to express my thanks to my supervisors Prof. Bart Vanrumste, Prof. Stijn Luca and Dr. Gerben Kooijman, for their guidance over the last four years. Prof. Bart's excitement for addressing complex challenges in machine learning for healthcare is remarkable, and I am grateful for his unwavering encouragement and remarkable patience throughout my graduate studies. Thank you for making yourself so approachable, even for the most minor technical or administrative issues. Prof. Stijn's incisive feedback and intellect are unparalleled and incredibly motivating. Dr. Gerben's input and persistent support during my time at Philips and in team meetings helped me a lot. I'd like to thank my supervisors for making our team meetings so stimulating and for thinking out loud with me. This has aided my development as a researcher, and I have learned a great deal from you. I will be eternally grateful for your kind assistance.

I would like to express my deepest appreciation to the members of my examination committee Prof. Jean-Marie Aerts, Prof. Joost Vennekens, Prof. Peter Karsmakers, Prof. Francesca Spigarelli, Prof. Milan Petkovic for agreeing to review my thesis. Thank you for taking the time to review my work and for providing helpful thoughts and comments to help me hone my work. Prof. David Moens deserves special thanks for agreeing to serve as chair of the examination committee and for organising the procedures so smoothly.

My graduate studies were part of the HEART project, which was funded by the European Union's Horizon 2020 research and innovation programme through the Marie Sklodowska-Curie grant. I'd like to thank

the project's partners, UNIMC and Philips Research, for their ongoing assistance in organising the practicalities and teaching me about the various interdisciplinary viewpoints.

Thanks to my colleagues, with whom I had a fantastic experience during my PhD adventure. Ahmed, Benjamin, Chunzhuo, Hannelore, Kymeng, Meng, and Sunil from the e-Media lab, and Heereen, Koustabh, Nuoya, Oleksandr, and Yuan from the HEART project. I sincerely value Koustabh and Yuan's friendship and long for the moments we spent together. Furthermore, I want to thank my close group of friends, Awadhesh, Devansh, Mahendra, Ashwin, Tushar, Sumit, Brijesh, Akhil, and Arpit - it is with you people that I learnt to be analytical. Special thanks go out to my dear friends Rituraj and Pratyusha, with whom I began my industrial research career. I dearly miss our breakfast and lunches together.

I want to express my sincere gratitude to everyone who supported me academically and encouraged me to pursue my interests.

Finally, I would like to express my gratitude to my lovely family for their constant support throughout my life. I am deeply indebted to my parents, Umesh and Kanchan, for always being there for me and supporting me to achieve my full potential. Tanya, my sister, deserves my heartfelt gratitude as well. Finally, I shall be eternally grateful to my wife Saswati for her never-ending support, patience, and optimism.

Chetanya Puri
January 5, 2023

Abstract

Healthcare services are being transformed by technological advancements and the availability of health-related data, from wearable device monitoring to treatment personalisation. Machine learning (ML) has the potential to harness this data by identifying patterns and developing prediction models to assist stakeholders and, ultimately, improve healthcare. The applications of machine learning in healthcare have grown exponentially, from drug discovery to preventative health. Given enough data, machine learning models can accurately predict or classify a disease. ML models can learn from longitudinal data collected over time and make predictions early enough to allow for the implementation of any necessary interventions.

Healthcare data, however, are susceptible to certain challenges that make the ML modelling difficult. In this dissertation, we would like to address some of the major challenges such as (i) the limited availability of data due to a small data corpus or the necessity to predict events in advance, (ii) personalisation of ML models that cater at an individual level as opposed to a *one-size-fits-all* approach, (iii) preserving privacy of an individual while maintaining a specific performance, and (iv) problems arising from missing data and how to handle them.

To demonstrate the pervasiveness of these challenges, a variety of healthcare applications are chosen. These applications encompass diverse health monitoring scenarios at an individual or institutional level. The modeling of weight gain in pregnant women during the course of their pregnancy to ensure a healthy pregnancy and postpartum life is an example of outside-hospital preventative health monitoring. Furthermore,

an application from a hospital setting is explored with the goal of predicting cognitive decline in Alzheimer's patients using a longitudinal dataset comprised of various data sources. Also, we investigate the prediction of infant mortality in a developing country from the perspective of population health management. Furthermore, we attempt to model the pain experienced by individuals performing repetitive tasks at work over time. The majority of these use cases require early prediction so that essential intervention can be carried out on time. As a result, it is imperative to develop machine learning models that can learn with only a few measurements from an individual.

The research developed in this thesis aims to address four research questions : (1) Can we predict a patient's health state with limited patient-specific time series data, (2) Can we detect infant mortality using structured tabular data with a very high percentage of missing data, (3) Can we create personalized machine learning models that can adapt over time to generate accurate predictions using few data points, and (4) Can we build machine learning models that can train in a secure manner while dealing with sensitive raw data without losing prediction performance? These broad research questions are further subdivided into individual application-based sub-objectives. To address these research questions and sub-objectives, we developed a number of techniques that can handle both N-dimensional time-series data and tabular data.

First, we propose a straightforward method for overcoming the limited availability of individual data where the underlying principle is to learn a non-person specific ML model from all the available individuals and then personalising it with the target user's available data.

Second, we propose a more complex method that follows the similar principle and combines a localised method for generating informative priors before learning a regression model. The localised method selects individuals from the training data, whose health history is similar to that of the individual of interest. This is then followed by a powerful Gaussian processes-based method of learning from the selected subset.

Thirdly, we offer a privacy-preserving learning paradigm based on the aggregation of ML models learned from an individual's data. This strategy differs from the conventional centralised technique in which raw

data is collected and shared to a central server. The findings of this study demonstrated that a good privacy-performance trade-off is feasible.

Through a case study, we discuss the current flaws in handling missing data with off-the-shelf techniques. We demonstrate the importance of identifying the mechanisms through which the data can be missed, as well as the inconsistencies that can creep into the model if these mechanisms are not properly studied during the exploratory phase. The findings of this case study suggest a technique for detecting biased features, which, if not handled carefully, can give the ML model a false sense of predictive power.

In conclusion, the concepts presented in this doctoral dissertation are relevant to addressing difficulties in modelling healthcare-related tasks using machine learning.

Beknopte samenvatting

Gezondheidszorgdiensten worden getransformeerd door technologische vooruitgang en de beschikbaarheid van gezondheidsgerelateerde gegevens, van monitoring van draagbare apparaten tot personalisatie van behandelingen. Machine learning (ML) heeft het potentieel om deze gegevens te benutten door patronen te identificeren en voorspellingsmodellen te ontwikkelen om belanghebbenden te helpen en uiteindelijk de gezondheidszorg te verbeteren. De toepassingen van machine learning in de gezondheidszorg zijn exponentieel gegroeid, van het ontdekken van medicijnen tot preventieve gezondheid. Met voldoende gegevens kunnen machine learning-modellen een ziekte nauwkeurig voorspellen of classificeren. ML-modellen kunnen leren van longitudinale gegevens die in de loop van de tijd zijn verzameld en voorspellingen vroeg genoeg doen om de implementatie van eventuele noodzakelijke interventies mogelijk te maken.

Gegevens in de gezondheidszorg zijn echter onderhevig aan bepaalde uitdagingen die de ML-modellering moeilijk maken. In dit proefschrift willen we enkele van de belangrijkste uitdagingen aanpakken, zoals (i) de beperkte beschikbaarheid van gegevens vanwege een klein gegevenscorpus of de noodzaak om gebeurtenissen van tevoren te voorspellen, (ii) personalisatie van ML-modellen die inspelen op een individueel niveau in tegenstelling tot een *one-size-fits-all*-benadering, (iii) het behoud van de privacy van een individu met behoud van een specifieke prestatie, en (iv) problemen die voortvloeien uit ontbrekende gegevens en hoe hiermee om te gaan.

Om de alomtegenwoordigheid van deze uitdagingen aan te tonen, wordt

gekozen voor een verscheidenheid aan toepassingen in de gezondheidszorg. Deze toepassingen omvatten diverse scenario's voor gezondheidsmonitoring op individueel of institutioneel niveau. Het modelleren van gewichtstoename bij zwangere vrouwen tijdens hun zwangerschap om een gezonde zwangerschap en postpartum leven te garanderen, is een voorbeeld van preventieve gezondheidsmonitoring buiten het ziekenhuis. Verder wordt een toepassing uit een ziekenhuisomgeving onderzocht met als doel cognitieve achteruitgang bij Alzheimerpatiënten te voorspellen met behulp van een longitudinale dataset bestaande uit verschillende databronnen. Ook onderzoeken we de voorspelling van kindersterfte in een ontwikkelingsland vanuit het perspectief van populatiegezondheidszorg. Bovendien proberen we de pijn te modelleren die wordt ervaren door individuen die repetitieve taken op het werk in de loop van de tijd uitvoeren. De meeste van deze use-cases vereisen vroege voorspelling, zodat essentiële interventie op tijd kan worden uitgevoerd. Als gevolg hiervan is het absoluut noodzakelijk om machine learning-modellen te ontwikkelen die kunnen leren met slechts een paar metingen van een persoon.

Het onderzoek dat in dit proefschrift is ontwikkeld, heeft tot doel vier onderzoeksvragen te beantwoorden: (1) Kunnen we de gezondheidstoestand van een patiënt voorspellen met beperkte patiëntspecifieke tijdreeksgegevens, (2) Kunnen we kindersterfte detecteren met behulp van gestructureerde tabelgegevens met een zeer hoog percentage van ontbrekende gegevens, (3) kunnen we gepersonaliseerde machine learning-modellen maken die zich in de loop van de tijd kunnen aanpassen om nauwkeurige voorspellingen te genereren met weinig datapunten, en (4) kunnen we machine learning-modellen bouwen die op een veilige manier kunnen trainen terwijl ze omgaan met gevoelige onbewerkte gegevens zonder de voorspellingsprestaties te verliezen? Deze brede onderzoeksvragen zijn verder onderverdeeld in individuele toepassingsgerichte deeldoelen. Om deze onderzoeksvragen en subdoelen te beantwoorden, hebben we een aantal technieken ontwikkeld die zowel N-dimensionale tijdreeksgegevens als tabelgegevens kunnen verwerken.

Ten eerste stellen we een eenvoudige methode voor om de beperkte beschikbaarheid van individuele gegevens te overwinnen, waarbij het onderliggende principe is om een niet-persoonsspecifiek ML-model te

leren van alle beschikbare personen en dit vervolgens te personaliseren met de beschikbare gegevens van de doelgebruiker.

Ten tweede stellen we een complexere methode voor die hetzelfde principe volgt en een gelokaliseerde methode combineert voor het genereren van informatieve priors voordat een regressiemodel wordt geleerd. De gelokaliseerde methode selecteert personen uit de trainingsgegevens waarvan de gezondheidsgeschiedenis vergelijkbaar is met die van de persoon van belang. Dit wordt vervolgens gevolgd door een krachtige op Gaussiaanse processen gebaseerde methode om te leren van de geselecteerde subset.

Ten derde bieden we een privacy-behoudend leerparadigma op basis van de aggregatie van ML-modellen die zijn geleerd van de gegevens van een persoon. Deze strategie verschilt van de conventionele gecentraliseerde techniek waarbij onbewerkte gegevens worden verzameld en gedeeld met een centrale server. De bevindingen van dit onderzoek laten zien dat een goede afweging tussen privacy en prestaties haalbaar is.

Aan de hand van een casestudy bespreken we de huidige tekortkomingen in het omgaan met ontbrekende gegevens met kant-en-klare technieken. We demonstreren het belang van het identificeren van de mechanismen waardoor de gegevens kunnen worden gemist, evenals de inconsistenties die in het model kunnen kruipen als deze mechanismen niet goed worden bestudeerd tijdens de verkennende fase. De bevindingen van deze casestudy suggereren een techniek voor het detecteren van vertekende kenmerken, die, als ze niet zorgvuldig worden behandeld, het ML-model een vals gevoel van voorspellende kracht kan geven.

Concluderend, de concepten die in dit proefschrift worden gepresenteerd, zijn relevant voor het aanpakken van problemen bij het modelleren van zorggerelateerde taken met behulp van machine learning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A high quality of life depends on one's health. Healthcare professionals evaluate individual health by gathering data (often clinical data) and analysing it. In addition to individual data, they also leverage data acquired from the general population to determine an individual's current state of health and recommend diagnoses to enhance it. Thus, data is crucial for analysing human health. Machine learning (ML) is a field of computer science that harnesses the power of data to automatically learn patterns from it. Several industries, such as computer vision, natural language processing and robotics, have been boosted by ML's use of data. Consequently, ML has enormous potential for automating the extraction of insights from raw data to aid caregivers and ultimately improve health. However, there are several challenges when applying machine learning to healthcare. This section introduces these difficulties inherent in the application of machine learning in healthcare, particularly preventative healthcare. The section 1.1 introduces the healthcare-related motivation followed by various challenges in applying machine learning in such scenarios. In section 1.2, we present a variety of healthcare-related applications that face these issues. Section 1.3 presents the key research objectives addressed in this thesis together with sub-objectives. Section 1.4 provides an outline of the subsequent chapters.

1.1 Introduction and challenges in healthcare informatics

Prognosis of a disease is central to the practice of medicine. The estimation of a disease's likelihood may assist health policymakers and physicians make decisions about identification, screening based on detected severity, and treatment in high-risk groups [1]. Evidence based medicine is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" [2]. The evidence-based approach integrates a medical practitioner's clinical expertise with the best available external evidence from systematic research [2]. Without clinical expertise, there are several risks that might occur in practice because strong evidence alone might not be applicable for an individual patient [2]. Similarly, the practice methods should be kept up to date by incorporating the most recent best evidence for the sake of a patient. Clinical prediction models (CPM) integrate patients' characteristics to evaluate the probability of health risk in certain individuals. For public health, targeted preventive interventions can be devised for at-risk individuals. In a clinical setting, CPMs can help the physicians in tracking the probability of a diagnosis for a certain patient [1].

A clinical or non-clinical prediction model can assist in determining the physical health of an individual in or out of a hospital setting. The prediction models can aid medical practitioners in identifying early signs of disease. They can then provide care at an individual level based on the posed severity of a prediction outcome. Typically, prediction models are built on population averages. *Precision medicine* is defined by National Research Council (US) as "*the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into sub-populations that differ in their susceptibility to a particular disease, in the biology or prognosis of those diseases they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not*" [3]. Fields such as 'precision medicine' are driving

huge research from a *one-size-fits-all* approach to a personalised approach that accounts for individual variability within groups of individuals [4], [5]. Predictive modelling is primarily done to estimate future values based on the historical data from either an individual or a population.

Machine learning (ML) is a branch of artificial intelligence that enables a computer to learn from data automatically by gradually improving its performance via experience, similar to how people learn. In order to gain insight into an individual, machine learning involves the collection and examination of data to identify patterns and significant trends. This pattern discovery and creation of prediction model by ML is enabled by the availability of sufficient annotated data around a specific task. With the digitisation of medical data and the advent of low-cost sensory hardware, a huge amount of data is made available for analysis. Typically, this information is stored throughout time to account for an individual's health history. This is termed longitudinal data since it has been collected over time. This data can range from wearable data for daily health monitoring, for example physical activity, heart rate to Electrocardiogram (ECG) or yearly/monthly data such as electronic health records (EHRs) in a hospital setting.

Machine learning (ML) applications in healthcare have seen a tremendous growth, from drug discovery [6] to preventive health [7], [8]. Much of this is made feasible by a massive shift from theoretical studies involving small proof-of-concepts to large-scale real-world applications, particularly in computer vision and natural language processing [9]. Combining large-scale data from heterogeneous sources from an individual (e.g. blood test report, X-ray, physician's notes, etc.) can help build machine learning models that are able to make individual predictions based on subjects' individual characteristics in a timely manner [10]. Despite these advances, there are several bottlenecks associated with ML applications in healthcare. Raw electronic health data, for example, are challenging to analyse using machine learning as they may contain a huge number of unrecorded values for one individual that would otherwise be present in another. This could be because the doctor ordered the tests based on past observations of an individual and might not be needed in another. Thus, many of the difficulties stem from the fact that data collection in healthcare is conducted largely to assist care, as opposed to allowing

data analysis [10].

In this dissertation, we will explore the training of machine learning algorithms while considering the issues that typical healthcare data may have, such as limited data availability, personalisation, missing data, and privacy concerns. These are discussed in detail in the following sections.

1.1.1 Limited data availability

Intervention in healthcare can be provided as soon as a condition is diagnosed. Some disorders simply require a snapshot of data at a single point in time to predict a health outcome, for example the association of high body mass index with greater cardiovascular risk [11]. Lifestyle based interventions are fruitful if the signs of a particular health-related deterioration are detected early. Physicians assess the previous health record of a person to identify a suitable trajectory of necessary lifestyle changes. Consequently, they are able to intervene in a timely manner by utilising the knowledge that they gained through time from treating other patients. A longitudinal study, as opposed to snapshot-based health prediction, consists of measurements where the health status is tracked over time. A machine learning model that can capture the differences in health status can be learned either through individual data or combining individual data with data from other subjects. If such a model can forecast a person's health outcome as early as possible, necessary interventions may be more successful. This is, however, difficult for the following reasons,

1. **Not enough subjects**: The statistical learning from data depends on the diversity in a dataset. If the study conducted has a small group of participants and the data being collected are only a few observations, they might not be a good representation of every individual's characteristics. This could result in less informative models that perform poorly with fresh data from an unseen subject. The capacity of ML models to perform well on unseen data is known as their *generalisation* ability. Machine learning requires that the collected dataset has enough observations that are a good

representation of the complete population to make a generalised model.

However, limited subjects alone might not always be an issue. For example, in human activity detection accelerometer data is collected that is sampled at very high-sampling rate (order of 100 measurements per second). This data is collected from a small number of participants engaging in a variety of activities (actions) with high intra-activity variability, enabling the development of generalised, high-performing models [12]. Unless the measurements are sampled at a high rate, limited subjects-based modelling is further aggravated by limited measurements within those subjects.

2. **Not enough measurements**: To forecast a state as early as possible, machine learning models must learn from a small number of observations of an individual. In a clinical situation when a patient makes several hospital visits, for instance, the ability to predict an individual's health state as early as the first or second visit presents a limited measurements problem.

3. **Not enough ground truth**: Usually the classification of the health condition requires training of a machine learning model against given data and its labels. Typically, two to three domain experts must annotate segments of lengthy recordings, such as an ECG or X-ray image, where a health abnormality is present. This combination of labels and data, known as labelled data, serves as training data for a machine learning model. Obtaining ground truth is a costly endeavour that needs the significant time and effort of domain specialists.

1.1.2 Personalisation

Personalisation in healthcare is an active field of interest, primarily because it is user-centric [13]. Precision medicine as described in section 1.1 offers solutions that can change the course of diagnosis of an individual from that of an average patient to a more individualistic approach [14]. Machine learning models attempt to create a global model based on the available labelled data, and based on these models,

inferences are formed about unlabelled data from an individual. The more diverse the data, the more variations can be captured for a specific learning task. This diversity is often achieved by collecting more data in different situations in a particular use-case which is then used to train such algorithms. If the model performs as well on an unseen data as it did on the training data, it is said to have generalised well. To produce personalised predictions for a user, one may collect data from various sources related to that user and develop models that are specific to each user. This can generate a user-specific model, but it takes a substantial amount of individual data. In addition, the model may suffer from bias with respect to a given user and be incapable of learning from different data. Consequently, models must be sufficiently generic to be applied to a large population but also be capable of adapting to the peculiarities of the individual.

But, the need for predicting early as stated below can limit the capabilities of a machine learning model for creating personalised models as it is required to predict with as little data as possible. For instance, it is vital to be able to forecast cognitive decline in Alzheimer's patients if individual patterns can be modeled as soon as the patient is admitted. This, however, severely limits the availability of historical data for that person, as no past history is available. We will discuss such use-cases in detail in chapters 2, 3 and 6. One way of tackling such a challenge is by incorporating more information. This can be done in two ways as follows,

Heterogeneous sources of data

Integrating data from multiple heterogeneous sources for a single patient can help obtain information about the characteristics of individual physiology. For example, instead of just modelling the historical data of the cognitive decline in an Alzheimer's patient over time, including other measurements like imaging modality based magnetic resonance imaging (MRI) or Positron emission tomography (PET) can improve the predictive performance of future values of cognitive decline. This will be covered in greater detail in the chapter 3.

Intra-subject discovery

As we discussed earlier, a doctor in an evidence-based approach combines external evidence with individual clinical expertise. Similarly, a machine learning model can be learnt by discovering patterns in the population and adapting these models with respect to limited personal information available from the individual of interest. For example, sepsis is a potentially life threatening hospital condition that occurs as a result of infection. Early detection and treatment of sepsis are crucial with each hour of delay associated with a 4-8% increase in mortality [15]. Although not particularly designed to diagnose sepsis, the National Early Warning Score (NEWS) is the standard score used to assess deterioration in a hospital setting. It detects in advance cardiac arrest or death by using certain thresholds and triggers for vital signs and other lab measures. There are many false alarms in an in-hospital setting since NEWS is not especially focused on sepsis. The authors in [16] trained a neural network architecture on a massive corpus of electronic health records (EHRs) data containing information about vitals (e.g. heart rate) and lab measurements measured over time. A deep neural network was trained on all the available subjects, as well as the test individual's information to learn the detection of sepsis prior to the event occurrence. When tuned with individual data, the patterns learned over time for the progression of vitals and lab measurements in other subjects achieved a significant performance boost while reducing false alarms in sepsis detection compared to the National Early Warning Score, which only uses individual data and compares to several thresholds based on average population risk scores.

1.1.3 Missing data

Healthcare data is evolving and becoming significantly more complex. Before being fed to any out-of-the-box machine learning estimation or classification model, population health datasets such as survey-based data and individual electronic health records (EHR) must be properly curated. Errors in recording entries or outliers may lead to considerable amount of missing data. Also, physicians may order specific diagnostic tests based on a patient's medical history, resulting in dynamic selection

of variables that are not present for every individual. Thus, there are dependencies in the creation of variables that must be carefully accounted for. If the dataset used to train machine-learning algorithms are biased, the system may be prejudiced in practice. These biases can occur as a result of collection of data or learning algorithms or both. For instance, it is straightforward to have more data from a normal individual than a specific diseased individual, hence generating an imbalance in the collected data to classify a specific disease vs a normal individual. If this imbalance in the dataset is not addressed effectively, machine learning models may be biased toward the normal class rather than the ill class. Similarly, ML algorithms may try to model the missingness, when data is missing for only one class as compared to others. Therefore, it is essential to handle the missing data with care.

Standard machine learning approaches can train models using either incomplete casewise deletion or missing data imputation to handle missing data. Incomplete casewise deletion refers to the elimination of instances that are missing one or more variables [17]. However, casewise deletion may result in the loss of statistical knowledge that could be informative. Therefore, missing data imputation, i.e. filling in missing data, is the most prevalent method for handling absent data. But it is important to understand the assumptions in which data can be missing. Using off-the-shelf imputation techniques without understanding the underlying mechanisms of the missing data might lead to biased results that might perform worse on an unseen dataset. When the underlying assumption is missing not at random (MNAR), the missingness of a variable is related to the unobserved data, for example, an illicit drug user would be hesitant to answer a drug usage related question in a survey. In such cases, modelling the missing data is the only way to obtain an unbiased estimate of the parameters, but this requires domain expertise in the missing variable, which is usually impractical. We will discuss this in detail in the chapter 5.

1.1.4 Privacy

Learning from data requires access to sensitive data from users, especially in a healthcare scenario. This is typically done, centrally where the data

is pooled at a central server and models are learned at this centralised server. There are several challenges associated with centralising the data, ranging from data protection and privacy challenges such as regulatory, ethical and legal challenges, to technical ones. For instance, regulatory constraints limit dataset quality by deleting information to protect the privacy of a specific user from a machine learning standpoint. Recital 26 [18] of the General Data Protection Regulation (GDPR), the European data protection law defines anonymous data as "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable" [19]. Complete anonymisation, restricting access, and exchanging healthcare data securely is a difficult and sometimes impossible task. In [20], the authors demonstrate that patients can be re-identified from an anonymised dataset with only a few demographical indicators, even when the dataset is highly incomplete, thereby challenging the technical and legal adequacy of de-identification release-and-forget models to meet the modern anonymisation standards established by regularity authorities, for instance General Data Protection Regulation (GDPR). As a result, these legal, ethical constraints, albeit necessarily, present a number of data-related issues when it comes to traditional centralised learning, further limiting the predictive ability of a machine learning model.

There has been growing research on building machine learning models where the learning paradigm is shifting from the traditionally centralised approach to a more decentralised approach [21] where a user has more control over their own data. Federated learning (FL) is described as the collaborative learning of a machine learning task by a federation of edge-devices by sharing updates of models learned locally on private data that are aggregated at a central server. After that, the aggregated updates are used to modify the local models. Federated learning (FL) was introduced by Google [21] and has become a promising approach because it addresses privacy and data governance issues by enabling machine learning from distributed data. Taking the model to the data and not the opposite has multiple advantages, for example reduction in data-duplication of large scale institutional data for local training.

Similarly, data transfer over borders has its own privacy concerns because

of different regulations in different countries. One way of utilising cross-border data within the scope of legal constraints is by learning a model in one geographical location such that it reveals close to zero personal information. This model can then be fine tuned with the data from another geographical location. This is known as transfer learning. We demonstrate in chapter 2 where a model learnt on population in one demographic is transferred to other demographic, much like fine-tuning a model, while keeping personal data private.

1.2 Use-cases

We discussed the key challenges that can arise when attempting to model a dataset. These challenges are not exclusive to the healthcare sector; they also exist in other fields where machine learning models must be learned when one or more of these challenges arise.For the purposes of this dissertation, only healthcare-related use cases will be considered. Before discussing the overarching research objective and the sub-objectives, we would like to explain the healthcare applications that this dissertation addresses and the related challenges.

1.2.1 Gestational weight gain management

Weight management is a crucial lifestyle-related issue that affects people of all ages and races in increasingly obesogenic societies [22]. Pregnant women are one of the most vulnerable population groups. The Institute of Medicine (IOM) has updated the suggested set of guidelines for how much weight women of various Body Mass Index (BMI) categories should acquire during pregnancy in order to promote optimal health for both the mother and her child [23]. Only about 30% of the pregnant women end up having normal weight gain in association with the recommended guidelines [24]. Several studies have found a link between gestational weight gain and pregnancy outcomes [25], [26]. Associated risks involve immediate and long-term dangers to mothers, including fetal macrosomia and post-partum weight retention, which can lead to maternal obesity.

Necessary lifestyle interventions can be devised if the weight gain trend is detected early in the pregnancy.

To accomplish so, we collected weight measurements from expecting women during their pregnancy and investigated the trajectory of weight gain. These measurements were acquired in a home environment using a mobile application linked to a Wi-Fi connected weighing scale. They were urged but not obliged to record measures at least once a week in order to continue their normal routine and be close to a realistic data collection scenario. While following the said data collection principle, the data collected had various challenges that are typically also present in other applications of machine learning. For example, because the data is self-reported, it is far more scarce than was thought to be. Even if a participant records daily, for an average pregnancy lasting about 40 weeks, measurements would not exceed 280 points in time. Some women were far more conscious than others, noting their weights on a regular basis, resulting in readings that were evenly spaced in time. Others had less incentive to do so, resulting in some cases with two nearest time points as widely apart as 4 weeks. Fig. 1.1 shows two densely sampled data (subject 9 and 65) vs sparsely sampled data (subject 16 and 53) for modelling.

Eighty subjects in the Netherlands and 153 subjects in China participated in this study. More details about the dataset are presented later in section 2.3. As outlined in the section 1.1.1, this data faces the challenge of limited avaiability as the number of subjects is small and the number of time measurements each subject possesses is also low. In addition, there is a lot of sparsity in a given time series making the time series non-uniformly sampled.

Furthermore, an intervention can be successful if the deviation in weight gain from the normal trajectory is detected as soon as possible. In terms of machine learning, this means that the learning algorithm only has access to an individual's measurements for training until a specified day, thus limiting the availability of personal data required to develop personalised models that can monitor individual trajectory. This challenge of personalisation detailed in section 1.1.2 further limits the availability of data.

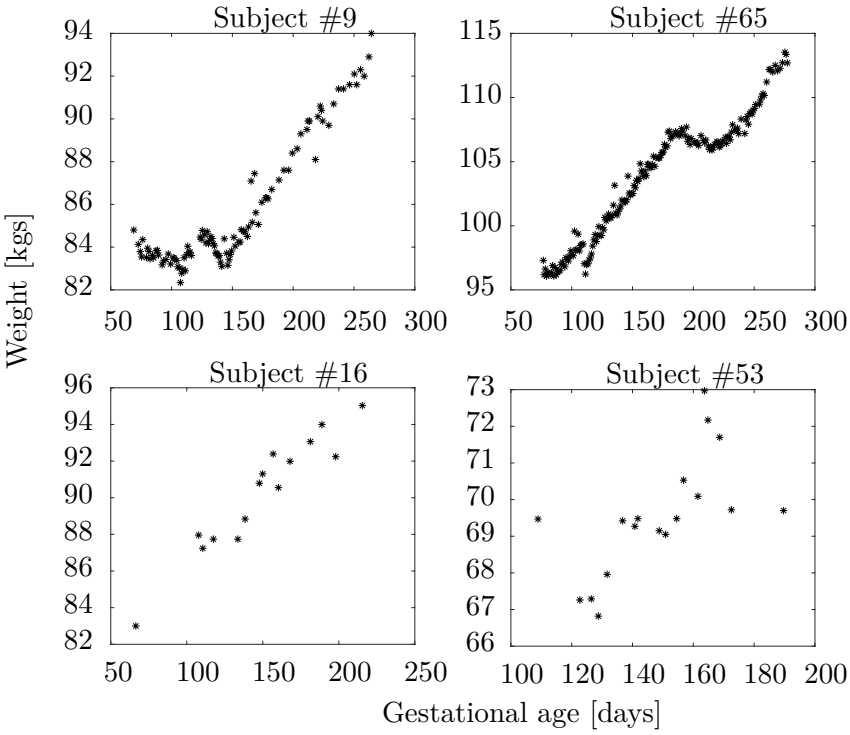


Figure 1.1: Gestational weight data with respect to time with subject 9 and 65 having abundant measurements vs data from subjects 16 and 53 that have very few measurements.

Typically, the raw weight gain data is shared to a central server along with other meta-data where this data is processed to learn efficient ML models. Since the weight gain data and other meta-data are personal and sensitive, there is a need to address the privacy concerns of sharing the raw weight data to a central data controller over time via a mobile application as described in section 1.1.4.

1.2.2 Alzheimer's disease prediction

Alzheimer's disease (AD) is the most prevalent form of dementia and a neurodegenerative ailment. It is urgent and difficult to predict the onset

of symptoms in the early stages of this progressive condition [27]. The design of clinical trials and the development of therapeutic interventions rely on the correct identification of patients in the earliest stages of disease, when therapies are most likely to be beneficial. The clinical status of an Alzheimer's patient is determined by regularly employed cognitive scores, specifically the mini mental state examination (MMSE) [28], the Washington University Clinical Dementia Rating Sum of Boxes score (CDRSB) [29], and the AD Assessment Scale-Cognitive subtest score (ADAS-Cog13) [30].

Generally, Alzheimer's disease data comprises numerous measurements taken over the course of multiple visits. This data might consist of data from several heterogeneous sources ranging from clinical notes to imaging data. Fig. 1.2 shows the dataset collected as part of the TADPOLE challenge [31] by the Alzheimer's Disease Neuroimaging Initiative (ADNI) consortium[1] [32].

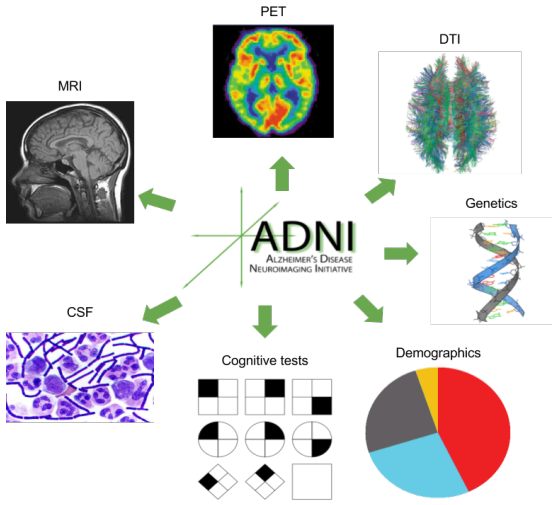


Figure 1.2: Multi modal dataset of an Alzheimer patient consists of image data, genetics, cognitive tests and demographic information taken over several visits to the hospital [32]

[1] http://adni.loni.usc.edu/

The data from 1737 patients collected every six months over a period of 120 months includes imaging modalities like magnetic resource imaging (MRI), positron emission tomography (PET) and diffusion tensor imaging (DTI), cerebro-spinal fluid (CSF) markers of amyloid beta and tau-deposition, cognitive assessments measured in the presence of a clinical expert, genetic information such as alipoprotein E4 (APOE4) status from DNA samples and general demographic information [31].

The Alzheimer's dataset consists of similar challenges as were presented for gestational weight gain (GWG) use-case in section 1.2.1. The GWG data is univariate, meaning there is only one input variable (time) that influences the output variable (weight gain), whereas the Alzheimer's data is multivariate (there are multiple input variables), where the output variable (cognitive decline) is also dependent on the multivariate input variables in addition to the historical output measurements. The aforementioned challenges present themselves as data from several visits are absent among individuals, and early prediction of cognitive decline reduces personal data available for training.

Multiple input variables are missing across several visits making the data highly sparse for modelling the cognitive decline. 95 subjects out of 1737 were selected such that data from at least ten visits (out of 24 total visits) are present and missing data are no more than 82.5% of the input variables. This decision was made in accordance with [33] so that we can compare our approach with their state-of-the-art results. Furthermore, in order to design effective clinical trials, it is critical to be able to predict cognitive deterioration as early as possible, hence limiting the accessible input data for developing personalised models. These difficulties were outlined in sections 1.1.1 and 1.1.2.

1.2.3 Infant mortality prediction

Child mortality remains a major challenge in India and is responsible for approximately 39.1 deaths per 1,000 live births in 2017 [34]. Child mortality as a pregnancy outcome is considered a major attribute in building efforts to preventive antenatal care thus reducing infant mortality. We chose a publicly available healthcare survey dataset conducted over women that underwent pregnancy in several states in India [35].

We select data from the open government platform in India where the Indian government has provided open access to datasets, documents, etc. for public use. This dataset is also collected as part of a joint initiative between government of India and US government. A number of 355 features in the Women pregnancy schedule (WPS) dataset [35] are present in the form of questionnaire, with fields related to social, economic, health status or demographic indicators as well as the outcome of pregnancy (live or stillbirth).

The data contains rows that represent individual subjects and columns that contain a particular questionnaire answer also referred to as a feature. Missing data (as described in section 1.1.3) is often handled by deleting the row containing missing values or by imputing the missing values by training an imputation algorithm on a subset of the given dataset.To obtain a reliable model, the subset utilised for training the imputation procedure should be complete (i.e. no missing data). However, there is no complete subset of data in the aforementioned questionnaire data. To construct robust infant mortality prediction models, it is therefore critical to deal with missing data in different ways. This is addressed in detail in chapter 5.

1.2.4 Pain Management at Workplace

Musculoskeletal disorders (MSD) are injuries and illnesses affecting the muscles, nerves, tendons, joints, cartilage, and spinal discs. Work-related musculoskeletal disorders (WMSD) are conditions where the work environment and work performance significantly contribute to the MSDs [36]. WMSDs are the most prevalent occupational health issue and the leading cause of absenteeism at workplace affecting about sixty percent of the workforce in Europe [37]. “Pain chronification” is the transformation of temporary pain into persistent pain at work and is one of the long-term effects of MSDs. Preventive pain management reduces the likelihood of developing chronic pain. Additionally, there are several models that integrate different biological, social and psychological factors to the perception of pain [38], [39], [40] that illustrate how various persons experience pain, which results in the subjectivity of pain assessments. Thus, predicting pain in a personalised manner is essential for preventing

pain persistence. It is crucial that both patients and medical practitioners have the education and abilities necessary to manage pain correctly [41]. Therefore, it is essential to conduct an accurate assessment of pain in advance, as this can help a person limit their expectations and fears about returning to work.

To achieve this, we recruited 99 participants from multiple sectors in Belgium and asked them to keep a daily diary of their work-related pain on a scale from 0 (no pain) to 100 (maximum pain). Due to the users' reluctance to record their pain levels at the end of the day, these *self-reported* pain levels are missing a significant amount of data across time. This presents several modeling difficulties for individual pain data. In order to predict pain as early as possible, only few measurements from an individual can be used for training a personalised model. In addition, this limited time-series data are sampled irregularly. These difficulties were discussed in section 1.1.1. Because pain data is self-reported, it is inherently subjective, making it difficult to create a one-size-fit-all model that can be used to predict pain levels in all the individuals. We will discuss in chapter 6 each person's pain fluctuates around *their baseline* pain measurement. This subjective baseline is based on their perception of pain and varies across individuals. Consequently, the pain-forecasting model requires a much-needed personalisation, as highlighted in section 1.1.2.

In this dissertation, we would like to define our research questions that lie at the intersection of these broad challenges with applications that span one if not multiple of these challenges. Figure 1.3 is a Venn diagram of these challenges in each use case.

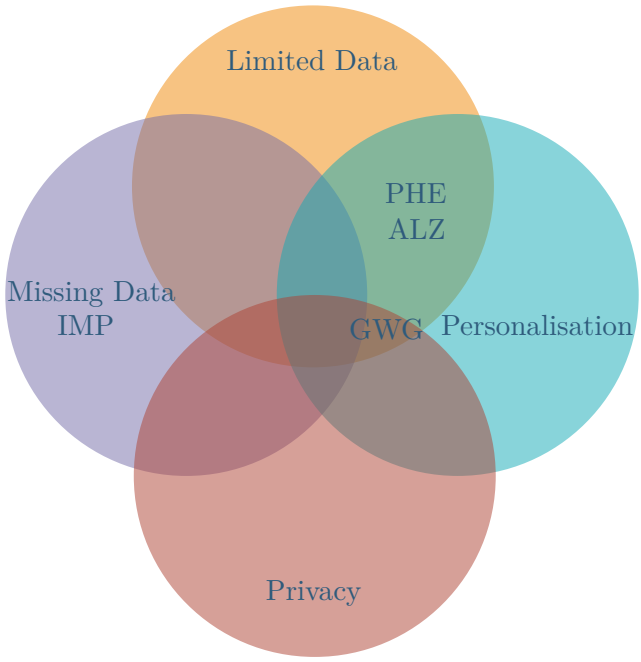


Figure 1.3: Challenges in Healthcare and use-cases addressing them. Abbreviations : Pain Management at Workplace (PMW), Alzheimer's disease estimation (ALZ), Gestational weight gain estimation (GWG), Infant mortality prediction (IMP)

1.3 Research Objective

When it comes to using machine learning to learn models, whether for classification or regression, the availability of data is now a must, and its abundance enables the deployment of data-intensive methods such as deep learning. In applications such as those discussed in this thesis, however, we shall see that this is not always the case. Data can be restricted for a variety of purposes, including measurement, privacy and/or personalisation as mentioned in section 1.1. *The primary objective of this dissertation is to investigate if machine learning models can attain reliable performance when given with a range of data-related challenges, notably in the context of time series data in healthcare.* With

this in mind, we wish to develop machine learning algorithms capable of addressing these obstacles, prompting us to formulate the following research questions (RQ):

RQ1: **Can we predict a patient's health state with limited patient-specific time series data?**

SUB-OBJECTIVES

(a) Can we reliably predict the gestational weight gain in expecting women from sparse weight data collected up to certain days in pregnancy (Chapters 2, 3)?

(b) Can we reliably predict the cognitive decline in Alzheimer's patients with sparse observations in time from multiple input sources (Chapter 3)?

(c) Can we reliably predict the pain measurements in workers from various sectors with sparse historical pain measurements (Chapter 6)?

RQ2: **Can we detect infant mortality using structured tabular data with a very high percentage of missing data?**

SUB-OBJECTIVES

(a) How to handle missing data when the missingness assumptions are biased (Chapter 5)?

(b) How to select features for unbiased handling of missing values? (Chapter 5)?

RQ3: **Can we create personalized machine learning models that can adapt over time to generate accurate predictions using few data points?**

SUB-OBJECTIVES

(a) Can we create personalised models that can reliably predict the gestational weight gain in expecting women, cognitive decline in alzheimer's patients and pain levels in a workplace (Chapters 2, 3, 6)?

(b) Can we develop an alignment technique for time series' from patients with Alzheimer's that accounts for the fact that individuals have variable degrees of cognitive decline at the time of recruitment? (Chapter 3)?

(c) How quickly can these models start predicting accurately the end-of-pregnancy weight gain (Chapter 2), cognitive decline in an alzheimer patient (Chapter 3), pain levels in a workplace (Chapter 6)?

RQ4: **Can we build machine learning models that can train in a secure manner while dealing with sensitive raw data without losing prediction performance?**

SUB-OBJECTIVES

(a) Can we construct time series forecasting models with a sufficient performance-to-privacy trade-off to predict the gestational weight gain in pregnant women? (Chapters 4)?

(b) Is it possible to transfer the model and fine-tune it without transferring the complete raw data across two geographical regions (Chapter 2)?

1.4 Outline

We would now like to outline the remainder of this dissertation, having provided an introduction to the difficulties associated with ML applications in healthcare.

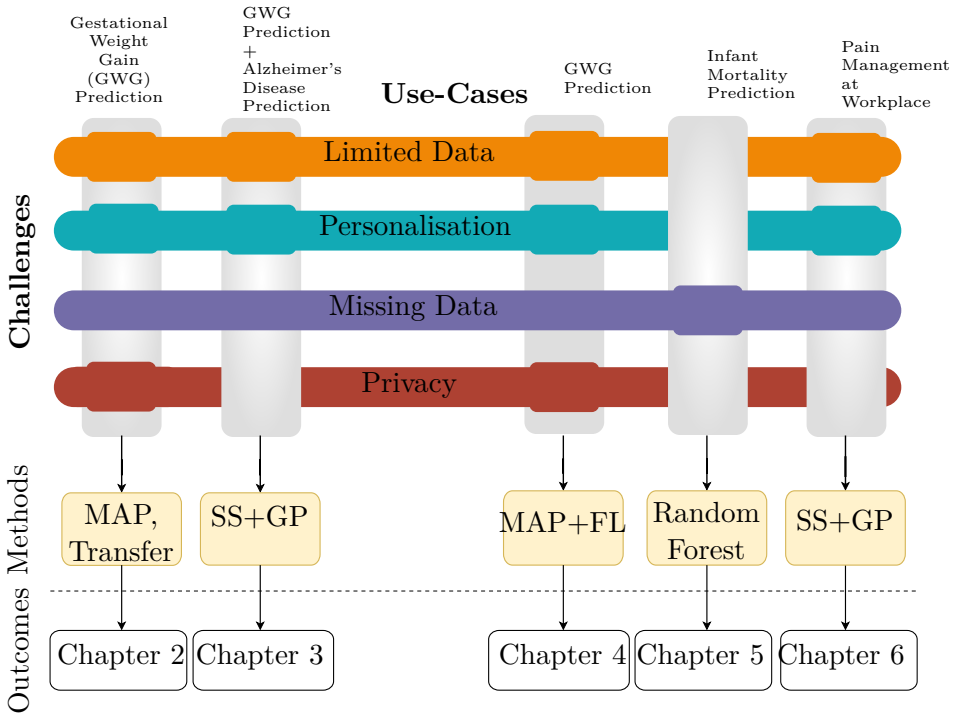


Figure 1.4: Outline schematic of the thesis spanning multiple challenges. The abbreviations are RQ : Research Questions; MAP : Maximum-a-posteriori approach ; SS : subset selection; GP : Gaussian Processes Regression; FL : Federated Learning

Chapter 2 introduces the problem of gestational weight gain estimation in expecting women. We introduce the problem as a multistep-ahead forecasting problem with the purpose of predicting multiple future steps. We introduce a unique weight measurements dataset collected during an individual's pregnancy that is sparse and non-uniformly sampled. We present a unique preprocessing technique in which raw data are transformed based on pre-pregnancy baseline data and medical guidelines. Then, a personalised machine learning model is learned by modifying the parameters of a general model based on available personal data.

Chapter 3 Introduces another human health-related application in which clinical trial design is affected by the early prediction of cognitive

loss in Alzheimer's patients. This longitudinal dataset comprising several modalities, such as imaging data and questionnaire data, is missing multiple visits, resulting in missing data. We present a novel localised learning based on dynamic time warping in which subjects similar to the test subject are used to learn priors and then forecasting model is learnt on this selected subset using Gaussian processes to achieve state-of-the-art forecasting results.

In **chapter 4**, we suggest a new federated approach as an alternative to the current centralised approach for predicting pregnancy weight gain. Instead of pooling raw data on the central server, this is accomplished by learning small local models that are aggregated on the central server.

Chapter 5 presents a large-scale questionnaire based dataset that concerns infant mortality prediction. This chapter highlights the importance of the underlying missing mechanisms before performing blind missing data imputation, especially in healthcare. A novel empirical approach is presented that can aid in the identification of biased features that, if incorrectly imputed, would create wrong models that do not perform well on a new data.

Chapter 6 presents data collected from individuals who reported their daily pain levels at work. This chapter describes pre-processing techniques that can be applied to data that takes into account the subjective nature of pain measurements. Then, we use the proposed localised learning strategy to select a subset of individuals whose pain data are similar to that of a given individual, as described in chapter 3 to model the pain in a personalised way.

Bibliography

[1] E. W. Steyerberg *et al.*, *Clinical prediction models*. Springer, 2019.

[2] D. L. Sackett, W. M. Rosenberg, J. M. Gray, R. B. Haynes, and W. S. Richardson, *Evidence based medicine: What it is and what it isn't*, 1996.

[3] National Research Council (US) Committee, *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (The National Academies Collection: Reports funded by National Institutes of Health). Washington (DC): National Academies Press (US), 2011, ISBN: 978-0-309-22222-8.

[4] N. J. Schork, "Personalized medicine: Time for one-person trials", *Nature*, vol. 520, no. 7549, pp. 609–611, 2015.

[5] S.-I. Lee, S. Celik, B. A. Logsdon, S. M. Lundberg, T. J. Martins, V. G. Oehler, E. H. Estey, C. P. Miller, S. Chien, J. Dai, *et al.*, "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia", *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[6] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, *et al.*, "Applications of machine learning in drug discovery and development", *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[7] J. Dunn, R. Runge, and M. Snyder, "Wearables and the medical revolution", *Personalized medicine*, vol. 15, no. 5, pp. 429–448, 2018.

[8] L. Neubeck, N. Lowres, E. J. Benjamin, S. B. Freedman, G. Coorey, and J. Redfern, "The mobile revolution—using smartphone apps to prevent cardiovascular disease", *Nature Reviews Cardiology*, vol. 12, no. 6, pp. 350–360, 2015.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health", *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020.

[11] C. Friedemann, C. Heneghan, K. Mahtani, M. Thompson, R. Perera, and A. M. Ward, "Cardiovascular disease risk in healthy children and its association with body mass index: Systematic review and meta-analysis", *BMJ*, vol. 345, 2012.

[12] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors", *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.

[13] S. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K.-S. Kwak, "The internet of things for health care: A comprehensive survey", *IEEE access*, vol. 3, pp. 678–708, 2015.

[14] K. B. Johnson, W.-Q. Wei, D. Weeraratne, M. E. Frisse, K. Misulis, K. Rhee, J. Zhao, and J. L. Snowdon, "Precision medicine, ai, and the future of personalized health care", *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, 2021.

[15] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock", *Critical care medicine*, vol. 34, no. 6, pp. 1589–1596, 2006.

[16] J. Futoma, S. Hariharan, K. Heller, M. Sendak, N. Brajer, M. Clement, A. Bedoya, and C. O'brien, "An improved multi-output gaussian process rnn with real-time validation for early sepsis detection", in *Machine Learning for Healthcare Conference*, PMLR, 2017, pp. 243–254.

[17] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[18] *Recital 26 - Not Applicable to Anonymous Data*, en-US. [Online]. Available: `https://gdpr-info.eu/recitals/no-26/` (visited on 08/04/2022).

[19] P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council", *Regulation (eu)*, vol. 679, p. 2016, 2016.

[20] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models", *Nature communications*, vol. 10, no. 1, pp. 1–9, 2019.

[21] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[22] N. Heart, Lung, B. Institute, N. I. of Diabetes, and K. D. (US), *Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: the evidence report*. National Heart, Lung, and Blood Institute, 1998.

[23] K. M. Rasmussen, P. M. Catalano, and A. L. Yaktine, "New guidelines for weight gain during pregnancy: What obstetrician/gynecologists should know", *Current opinion in obstetrics & gynecology*, vol. 21, no. 6, p. 521, 2009.

[24] R. F. Goldstein, S. K. Abell, S. Ranasinha, M. Misso, J. A. Boyle, M. H. Black, N. Li, G. Hu, F. Corrado, L. Rode, *et al.*, "Association of gestational weight gain with maternal and infant outcomes: A systematic review and meta-analysis", *Jama*, vol. 317, no. 21, pp. 2207–2225, 2017.

[25] R. Gaillard, B. Durmuş, A. Hofman, J. P. Mackenbach, E. A. Steegers, and V. W. Jaddoe, "Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy", *Obesity*, vol. 21, no. 5, pp. 1046–1055, 2013.

[26] L. A. Gilmore, M. Klempel-Donchenko, and L. M. Redman, "Pregnancy as a window to future health: Excessive gestational weight gain and obesity", in *Seminars in Perinatology*, Elsevier, vol. 39, 2015, pp. 296–303.

[27] J. L. Cummings, "Challenges to demonstrating disease-modifying effects in Alzheimer's disease clinical trials", *Alzheimer's & Dementia*, vol. 2, no. 4, pp. 263–271, 2006.

[28] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state": A practical method for grading the cognitive state of patients for the clinician", *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[29] C. P. Hughes, L. Berg, W. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia", *The British journal of psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.

[30] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease.", *The American journal of psychiatry*, 1984.

[31] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, D. C. Alexander, *et al.*, "Tadpole challenge: Prediction of longitudinal evolution in Alzheimer's disease", *arXiv preprint arXiv:1805.03909*, 2018.

[32] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative", *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, 2005.

[33] K. Peterson, O. Rudovic, R. Guerrero, and R. W. Picard, "Personalized gaussian processes for future prediction of alzheimer's disease progression", *NeurIPS Workshop on Machine Learning for Healthcare.*, 2017.

[34] *India demographics profile*, en, `https://www.indexmundi.com/india/demographics_profile.html`. (visited on 03/08/2021).

[35] *Census of india : Annual health survey 2010 - 11 fact sheet*, `https://www.censusindia.gov.in/vital_statistics/AHSBulletins/Factsheets.html`. (visited on 02/11/2021).

[36] Centers for Disease Control and Prevention and others, "Work-related musculoskeletal disorders & ergonomics", *Centers for Disease Control and Prevention, Atlanta*, 2020. [Online]. Available: https://www.cdc.gov/workplacehealthpromotion/health-strategies/musculoskeletal-disorders/index.html.

[37] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Safer and Healthier Work for All — Modernisation of the EU Occupational Safety and Health Legislation and Policy*. COM, 2017, p. 9. [Online]. Available: https://ec.europa.eu/social/BlobServlet?docId=16874&langId=en.

[38] J. W. Vlaeyen and S. J. Linton, "Fear-avoidance and its consequences in chronic musculoskeletal pain: A state of the art", *Pain*, vol. 85, no. 3, pp. 317–332, 2000.

[39] M. I. Hasenbring, D. Hallner, B. Klasen, I. Streitlein-Böhme, R. Willburger, and H. Rusche, "Pain-related avoidance versus endurance in primary care patients with subacute back pain: Psychological characteristics and outcome at a 6-month follow-up", *Pain*, vol. 153, no. 1, pp. 211–217, 2012.

[40] S. Bunzli, A. Smith, R. Schütze, I. Lin, and P. O'Sullivan, "Making sense of low back pain and pain-related fear", *journal of orthopaedic & sports physical therapy*, vol. 47, no. 9, pp. 628–636, 2017.

[41] B. Morlion, F. Coluzzi, D. Aldington, M. Kocot-Kepska, J. Pergolizzi, A. C. Mangas, K. Ahlbeck, and E. Kalso, "Pain chronification: What should a non-pain medicine specialist know?", *Current Medical Research and Opinion*, vol. 34, no. 7, pp. 1169–1178, 2018.

Chapter 2

Personalised Modelling For Early Prediction of Gestational Weight Gain

This chapter was previously published as:
C. Puri, G. Kooijman, F. Masculo, S. V. Sambeek, S. D. Boer, J. Hua, N. Huang, H. Ma, Y. Jin, F. Ling, G. Li, D. Zhang, X. Wang, S. Luca, and B. Vanrumste, "A personalized bayesian approach for early intervention in gestational weight gain management toward pregnancy care", *IEEE Access*, vol. 9, pp. 160 946–160 957, 2021

Abstract

Pre-pregnancy body mass index and weight gain management are associated with pregnancy outcomes in expecting women. Poor gestational weight gain (GWG) management could increase the risk of adverse complications. These risks can be alleviated by lifestyle-based interventions if an undesired GWG trend is detected early on in the pregnancy. Current literature lacks analysis of gestational weight gain data and tracking the pregnancy over time. In this work, we

collected longitudinal gestational weight gain data from women during their pregnancy and model their weight measurements to predict the end-of-pregnancy weight gain and classify it in accordance with the medically recommended guidelines. The measurement frequency of the weights is often very variable such that segments of data can be missing and the need to predict early utilising few data points complicates data modelling. We propose a Bayesian approach to forecast weight gain while effectively dealing with the limited data availability for early prediction. We validate on diverse populations from Europe and China. We show that utilising individual's data only up to mid-way through the pregnancy, our approach produces mean absolute errors of 2.45 kgs and 2.82 kgs in forecasting end-of-pregnancy weight gain on these populations respectively, whereas the best of state-of-the-art yields 8.17 and 6.60 kgs on respective populations. The proposed method can serve as a tool to keep track of an individual's pregnancy and achieve GWG goals, thus supporting the prevention of excessive or insufficient weight gain during pregnancy.

2.1 Introduction

In this increasingly obesogenic society, weight management is a key lifestyle-related condition that affects people of all ages and ethnicities. One of the most important demographic groups affected by this is pregnant women. 47% of the pregnant women gain too much weight over the gestational period and around 23% tend to gain too little weight during their pregnancy [2]. Institute of Medicine (IOM) updated the recommended set of guidelines [3] on how much weight women in different BMI categories should gain during their pregnancy to encourage optimal health for the mother and her child (Table 2.1). With only 30% of the women in the normal weight category after pregnancy [2], most of the women do not follow the guidelines or realize too late in the pregnancy that an intervention or control of the weight gain is necessary.

Risks associated with undesired weight gain: There have been several studies that associate gestational weight gain with pregnancy related outcomes. For example, excessive Gestational Weight Gain (GWG) can pose several short and long term risks for the mothers such

Table 2.1: 2009 IOM guidelines [3] for weight gain and rate of weight gain during pregnancy with respect to BMI. The guidelines assume a weight gain of $0.5 - 2$ kg in the first trimester of pregnancy.

Pre-pregnancy Body Mass Index (BMI) category	Mothers of singletons		
	Total weight gain (in kgs)	Weight gain in the first trimester (kgs)	Rate of weight gain in the second and third trimesters (kg/wk)
Underweight ($<$(18.50 kg/m^2)	$12.70 - 18.14$	$0.50 - 2.00$	$0.45 - 0.59$
Normal-weight ($18.50 - 24.90$ kg/m^2)	$11.34 - 15.88$		$0.36 - 0.45$
Overweight ($25.0 - 29.9$ kg/m^2)	$6.80 - 11.34$		$0.23 - 0.32$
Obese (30.0 kg/m^2)	$4.99 - 9.07$		$0.18 - 0.27$

as fetal macrosomia and post-partum weight retention leading to maternal obesity [4]. Women entering into pregnancies with high pre-pregnancy Body Mass Index (BMI) are at increased risk for gestational diabetes [5]. It can also result in large-for-gestational-age infants and/or caesarean delivery or other labor and delivery complications [2]. In terms of risks for the offsprings, Oken et al. [6] and Sridhar et al. [7] found that exceeding the recommended guidelines was associated with a 46% increase in odds of having an overweight/obese child after adjusting for maternal pre-pregnancy BMI, race/ethnicity, age at delivery, education, child age, birth-weight, gestational age at delivery, gestational diabetes, parity, infant sex, total metabolic equivalents, and dietary pattern. Additionally, adverse cardiovascular diseases in later stages of the offspring's life is also reported in [8]. On the contrary, gaining too little weight during pregnancy is also not considered healthy. Evidence for a correlation exists between inadequate weight gain and perinatal mortality. Davis et al. studied over 100,000 records from the National Center for Health Statistics (NCHS) 2002 Birth Cohort Linked Birth/Infant Death Data and indicated that inadequate gestational weight gain is highly associated with increased odds of infant death up to 1 year after death [9]. Other reported risks include increased risk of pre-term birth or small-for-gestational-age infants [2] or failure to initiate breastfeeding [3].

There have been several factors associated with the undesired gestational

weight gain such as age, ethnicity, genetics [10], [4] which are fixed. Apart from these fixed factors, modifiable factors related to lifestyle such as amount of physical activity and food intake also show a high correlation with the gestational weight gain [11]. Several intervention studies [12], [13] showed that lifestyle based interventions can improve the outcome of gestational weight gain, if the intervention is timely, preferably initiated before the start of the pregnancy [14].

In this work, we aim to reliably predict the gestational weight gain using the weight measurements from initial days of the pregnancy. Our proposed approach uses the weight gain measurements from other subjects in the training data to generate prior information about the (personal) model of the test subject. The model is then trained on the available limited data of the test subject along with the generated prior information resulting in an increase in the performance of the overall system, which we discuss later. Our proposed solution can help prenatal care providers in risk assessment during a pregnancy and provide adaptive coaching to the mothers. Moreover, mothers can track the *rate* of weight gain and use the model to monitor weight gain, thus reducing GWG related risks at the end of their pregnancy.

Real life weight measurements are used that are mostly *self-reported* (measurements consistent with regular mid-wife/ hospital visits) by 233 expecting mothers during their pregnancy in Europe and China. We formulate this as an absolute weight prediction problem with the end goal of predicting the weight at the end of the pregnancy and classifying if the weight is within the IOM recommended guidelines or not. We have restricted our analysis to the mothers with singleton pregnancy for this study. Data from mothers expecting more than one child is very rare to obtain. Also, the guidelines for gestational weight gain consider singleton mothers [3].

Lifestyle interventions can be done in the form of personal coaching by traditional health-care providers, or eHealth mobile-application based coaching or a mix of both [15], [16]. A schematic diagram of the solution following a mix of both is provided in Fig. 2.1 where recorded weight measurements are sent for processing along with meta-data and feedback/alerts can be shared with the individual and/or caregivers. Recommended weight gain during pregnancy varies from person to person

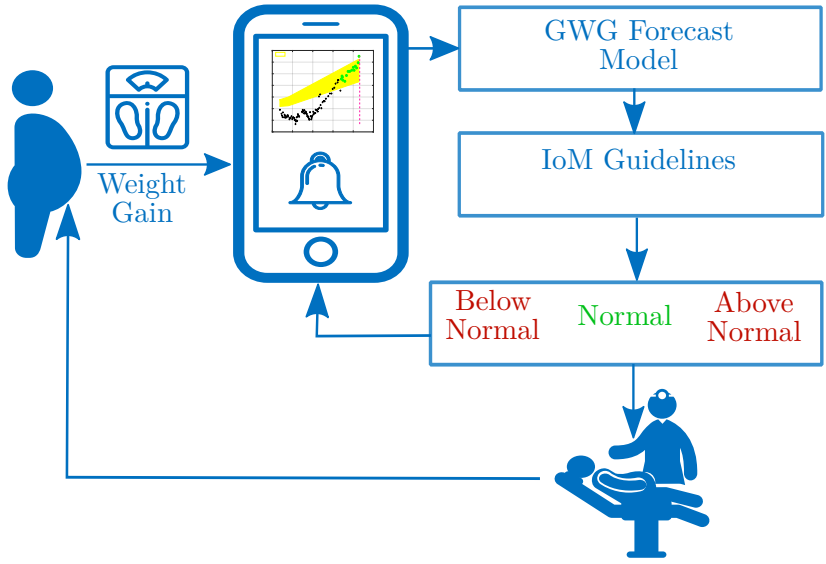


Figure 2.1: Schematic diagram of GWG weight estimation

based on their BMI ranges. Women with underweight pre-pregnancy BMI are expected to gain weight at a higher rate than women that were overweight before pregnancy. This calls for personalization of the learning method. Additionally, it is important to note that the problem of estimation is a multi-step forecasting problem, which means that we train a model using self-reported weights at the start of the pregnancy period (e.g. first 180 days) and use this model to forecast the weight at the end of pregnancy (around day 270-280).

The primary contributions of this paper are,

- collecting weight gain data from women *across time* during the course of their pregnancy in a practical scenario (example, via self reporting),
- building personalised model for GWG trend prediction using as little personal data as possible,
- unique raw weight gain transformation approach that reduces inter-BMI class variance for accurate GWG modelling.

- validating the proposed approach across *different geographical regions* and examine the model transfer to evaluate the generalizability of the approach.

2.2 Related Works

Various works [2], [17] study the association of pre-pregnancy BMI, the amount of weight gain during pregnancy and the health risks to mothers and infants. Diana et. al. propose a differential equation model for pregnant women in different pre-pregnant BMI category that predicts GWG that results from changes in energy intakes [18]. This method helps predict the impact of changes in dietary energy intake on GWG in these BMI categories. Although this tool helps in understanding the dietary needs, there exists no studies that helps pregnant women understand and track the absolute weight gain during their pregnancy in a personalised manner based on individual's weight gain data.

Several time series forecasting methods exist in the literature such as state-space approaches e.g. Kalman filtering [19] and Autoregressive Integrated Moving Average (ARIMA) [20] that learn structures from the time series data for few-step ahead predictions, given sufficient historical personal data. However, they tend to converge towards the mean as the forecast horizon increases, thus giving inaccurate predictions [21]. Alternatively, a polynomial model of lower order (1, 2, or 3) can be used to estimate the end-of-pregnancy weight gain using weight measurements from the start of the pregnancy period, if enough reliable weight measurements collected uniformly over time are available for training. However, there are two major challenges i) weight measurement data are often noisy, incomplete, sparse and non-uniformly sampled due to the self-reported nature, ii) available data from the initial few days of the pregnancy are often limited, complicating the training of a model. Polynomial fit using maximum likelihood estimation (MLE) or ARIMA suffer from at least one of these challenges. In the recent decade, deep learning approaches such as Long short-term memory (LSTM) networks [22] have become popular and they are known to model the non-linearity among the datasets very well for forecasting. However, lack of availability of individual training

data pertaining to early prediction in our case, and high number of trainable parameters associated makes them unsuitable in the practical scenario at hand. Fig. 2.2 illustrates the early prediction of weight gain measurements for two subjects using state-of-the-art methods.

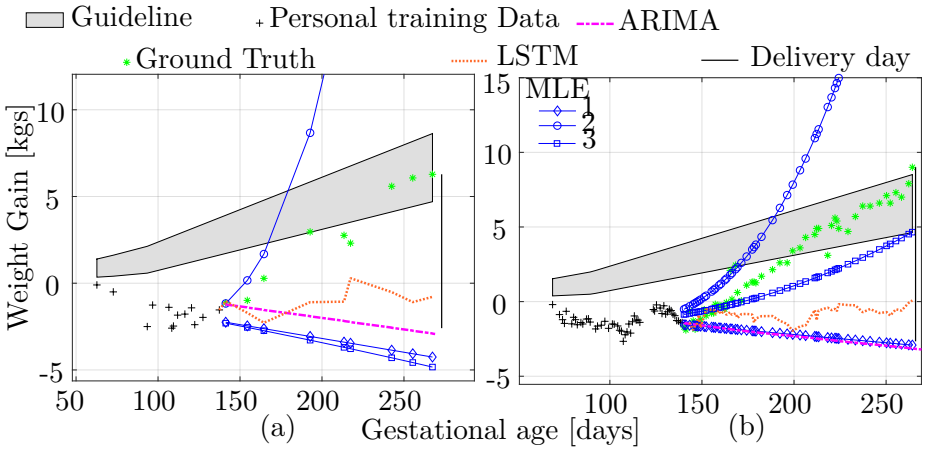


Figure 2.2: State-of-the-art methods, MLE with (order = 1, 2, 3), ARIMA, LSTM to predict the end-of-pregnancy weight-gain for i^{th} subject. The prediction accuracy that can be obtained from the data shown in left subplot is superior to the accuracy using the data that is shown in the right. The data shown in (a) is of a higher quality at the start of the pregnancy period (i.e. more uniformly sampled, less sparse).

In this paper, we experiment with *parametric* Bayesian regression to model the time series data. In contrast to the previous work [23], our algorithm incorporates meta-data such as pre-pregnancy weight and BMI to improve the efficacy. We also test the generalization capability of our proposed algorithm on new data from a different geographic region by training our proposed approach on data from one region and testing the learned model on another region. We show that our approach outperforms state-of-the-art in early weight gain prediction by using data from training subjects to create an a-priori model estimate and then tuning it to model the test subject's limited available personal training observations. To our knowledge, this is the first study that uses few weight measurements from the early days of pregnancy to estimate the end-of-pregnancy weight gain.

2.3 Database

Data from diverse pregnant women were collected in Europe ($\mathcal{D}_E$) and China ($\mathcal{D}_C$). Women that were in their gestational week 5 or later were recruited randomly from midwife practices in Europe and private hospitals in China. The details of these datasets are described below:

$\mathcal{D}_E$

Two midwife locations recruited 90 participants in Eindhoven, The Netherlands over a period of three months. However, data from only 80 women were considered for the final analysis as 10 subjects dropped out of the study due to miscarriage or technical problems. 40% of the women were experiencing their first pregnancy, while for another 40% it was their second and 20% had more than two previous pregnancies. Education level was generally high with more than 60% having at least college degree. This means that women with low and no education are under-represented in this data. This may be relevant as it is well known that Socio Economic Status (SES) is correlated with nutrition, weight-gain and lifestyle factors in general. 9% of women reported smoking. The weight data was collected using a WiFi-connected weight scale, Withings WS30[1]. The participants were asked to log their weights weekly and the recorded weight data was sent to the cloud via a mobile application. Participants were instructed to weight themselves at least once per week. However, post-hoc analysis shows that participants recorded 2.0 ± 1.4 measurements per week. Overall, 86% of participants were adherent to the study measurement protocol with most of the women measuring more than 1 time per week.

$\mathcal{D}_C$

Two hospitals recruited 366 subjects living in Shanghai, China. After filtering the subjects that had a disease or left the study in the middle, 153 women's pregnancy weight gain data were considered. About 2/3 of

[1] https://www.withings.com/

subjects were having their first pregnancy and only few were pregnant for the 3rd or 4th time. The overwhelming majority of the subjects have received at least college degree, which together with a median household income of $2811 - 4200$ US$ per month indicates their relatively high socio-economic status. The weight data were collected weekly in home as well as on regular visits to the hospital. The in-hospital weight data was highly correlated with the in-home collected data, indicating that the in-home measured data were reliable for further analysis.

Additional meta-data such as age, height and pre-pregnancy weight were also collected for both the datasets. The participants provided an informed consent pre-data collection, and the study was approved by the Internal Ethics Committee for Biomedical Experiments of the involved organizations (ICBE Reference number 2015-0079 and 2017-0189 for $\mathcal{D}_E$ and $\mathcal{D}_C$ respectively).

Table 2.2: Dataset description for data from different geographies

Dataset Attribute	$\mathcal{D}_E$ (80 Subjects) Mean $\pm$ Std	$\mathcal{D}_C$ (153 Subjects) Mean $\pm$ Std
Age (years)	31.01 ± 3.50	32.10 ± 3.51
Height (meters)	1.69 ± 0.07	1.64 ± 0.05
Pre-pregnancy weight (kgs)	69.01 ± 15.10	57.90 ± 9.77
Pre-pregnancy BMI (kgs/m^2)	24.11 ± 4	21.40 ± 3.22
Delivery (days)	277.00 ± 10.00	273.20 ± 12.20
Weight Gained (kgs)	13.70 ± 4.70	14.10 ± 4.30
Number of recorded weight gain samples	59.83 ± 41.02	17.21 ± 7.30

It is important to note that $\mathcal{D}_C$ is sparser than $\mathcal{D}_E$ in time. The maximum number of samples for an individual present in $\mathcal{D}_C$ is 37 and in $\mathcal{D}_E$ this is 230. This is one of the reasons why modelling such a data is difficult. The data in $\mathcal{D}_C$ shows less variability among individual subjects in terms of pre-pregnancy BMI class (Table 2.2).

Table 2.3 shows the data distribution in our sample dataset pre and post-pregnancy for under, within and over guidelines. Interestingly, our sample data-set's distribution is close to that in [2], which is obtained

from a large population of more than a million women, with almost half of the women gaining above the recommended guidelines. This further strengthens the need for this study.

Table 2.3: End of pregnancy weight class with respect to IOM guidelines for both datasets (represented as $\mathcal{D}_E(\mathcal{D}_C)$)

Pre-pregnancy BMI class	#Sub	Distribution post-pregnancy		
		Underweight	Normal	Overweight
Underweight	3 (23)	1 (5)	2 (13)	0 (5)
Normal	45 (110)	11 (15)	15 (50)	19 (45)
Overweight	32 (20)	4 (3)	8 (5)	20 (12)
80 (153) subjects		16 (23)	25 (68)	39 (62)
(Class %)		**20%** (**15.1%**)	**31.2%** (**44.4%**)	**48.8%** (**40.5%**)

2.4 Methods

Notation. We are given a population of $N-1$ subjects that, by means of self-reporting tools, acquired $N-1$ time series of gestational weight gain measurements as $\mathcal{X} = \{(\mathbf{x}^1, \mathbf{y}^1), \cdots\text{-}(\mathbf{x}^{N-1}, \mathbf{y}^{N-1})\}$, where $\mathbf{x}^i = [t_1^i, t_2^i, t_3^i, \cdots, t_{m_i}^i]$ represents the input gestational days up to delivery day $t_{m_i}^i$ and $\mathbf{y}^i = [y_1^i, y_2^i, y_3^i, \cdots, y_{m_i}^i]$ represents the output weight gain for i^{th} subject, where $y_k^i = y(t_k^i)$. It is important to note here that t_1^i *does not necessarily equal* $t_1^j, i, j \in \{1, 2, \cdots, N-1\}$. This is because each subject acquires measurements at different times according to their personal preferences and adherence to data collection.

Additionally, we are given individual weight measurements from test subject's (N^{th} subject) initial t_d^+ days of pregnancy data, $\mathcal{D} = \{(t_1^+, y_1^+), (t_2^+, y_2^+), \cdots, (t_d^+, y_d^+)\}$. We call this the *personal-training data*. Weight gain data from $N-1$ training subjects over entire gestational period is called the *public-training data*.

The objective is to try to learn function(s) f from given public and individual training data, such that,

$$y_i^+ = f(t_i^+) + \epsilon_i \tag{2.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent and identically distributed (i.i.d) according to a Gaussian.

Our parametric approach learns parameters' information a-priori from the public-training data. We then use this generated prior-knowledge along with the personal-training data to build personalised models and learn f. The individual weight gain in future at delivery time t_m^+ is forecasted using the learned model f and $y_m^+ = y(t_m^+) \approx f(t_m^+)$.

Firstly, before we discuss the parametric regression, we introduce a pre-processing technique for transformation of input data using IOM guidelines.

2.4.1 Transformation using IOM guidelines

We subtract the pre-pregnancy weight to calculate the weight gain data. After using pre-pregnancy weight to standardize the data, we propose to transform the obtained weight gain data by introducing a non-linear trend controlled by a subject's pre-pregnancy BMI. This trend is based on the pre-pregnancy-BMI classes and their respective expected rate of weight gains in accordance with IOM guidelines. Lower and upper guidelines are obtained using linear interpolation based on the total weight gain and the rate of weight gain that are suggested by the IOM guidelines (Table 2.1). For the i^{th} subject with pre-pregnancy BMI class bmi^i at time t_k, this means that the following extrapolation is proposed.

$$L_{bmi}(t_k) = \begin{cases} \left(\dfrac{\Delta_{min} * t_k}{90}\right) & 0 \leq t_k \leq 90, \\ \Delta_{min} + \left(\dfrac{(\alpha_{min}^{bmi^i} - \Delta_{min}) * (t_k - 90)}{t_{max} - 90}\right) & 90 \leq t_k \leq t_{max} \end{cases} \tag{2.2}$$

$$U_{bmi}(t_k) = \begin{cases} \left(\frac{\Delta_{max} * t_k}{90}\right) & 0 \leq t_k \leq 90, \\ \Delta_{max} + \left(\frac{(\alpha_{max}^{bmi^i} - \Delta_{max}) * (t_k - 90)}{t_{max} - 90}\right) & 90 \leq t_k \leq t_{max} \end{cases} \tag{2.3}$$

$$\rho_{bmi}(t_k) = \frac{U_{bmi^i}(t_k) - L_{bmi^i}(t_k)}{2} \tag{2.4}$$

where $bmi^i = \{$'underweight','normal','overweight','obese'$\}$ is calculated using pre-pregnancy BMI, $\Delta_{min} = 0.5$ kgs, $\Delta_{max} = 2$ kgs are the first trimester (90 days) minimum and maximum gains respectively according to the guidelines (Table 2.1). α_{min}^{bmi} and α_{max}^{bmi} are the minimum and maximum allowed weight gains during second and third trimester in IOM guidelines (Table 2.1). For example, for $bmi =$ 'underweight' class, $\alpha_{min}^{underweight} = 12.7,\ \alpha_{max}^{underweight} = 18.14$. Assuming $t_{max} = 280$ days as the day of delivery, Fig. 2.3 show the guidelines and ρ_{bmi} for different BMI classes following eqn. (2.2), (2.3) and (2.4) respectively. The transform and inverse-transform weight-gain operation can be performed respectively using eqn. (2.4) as follows:-

$$y_{transform}(t_k^i) = y(t_k^i) \times \rho_{bmi}(t_k^i) \tag{2.5}$$

$$y_{detransform}(t_k^i) = \frac{y_{transform}(t_k^i)}{\rho_{bmi}(t_k^i)} \tag{2.6}$$

It should be noted that we are introducing a non-linear trend in our pre-processing approach by multiplication with $\rho(t)$ instead of standard division-based normalisation. As Fig. 2.3a and Table 2.1 suggests, an underweight woman is allowed a larger weight-gain bandwidth than an obese woman. We multiply the original weight gain data with this bandwidth factor ρ calculated based on pre-pregnancy BMI class that allows an underweight woman to have a wider window of weight gain than an obese woman (Fig. 2.3b). Such scaling ensures that the data across different subjects and BMIs are closer to each other in transformed space for a better fit. Fig. 2.4 shows how original and transformed data scale across each BMI class among all the subjects in dataset $\mathcal{D}_E$.

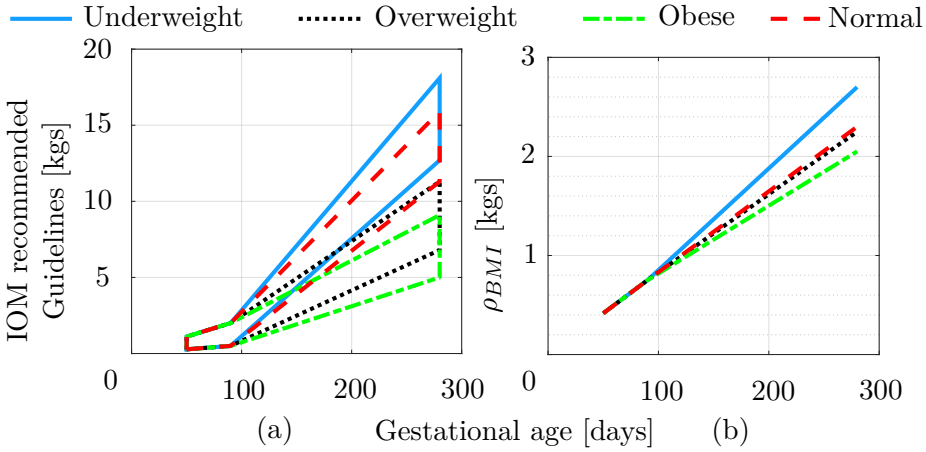


Figure 2.3: Transforming the weight gain data using extrapolated guidelines based on different BMI classes

2.4.2 Regression

We can fit a p^{th}-order polynomial with $f = w_0 + w_1 t + w_2 t^2 + \cdots + w_p t^p$ in eq. (2.1) and estimate the coefficients $\mathbf{w} = [w_0, w_1, \cdots, w_p]^T$ by maximizing the likelihood ($\mathcal{L}$) over an individual's personal-training data $\mathcal{D}$, $\mathcal{L}(\mathbf{w}) = P(\mathcal{D}|\mathbf{w})$,

$$\hat{\mathbf{w}}_{MLE} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{d} p(y_i^+|t_i^+; \mathbf{w}) \tag{2.7}$$

Eq. (2.7) refers to the model learnt from the individual's sparse limited observations up to given t_d days. Next, we exploit the public-training data and find the maximum likelihood point estimates (MLE) of $\hat{\mathbf{w}}^i$ for each individual time series in the public-training data following eq. (2.7). If we assume gaussianity over the distribution of $\mathbf{w}$ such that $\mathbf{w} \sim \mathcal{N}(\mu_{\hat{\mathbf{w}}}, \Sigma_{\hat{\mathbf{w}}})$, we can find a closed-form solution of maximum-a-posterior (MAP), $\mathbf{w}_{MAP}$ analytically. Here, $\mu_{\hat{\mathbf{w}}} = mean([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^{N-1}]^T)$, $\Sigma_{\hat{\mathbf{w}}} = cov([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^{N-1}]^T)$ are mean and covariances of the polynomial coefficients $\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^{N-1}$ that are each obtained using the individual gestational weight gain data from each of the $N-1$

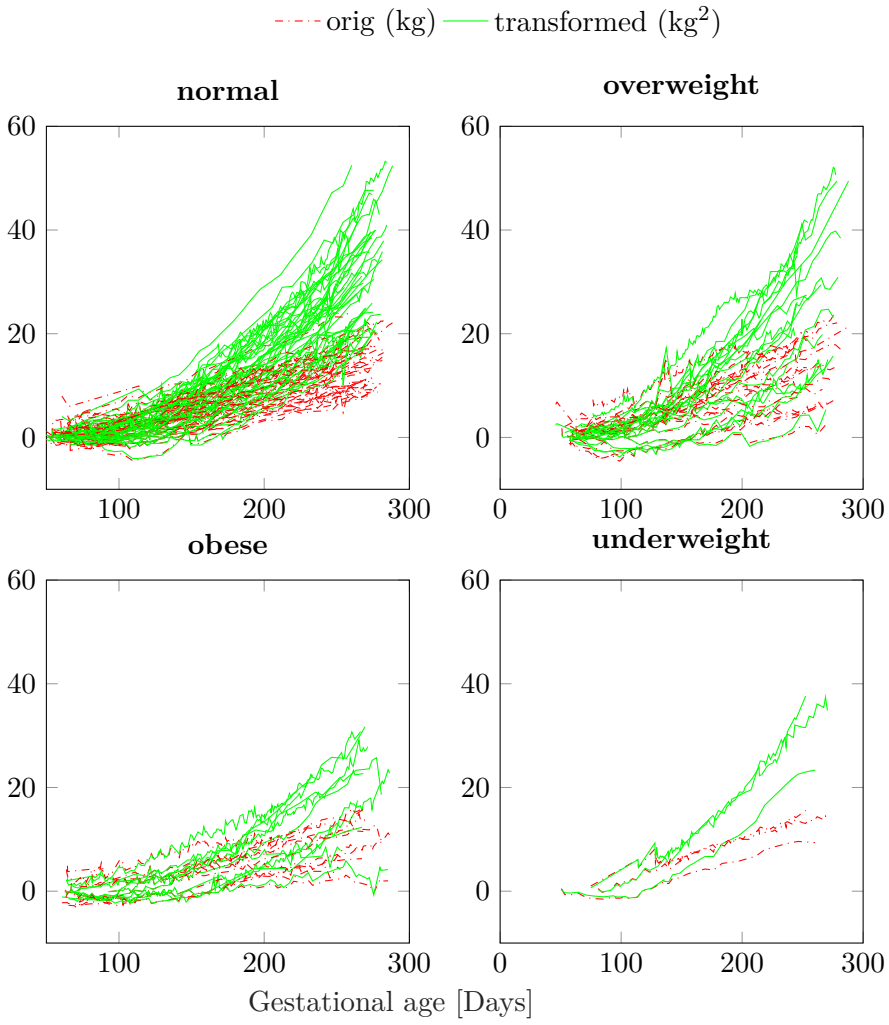


Figure 2.4: Pre-pregnancy based BMI class based transformation of subjects' weight gain from dataset $\mathcal{D}_E$.

subjects in the public-training data. This distribution over the MLE estimates of the coefficients, $p(\mathbf{w})$ is acquired from the $N-1$ subjects in the public-training data as an *a-priori* estimate. The likelihood learnt from the self-training data and the *a-priori* distribution learnt from the population data are then combined using bayes theorem to calculate the

maximum-a-posteriori (MAP) estimate of the coefficients $p(\mathbf{w}|\mathcal{D})$.

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, \frac{P(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{P(\mathcal{D})} \tag{2.8}$$

We can ignore $P(\mathcal{D})$ in eqn. (2.8) as it doesn't depend on $\mathbf{w}$. The forecast at time t_m^+ is given by $\hat{\mathbf{w}}_{MAP}[t_m^+ \, {t_m^+}^2 \cdots {t_m^+}^p]^T$.

2.4.3 Classification using guidelines

We further extend the prediction results for better interpretation by classifying the predicted weight gain into three classes, 'underweight', 'normal', and 'overweight' represented as integer values '-1', '0' and '1' respectively. For this purpose, we compare the predicted weight gain with the recommended weight-gain guidelines at the delivery day t_d to get the 3-class classification output. Following eq. (2.2) and (2.3), classification function $c(t^i, y^i(t^i))$ for i^{th} subject is defined as a function of time t^i and weight gain value $y^i(t^i)$:

$$c(t^i, y^i(t^i)) = \begin{cases} -1 & y^i(t^i) < L_{bmi^i}(t^i), \\ 0 & L_{bmi^i}(t^i) \leq y^i(t^i) < U_{bmi^i}(t^i), \\ 1 & U_{bmi^i}(t^i) \leq y^i(t^i) \end{cases} \tag{2.9}$$

2.5 Experiments

We experiment with 1^{st} to 5^{th} order to fit our weight-gain data. We empirically chose a third order polynomial as it obtains the minimum prediction error among all other orders in cross-validation. However, with transformation based pre-processing, we choose order 2 for modelling $y_{transformed}$ as the transformation itself adds to the non-linearity by order 1.

2.5.1 State-of-the-art

ARIMA. This is a time series forecasting approach [20] that exploits correlations in historical data. Forecasting using ARIMA methods requires uniformly spaced samples of the time series. We introduce uniformity in personal training data by linear interpolation between samples. We fit an ARIMA*(p,d,q)* model by i) enforcing equi-spaced sampling by linear interpolation, ii) performing a grid search over the hyper-parameters [24] to find an optimal autoregressive order, degree of differencing, and moving average order, iii) forecasting multi-steps ahead in time to find the end-of-pregnancy gestational weight gain using the optimised hyper-parameters over the training part (GWG data until day t_d).

LSTM. We evaluate LSTM based regression network with 200 hidden units by training them to minimise the mean absolute error using the 'adam' optimization method [25]. The hyperparameters search space was set as follows, epochs = $\{50, 100, 150, 200, 250\}$, learning rate = $\{0.0001, 0.0005, 0.001, 0.005\}$, batch size = $\{16, 32, 64, 128\}$.

MLE. We also tested a polynomial fitting approach following maximum likelihood estimation (MLE) with different order polynomials. Order 2 produces best results (among the orders 1 to 5).

Each method in the state-of-the-art is trained using either raw data or data processed using the pre-processing strategy given in the section 2.4.1. For brevity, the findings for only the best performing data are maintained in the following results. With transformed data, LSTM and MLE performed best, however ARIMA performed best when given input from raw data.

2.5.2 Evaluation Metric

The performance of regression was computed using Mean Absolute Error (MAE),

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y(t^i_{m_i}) - y_{pred}(t^i_{m_i})|$$

We use accuracy *acc* as the desired metric for evaluating classification performance defined using eq. (2.9) as

$$\begin{aligned} acc &= \frac{1}{N}\sum_{i=1}^{N} I\left(c\Big(t^i_{m_i}, y^i_{pred}(t^i_{m_i})\Big) = c\Big(t^i_{m_i}, y^i(t^i_{m_i})\Big)\right) \\ &= \frac{\#\text{correct predictions in recommended guidelines}}{\#\text{total subjects}} \end{aligned} \tag{2.10}$$

where I is the indicator function such that $I(A) = 1$, if event A occurs and 0 otherwise and $t^i_{m_i}$ is the delivery day for i^{th} subject. Accuracy *acc* at a time t_j is the accuracy (averaged over N users) calculated using eq. (2.10) when personal-training data for the i^{th} subject is considered to be available only until the day t_j. Next, we calculate the normalized area under the accuracy curve ($AuAC$) to evaluate the performance of a given approach with respect to the available training data between days T_0 to T_1 as

$$AuAC_{T_1-T_0} = \frac{\int_{T_0}^{T_1} acc(t)\ \mathrm{d}t}{\int_{T_0}^{T_1} 1\ \mathrm{d}t} \;=\; \frac{\int_{T_0}^{T_1} acc(t)\ \mathrm{d}t}{T_1 - T_0}$$

We omit T_0 from the notation $AuAC_{T_1-T_0}$ and use $AuAC_{T_1}$ to denote $AuAC$ until day T_1 for simplicity as $T_0 = 120$ is fixed in our analysis. This is because atleast one subject exists with no recorded weight gain measurement before day 120. Fig. 2.5 shows two exemplary curves A and B with B being better at early prediction than A, hence $AuAC^B_{160} > AuAC^A_{160}$.

2.6 Results

We evaluate the performance of the described approaches in terms of MAE and accuracy of the predicted weight gain (class) against the actual end-of-pregnancy weight gain (class). To validate the performance, we perform leave-one-subject-out cross validation, where training dataset in each iteration consists of public-training data (weight-gain from $N-1$ subjects) and personal-training data from the test subject as defined in section 2.4. We experiment by varying the amount of available personal-training

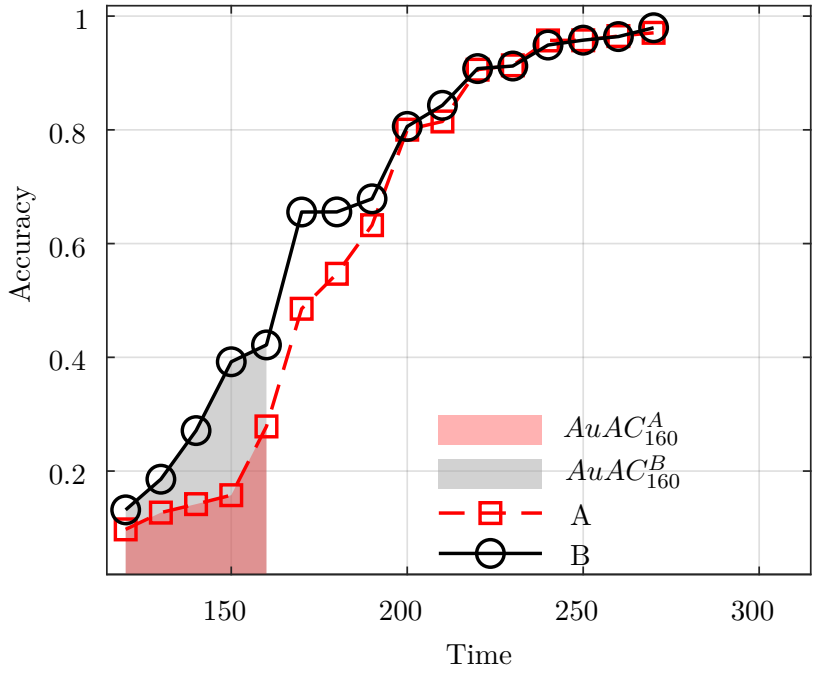


Figure 2.5: $AuAC$ for two exemplary accuracy curves A and B. The higher the accuracy with respect to time, the higher the $AuAC$.

data until a certain day in pregnancy and perform cross-validation to evaluate the performance of different approaches against training data availability. We also present performance measures for early prediction by taking day '140' as the early threshold as it is mid-way through the pregnancy. Finally, we study the effects of transferring model learnt from one geographic region to infer the data from subjects in another geographic region.

2.6.1 Weight gain trend visualisation

We predict the trend of weight gain on both the datasets $\mathcal{D}_E$ and $\mathcal{D}_C$ and present in Fig. 2.6 how such a prediction looks like with limited

training data. Fig. 2.6 shows the personal-training data up to 140 days into the pregnancy and the best and worst prediction results in terms of mean absolute error alongside the actual weight gain measurements during the later stages of pregnancy using the proposed approach with transformation. Since we are concerned about the end-of-pregnancy weight gain, we calculate the MAE right before the delivery date between actual and predicted weight gain while also show the predicted trend of weight gain for these subjects. The errors in prediction for the (best, worst) cases among the $\mathcal{D}_E$ and $\mathcal{D}_C$ are (0.93, 9.24) and (0.03, 11.42) kgs respectively. One can see that in Fig. 2.6(c) and (d), there is only single training observation before day 140. In Table 2.4 the confusion matrix for predicting different classes according to recommended guidelines on the both the datasets with training data until day 140. Also, Table 2.4(c) shows the confusion matrix based on model learnt from dataset $\mathcal{D}_E$ and tested on $\mathcal{D}_C$. Next, we perform leave-one-subject-out cross validation

Table 2.4: Confusion matrices for classification of end-of-pregnancy weight gain (underweight(u), normal(n) and overweight(o)) based on personal-training data up to only 140 days into the pregnancy using proposed method (P_T) in (a) LOOCV for dataset $\mathcal{D}_E$, (b) LOOCV for dataset $\mathcal{D}_C$ and (c) transferring model learn on dataset $\mathcal{D}_E$ to dataset $\mathcal{D}_C$.

		Pred		
		u	n	o
True	u	8	6	2
	n	4	15	6
	o	1	7	31

(a) $\mathcal{D}_E$

		Pred		
		u	n	o
True	u	1	22	0
	n	0	49	19
	o	1	23	38

(b) $\mathcal{D}_C$

		Pred		
		u	n	o
True	u	14	8	0
	n	7	48	10
	o	3	26	31

(c) $f_{\mathcal{D}_E} \rightarrow f_{\mathcal{D}_C}$

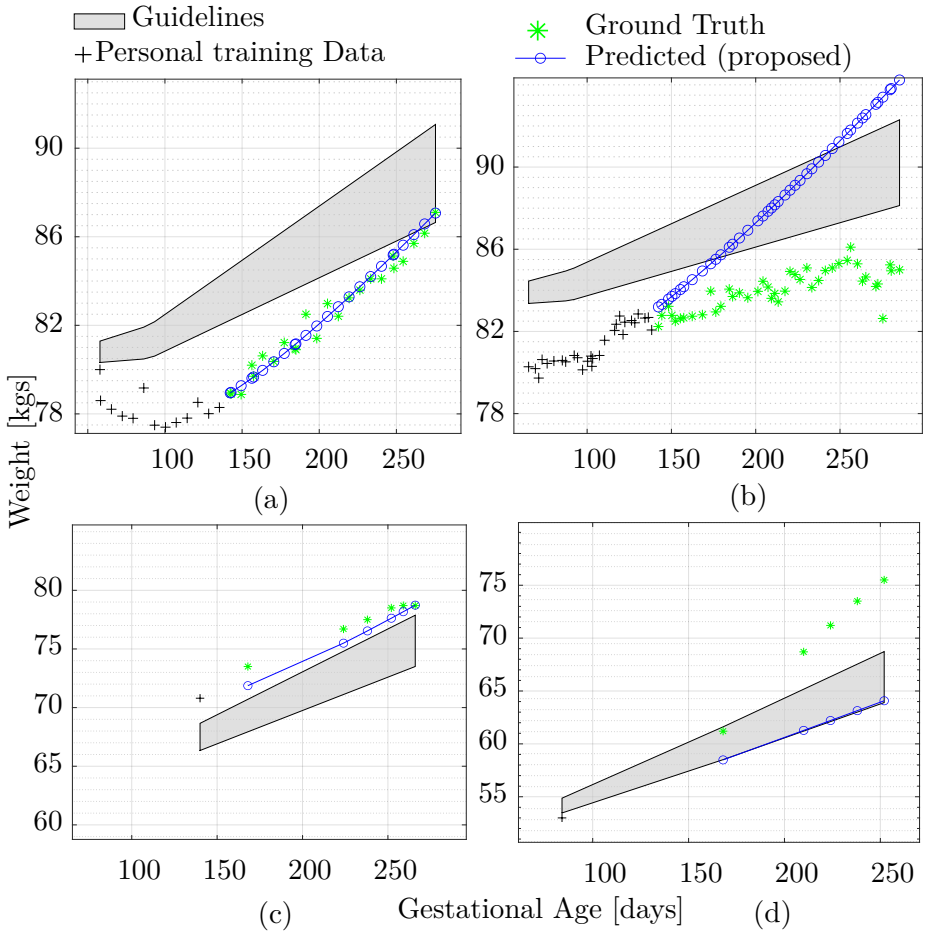


Figure 2.6: Proposed approach with transformation (P_T) to forecast weight gain with best (a), (c) and worst (b), (d) predictions with the actual weight gain data and recommended guidelines with number of training days = 140 on dataset $\mathcal{D}_E$ (a), (b) and $\mathcal{D}_C$ (c), (d).

(LOOCV) over all the subjects in each of the dataset by varying the availability of personal-training data before a given day in gestational age and calculate the performance averaged over all the subjects.

2.6.2 Comparison with State-of-the-art

To compare the performance of the proposed approach with the state-of-the-art methods, we study Mean absolute error (MAE) and accuracy (acc) against different amount of available personal-data. Fig. 2.7 shows that our proposed method outperforms the state-of-the-art approach in early detection (until day 160). All the improvements of the proposed method P_T are statistically significant based on a paired t-test with equal variances and $p < 0.05$ on both the datasets $\mathcal{D}_C$ and $\mathcal{D}_E$ compared to state-of-the-art.

Furthermore, ARIMA models' results are statistically insignificant as compared to proposed method for available training data from day 170 to 210 for both the datasets. Additionally, from Fig. 2.7, it can be observed that the MAE reduces, and accuracy increases with increasing availability of personal-data. Paired t-test with equal variances suggest that these improvements are statistically significant only for dataset $\mathcal{D}_E$ when sufficient training data is available (day 190 onwards) and is never statistically significant for $\mathcal{D}_C$.

Next, in addition to accuracy we try to quantify the performance of all the approaches against different availability of training data using a single metric by calculating $AuAC$ between day 120 to day 140. These values for different methods are presented in Table 2.5. Also, the accuracy score with training data until day '140' reported in Table 2.5 suggests an improvement of around 25.9% and 31.1% over the best of state-of-the-art for datasets $\mathcal{D}_E$ and $\mathcal{D}_C$ respectively.

Table 2.5: $MAE(t_{140})^{\ddagger}$, $AuAC^{\dagger}_{140}$ and $acc(t_{140})^{\dagger}$ for proposed technique v/s state-of-the-art (Best values in **bold**, ‡Lower is better, † Higher is better).

Method	**Proposed**				**State-of-the-art**					
	$\mathbf{P_T}$		**P**		**ARIMA**		**LSTM**		**MLE**	
Dataset	$\mathcal{D}_E$	$\mathcal{D}_C$	$\mathcal{D}_E$	$\mathcal{D}_C$	$\mathcal{D}_E$	$\mathcal{D}_C$	$\mathcal{D}_E$	$\mathcal{D}_C$	$\mathcal{D}_E$	$\mathcal{D}_C$
$MAE(t_{140})^{\ddagger}$	**2.45**	2.60	2.82	**2.57**	16.22	6.60	12.10	16.01	8.17	54.76
$AuAC^{\dagger}_{140}$	**0.65**	0.53	0.59	**0.56**	0.43	0.33	0.43	0.22	0.51	0.32
$acc(t_{140})^{\dagger}$	**0.68**	**0.59**	0.61	0.46	0.51	0.45	0.43	0.35	0.54	0.41

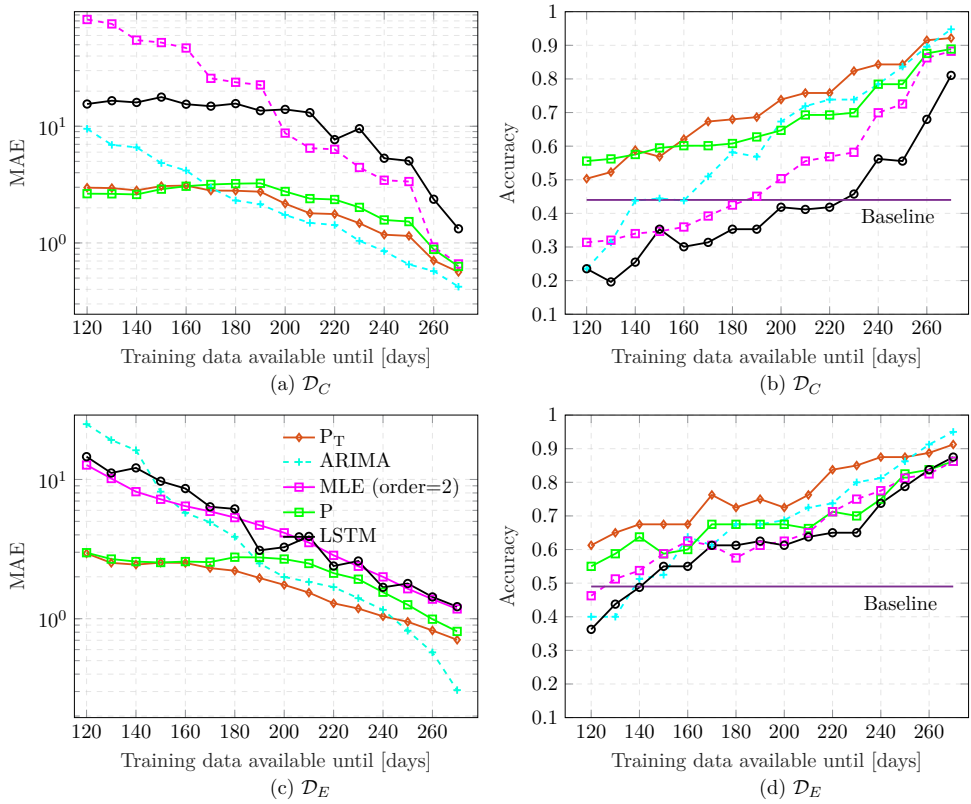


Figure 2.7: Performance scores (mean absolute error and accuracy) for the proposed approach with respect to state-of-the-art on $\mathcal{D}_E$ and $\mathcal{D}_C$. A single (abscissa, ordinate) pair in the figure represent the performance score (ordinate) averaged over all the subjects with respect to availability of training data until a certain day (abscissa). MAE reduces (a,c) and accuracy increases (b,d) as availability of training data increases. Majority label percentage in respective datasets is taken as the accuracy baseline.

2.6.3 Effect of model transfer between datasets

We test the proposed approach in two settings to test the model transfer as follows, i) we train the MAP model on $\mathcal{D}_E$ and test the model learnt

on $\mathcal{D}_C$, ii) we perform leave-one-out cross validation (LOOCV) on $\mathcal{D}_C$ where no subjects from $\mathcal{D}_E$ were taken into account. Fig. 2.8 shows the comparison of model transfer with or without the transformation based processing step.

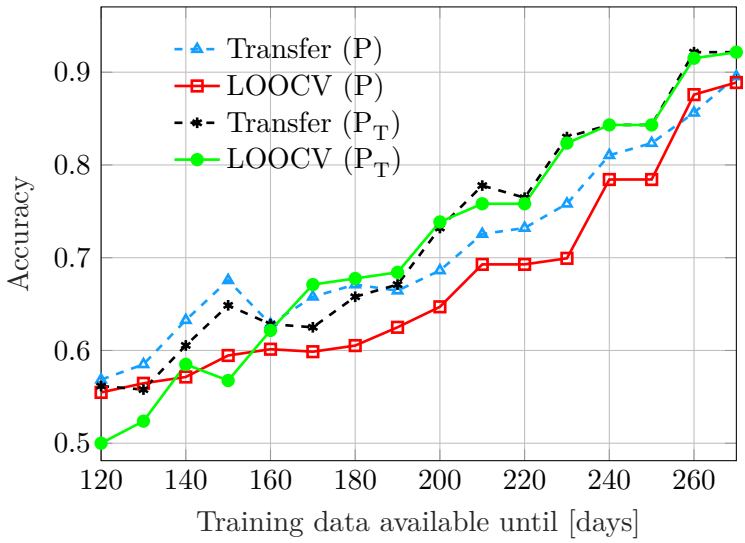


Figure 2.8: In early prediction, accuracy assessed on $\mathcal{D}_C$ with model transfer from $\mathcal{D}_E$ is superior to accuracy with LOOCV (with only $\mathcal{D}_C$).

It can be observed in Fig. 2.8 that accuracy of model transfer based on P_T is greater than LOOCV until day 160 i.e in early prediction. However, accuracy of the proposed MAP approach without the transformation is almost always better with model transfer.

2.7 Discussion

Predicting weight gain reliably in pregnant women as early as possible is at the heart of this study. In this study, we experiment by first collecting weight-gain datasets in two different geographies and building prediction models that utilise prior information generated from public-training dataset to tune the personal-model for accurate estimation of the end-of-

pregnancy weight gain. The total percentage of the most represented class post-pregnancy is set as a baseline for comparing prediction accuracy. According to Table 2.3, this baseline is 0.49 for $\mathcal{D}_E$ and 0.44 for $\mathcal{D}_C$ marked in Fig. 2.7.

With limited amount of available personal-training data for prediction of weight gain, our MAP based Bayesian approach forms an a-priori estimate of model coefficients based on public-training data model coefficients. This addition of prior in the model also acts as a type of regularization. This results in high performance gains in *early* prediction of around 25.9% and 31.1% over the best of state-of-the-art for datasets $\mathcal{D}_E$ and $\mathcal{D}_C$ respectively. Additionally, including the transformation based processing step improves the performance further (Table 2.5). This is because our transformation step introduces a non-linearity in time based on pre-pregnancy BMI that scales each subject's raw weight gain data with respect to the allowed rate of weight gain thus scaling each time series to similar range. Also, the polynomial fit for transformed time-series is done with one lower order ($p = 2$) than the ordinary MAP fit ($p = 3$) which improves the generalization ability of the fit. It is evident from Fig. 2.6(b), the worst result occurs when the person's weight gain trend is different from any of the available subjects in public-training data and the personal-training data (until day 140) is also insufficient to capture this trend. We think that there are two ways in which this can be addressed 1) increasing the amount of personal-training data and/or 2) increasing the size of public-training data by adding more subjects that reduces the variance of the model. Fig. 2.6(c) and (d) show the best and worst result on dataset $\mathcal{D}_C$. It can be observed that in both the cases only a single personal-training observation is present before day 140 with it being present close to test data in time in the best case (Fig. 2.6(c)) and being further away in time to the forecast horizon in the worst case (Fig. 2.6(d)). One can infer that the points close in time to the forecast horizon have more importance in reliable prediction than the ones farther away in time.

Table 2.4(a) and (b) suggest that most of the prediction errors are to the neighbouring classes. The accuracy is lowest for the underweight class as it is the most under-represented class (Table 2.3) in our dataset. Fig. 2.7a,c and Fig. 2.7b,d show the mean absolute error in prediction

and the accuracy for different datasets averaged over all the subjects. Fig. 2.7 shows that the prediction error reduces and accuracy improves as the personal-training data availability increases.

Although at a glance at Fig. 2.7(a), it might look like ARIMA's MAE for dataset $\mathcal{D}_C$ is less than proposed P_T when training data is available as early as day 170. However, as described in subsection 2.6.2 the low mean absolute error is statistically insignificant as compared to P_T until day 210. As more personal-training data is available by day 210 for dataset $\mathcal{D}_C$ and by day 240 for dataset $\mathcal{D}_E$, the personal models based on ARIMA tend to become more accurate than the proposed approach. Although, this could be of importance in problems with low forecast-horizon, but in cases where early forecast is needed such as ours, proposed approach outperforms ARIMA.

We can observe from Fig. 2.7 that for such a smaller dataset with non-uniformly sampled time series, even ARIMA performs better than LSTM. LSTM based deep learning approaches perform better when huge amount of data is available and enough training data is present. In our case, this availability of personal-training data is not present because of two reasons i) the data is sampled irregularly and has a very low sampling frequency and ii) early intervention requires using as little personal-training data as possible. We believe that when more subjects participate, our approach will scale better than LSTM based approach because of the aforementioned reasons.

Remark that the methods MLE, ARIMA, and LSTM provided as state-of-the-arts train a model using only the test subject's limited personal data, as these algorithms can handle the time series data from a single participant at a time in a personalised manner. A multivariate treatment of MLE or LSTM might involve using data from all the subjects and then training a model that uses information of all the subjects at a different time instants. However, in the case of non-uniformly sampled data the training data might appear noisy as not all the subjects' information is available at a given time instant and the model created using the training data is a very general one that cannot be personalised using the test subjects' available data.

The proposed approach (MAP) is an extension of the MLE method that

utilises other subjects' data along with personal data to improve the performance.

Fig. 2.8 shows that model 'Transfer' works better than 'LOOCV' irrespective of pre-processing. Table 2.4(c) shows that there is a huge improvement in predicting the class "underweight" with this model transfer without compromising the performance of other classes. The dataset $\mathcal{D}_E$ exhibits more variability in terms of capturing weight-gain trend among different BMI classes with pre-pregnancy BMI ranging from 20 to 28 kg/m^2. This might be one of the causes that model trained on this dataset generalizes well on $\mathcal{D}_C$.

Our proposed Bayesian approach with pre-processing has a prediction MAE of only 2.45 kgs ($\mathcal{D}_E$) and 2.82 kgs ($\mathcal{D}_C$) and a classification accuracy of 67.5% ($\mathcal{D}_E$) and 58.9% ($\mathcal{D}_C$) at day '140'(mid way through the pregnancy) for early intervention as compared to state-of-the-art approaches, best of which has an MAE of 8.17 ($\mathcal{D}_E$) and 6.60 kgs ($\mathcal{D}_C$) and an accuracy up to 53.8% ($\mathcal{D}_E$) and 44.8% ($\mathcal{D}_C$). Fig. 2.7 shows that our approach predicts better than the state-of-the-art when training from data using 120-240 days, and predicts close to state-of-the-art during the very last few days of the pregnancy. $AuAC_{140}$ can be thought of as an early intervention score that measures how accurate the classification performance is with varying amount of training data from day 120 until day 140. In other words, the early prediction performance of our technique with transformation has an $AuAC_{140}$ of 0.65 ($\mathcal{D}_E$) and 0.53 ($\mathcal{D}_C$). Another key step in this work was to apply model transfer to test the generalisation capability of the model between two different geographic regions that further improves the prediction capability on the sparser dataset $\mathcal{D}_C$.

2.8 Conclusion

In this study, we propose an efficient early-weight gain prediction system in pregnant women. We validate and show the efficacy of our proposed approach over this unique dataset from two diverse geographical regions. Our approach utilises the power of combining a-priori information learnt from the public-training data and tunes the

parameters of personal training data based on this prior information. Additionally, we incorporate a pre-processing step to scale our data using meta-data such as pre-pregnancy weight and BMI to achieve additional boost in our performance. Our results show the reliable estimation of end-of-pregnancy weight gain that can help to provide proper interventions by pre-natal care providers and to reduce risks of adverse maternal and neonatal effects of excessive or inadequate GWG.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

Bibliography

[1] C. Puri, G. Kooijman, F. Masculo, S. V. Sambeek, S. D. Boer, J. Hua, N. Huang, H. Ma, Y. Jin, F. Ling, G. Li, D. Zhang, X. Wang, S. Luca, and B. Vanrumste, "A personalized bayesian approach for early intervention in gestational weight gain management toward pregnancy care", *IEEE Access*, vol. 9, pp. 160 946–160 957, 2021.

[2] R. F. Goldstein, S. K. Abell, S. Ranasinha, M. Misso, J. A. Boyle, M. H. Black, N. Li, G. Hu, F. Corrado, L. Rode, *et al.*, "Association of gestational weight gain with maternal and infant outcomes: A systematic review and meta-analysis", *Jama*, vol. 317, no. 21, pp. 2207–2225, 2017.

[3] K. M. Rasmussen, P. M. Catalano, and A. L. Yaktine, "New guidelines for weight gain during pregnancy: What obstetrician/gynecologists should know", *Current opinion in obstetrics & gynecology*, vol. 21, no. 6, p. 521, 2009.

[4] R. Gaillard, B. Durmuş, A. Hofman, J. P. Mackenbach, E. A. Steegers, and V. W. Jaddoe, "Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy", *Obesity*, vol. 21, no. 5, pp. 1046–1055, 2013.

[5] L. A. Gilmore, M. Klempel-Donchenko, and L. M. Redman, "Pregnancy as a window to future health: Excessive gestational weight gain and obesity", in *Seminars in Perinatology*, Elsevier, vol. 39, 2015, pp. 296–303.

[6] E. Oken, S. L. Rifas-Shiman, A. E. Field, A. L. Frazier, and M. W. Gillman, "Maternal gestational weight gain and offspring

weight in adolescence", *Obstetrics and gynecology*, vol. 112, no. 5, p. 999, 2008.

[7] S. B. Sridhar, J. Darbinian, S. F. Ehrlich, M. A. Markman, E. P. Gunderson, A. Ferrara, and M. M. Hedderson, "Maternal gestational weight gain and offspring risk for childhood overweight or obesity", *American journal of obstetrics and gynecology*, vol. 211, no. 3, 259–e1, 2014.

[8] A. Fraser, K. Tilling, C. Macdonald-Wallis, N. Sattar, M.-J. Brion, L. Benfield, A. Ness, J. Deanfield, A. Hingorani, S. M. Nelson, *et al.*, "Association of maternal weight gain in pregnancy with offspring obesity and metabolic and vascular traits in childhood", *Circulation*, vol. 121, no. 23, p. 2557, 2010.

[9] R. R. Davis and S. L. Hofferth, "The association between inadequate gestational weight gain and infant mortality among us infants born in 2002", *Maternal and child health journal*, vol. 16, no. 1, pp. 119–124, 2012.

[10] B. Abrams, S. Carmichael, and S. Selvin, "Factors associated with the pattern of maternal weight gain during pregnancy", *Obstetrics & Gynecology*, vol. 86, no. 2, pp. 170–176, 1995.

[11] E. Althuizen, M. N. van Poppel, J. C. Seidell, and W. van Mechelen, "Correlates of absolute and excessive weight gain during pregnancy", *Journal of Women's Health*, vol. 18, no. 10, pp. 1559–1566, 2009.

[12] I. Streuling, A. Beyerlein, E. Rosenfeld, H. Hofmann, T. Schulz, and R. Von Kries, "Physical activity and gestational weight gain: A meta-analysis of intervention trials", *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 118, no. 3, pp. 278–284, 2011.

[13] A. Hui, L. Back, and e. a. Ludwig, "Lifestyle intervention on diet and exercise reduced excessive gestational weight gain in pregnant women under a randomised controlled trial", *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 119, no. 1, pp. 70–77, 2012.

[14] P. Brawarsky, N. Stotland, R. Jackson, E. Fuentes-Afflick, G. Escobar, N. Rubashkin, and J. Haas, "Pre-pregnancy and pregnancy-related factors and the risk of excessive or inadequate gestational weight gain", *International Journal of Gynecology & Obstetrics*, vol. 91, no. 2, pp. 125–131, 2005.

[15] L. M. Redman, L. A. Gilmore, J. Breaux, D. M. Thomas, K. Elkind-Hirsch, T. Stewart, D. S. Hsia, J. Burton, J. W. Apolzan, L. E. Cain, *et al.*, "Effectiveness of smartmoms, a novel ehealth intervention for management of gestational weight gain: Randomized controlled pilot trial", *JMIR mHealth and uHealth*, vol. 5, no. 9, e133, 2017.

[16] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, ACM, 2019, pp. 398–399.

[17] E. Y. Lau, J. Liu, E. Archer, S. M. McDonald, and J. Liu, "Maternal weight gain in pregnancy and risk of obesity among offspring: A systematic review", *Journal of obesity*, vol. 2014, 2014.

[18] D. M. Thomas, J. E. Navarro-Barrientos, D. E. Rivera, S. B. Heymsfield, C. Bredlau, L. M. Redman, C. K. Martin, S. A. Lederman, L. M Collins, and N. F. Butte, "Dynamic energy-balance model predicting gestational weight gain", *The American journal of clinical nutrition*, vol. 95, no. 1, pp. 115–122, 2012.

[19] P. Guo, D. E. Rivera, J. S. Savage, E. E. Hohman, A. M. Pauley, K. S. Leonard, and D. S. Downs, "System identification approaches for energy intake estimation: Enhancing interventions for managing gestational weight gain", *IEEE Transactions on Control Systems Technology*, 2018.

[20] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[21] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2017.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278.

[24] R. Shibata, "Selection of the order of an autoregressive model by akaike's information criterion", *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

Chapter 3

Modelling Time Series Through Informative Subset Selection

This chapter is published as:
C. Puri, G. Kooijman, B. Vanrumste, and S. Luca, "Forecasting time series in healthcare with gaussian processes and dynamic time warping based subset selection", *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6126–6137, 2022. DOI: 10.1109/JBHI.2022.3214343

The chapter is also inspired by:
C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278

Abstract

Modelling real-world time series can be challenging in the absence of sufficient data. Limited data in healthcare, can arise for several reasons, namely when the number of subjects is insufficient or the observed time series is irregularly sampled at a very low sampling frequency. This is especially true when attempting to develop personalised models, as there are typically few data points available for training from an individual subject. Furthermore, the need for early prediction (as is often the case in healthcare applications) amplifies the problem of limited availability of data. This article proposes a novel personalised technique that can be learned in the absence of sufficient data for early prediction in time series. Our novelty lies in the development of a subset selection approach to select time series that share temporal similarities with the time series of interest, commonly known as the test time series. Then, a Gaussian processes-based model is learned using the existing test data and the chosen subset to produce personalised predictions for the test subject. We will conduct experiments with univariate and multivariate data from real-world healthcare applications to show that our strategy outperforms the state-of-the-art by around 20%.

3.1 Introduction

Time series forecasting is an extensive field of research for diverse applications with possibilities in economics, physical or environmental sciences, or healthcare. Traditional treatment of time series includes multiplicative methods such as the auto-regressive integrated moving average model (ARIMA) and its multivariate treatment or state-space models such as the Kalman filter and generalised autoregressive conditional heteroskedasticity (GARCH) process that are additive [3]. These methods are well suited for modelling time series when the data are uniformly sampled. However, as the number of time series in a dataset increases, these methods do not scale well because each time series must be trained individually. Moreover, it is difficult to model the shared temporal patterns across various time series in the whole dataset during training and forecasting.

Modelling real-world healthcare related time series for forecasting is often difficult owing to the limited availability of data due to practical constraints. For example, if a study is conducted only with a small number of participants, then the dataset might not always be a complete representation of a given task. However, limited subjects alone might not be the only issue. For instance, accelerometer-based time series data gathered at high frequency (on the order of 25 Hz) to classify human activity recognition. Even if the label information is present every 5 seconds in a recording of 5 minutes from a single subject, there are 125 points in time to model each label and 60 such instances can be acquired from a single recording. As a result, even with a small number of participants, it is possible to develop generalizable, high-performing models [4]. If individual time-series are sampled at very low sampling rate from a small number of subjects, the modelling becomes difficult, e.g. modelling daily weight gain over a period of pregnancy. The problem of limited subjects and low sampling frequency is further aggravated when the observed time series, univariate or multivariate, are sporadic in nature, i.e., they are noisy and contain missing values. Few examples include sensor failure, data artifacts in climate time series, or in healthcare use-cases. For example, a patient can skip regular health check-up appointments for intentional or unintentional reasons resulting in multiple missing entries in the electronic health record (EHR) [5]. Furthermore, the individual forecasts must be performed as quickly as possible so that timely interventions can be implemented. This further restricts the availability of the personal data required to learn individual patterns.

Modern deep learning techniques have gained traction in time series forecasting because they can utilise multiple time series from the training data to discover non-linear temporal patterns [6]. However, deep learning models expect huge amounts of training data to learn these patterns [7], [8]. Additionally, state-of-the-art deep learning models for time series forecasting still suffer when the training data is sporadic in nature and multi-step forecasting is difficult in the presence of insufficient time series data [9], [10].

This work will develop methods for forecasting time series data in healthcare applications where for each participant time series data is

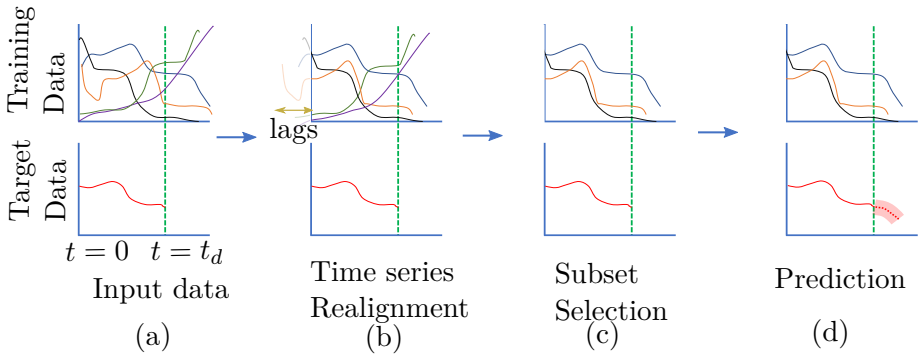


Figure 3.1: An example to illustrate our SS-GP approach. (a) The training and target time series that are considered (the green dotted line shows that target data is only available until time t_d^+). (b) The training data that are aligned in time with the target time series. (c) A subset of the training data that share similar temporal characteristics with the target data (the purple and the dark green curves are therefore discarded). (d) The training data and the available target data are used to predict a sequence of future values in the target time series (red dotted line).

available. Given a target time series (of a test subject) of non-uniformly sampled instances, the main aim is to predict future values over a period of time. In particular, we will treat the following challenges which are often encountered in healthcare data: (a) time series forecasting when for each subject the data are non-uniformly sampled and can have a very low sampling frequency and (b) the estimation of individualised models while very little individual data are available. Note that these data-related difficulties are even more challenging when at the same time the number of subjects is low.

The solution proposed in this paper consists of a subset selection (SS) approach to select time series from the training data (of other subjects) that share temporal similarities with the target time series. This subset of time series is then used to train a non-parametric Gaussian process (GP) in a Bayesian way [11]. Modelling unevenly sampled time series with Gaussian process-based techniques eliminates the need to impute data to make them uniformly sampled. We will show that this approach (further referred to as SS-GP) can improve the target series' forecasting

performance, especially when the time series in the selected subset are aligned in time with the target time series.

Fig. 3.1 showcases an example: first, the training data are aligned with the target time series; second, a subset of time series from the training data is selected that share similar temporal characteristics with the target time series. The subset is then used to train a GP for multi-step ahead prediction, i.e., for predicting a sequence of future values in the target time series.

We experiment with two real-life time-series datasets from healthcare to prove the efficacy of the proposed solution in multi-step time series forecasting. We further demonstrate the implications of limited individual data on training by varying the availability of data in time and assessing the prediction error. We empirically show that our approach not only reliably predicts in the case of missing observations but also accurately predicts multiple steps ahead in time in the case of limited personal data.

The main contributions of this paper are:

- We propose a new multi-step time series prediction approach that can handle time series with non-uniformly sampled time series data in limited datasets.
- We design a time series realignment technique that tackles time series in a training set that were initiated at different times. In other words, when time t_0 of the training time series is different, realigning them with respect to each other prior to modelling leads to a more exact pattern match and a more precise forecast.
- We suggest dynamic subset selection, which takes advantage of shared temporal patterns to dynamically select a smaller subset of time-series from the training data.
- Finally, we empirically show that the SS-GP approach outperforms state-of-the-art approaches on two real-world healthcare datasets where there is a need to predict early and where missing data are inevitable.

3.2 Related Work

Time series literature consists of widespread approaches for forecasting ranging from classical works from the 1960s to contemporary works [12], [13]. Classical works like state-space or autoregressive approaches such as ARIMA for univariate and VARIMA for multivariate approaches exist that predict the individual observations in time series [14]. Much of these approaches are applied in an auto-regressive manner where one step predictions are achieved by applying the learned model recursively. This tends to achieve significant errors in prediction if the forecast horizon is large. Currently, deep learning-based methods such as recurrent neural networks (RNNs) are popular due to their automatic feature extraction abilities in sequence modelling. Improved variants of RNNs that alleviate vanishing gradient problems such as long-short term memory networks [6] and gated recurrent units (GRU) [15] are capable of capturing long term dependency with uniformly sampled sequence data. Authors in [16] create a mask where the data is missing and use this mask along with available data as input thus utilising missingness in data as informative features to train RNNs and cope with missingness in the data.

Multiple approaches in deep learning have focused on time series classification and regression in healthcare ranging from ECG classification [17] to glucose forecasting [18]. Authors in [16], [19] have presented works that are able to diagnose a condition, such as sepsis in an intensive care unit environment, by learning from multivariate clinical data using resources such as electronic health records (EHRs). The majority of these methods that can manage missing data have been trained on a significant amount of data, providing them an advantage. However, when insufficient training data is available, traditional machine learning strategies outperform deep learning strategies [20]. There have not been any systematic work that handles limited data availability. We attempt to address such deficiency in training data that stems from either (a) the irregularly sampled time series, or (b) the limited number of samples of an individual time series resulting from the necessity to predict as soon as possible.

Gaussian processes (GPs) provide a framework to model time series in the presence of such irregularly sampled instances and can quantify the

uncertainty of predictions. For example, GP models are used in clinical time series classification and imputation [21].

This work proposes a personalised approach for multi-step time series forecasting that can handle non-uniformly sampled time series through Bayesian learning.

3.3 Notation

Let us assume, N subjects are studied and the *training data* consists of time series data of K predictor variables denoted by x at time t for each subject $1 \leq j \leq N$:

$$x_1^j(t), ..., x_K^j(t).$$

Our goal is to make predictions about a response variable $y^j(t)$ based on such feature data. For each subject however the time series are sampled at i different times, t_i^j, such that,

$$t_1^j < t_2^j < \cdots < t_{m^j}^j,$$

where m_j denotes the number of measurements of the feature $x_k^j (1 \leq k \leq K)$ that are available for the j^{th} subject. Remark that, for a given subject j, all predictor variables are measured at the same time instances.

In what follows, the feature data is denoted in matrix notation:

$$\mathbf{X}^j = [x_k^j(t_i^j)]_{ik}$$

denoting a $m^j \times K$ matrix of which the k^{th} column contains the data of the k^{th} feature of the j^{th} subject over all the time instances.

The measurements of the responses of a subject are collected in a vector:

$$\mathbf{y}^j = [y^j(t_1^j) \ \ldots \ y^j(t_{m^j}^j)].$$

Note that the response variable for subject j is sampled at the same time instances as the predictor variables of subject j.

There are two ways that data in time series might go missing:

- *missing observations within a time-series* : time series in the j^{th} instance of training or target data might not be evenly spaced, i.e., $t^j_{(i+1)} - t^j_i \neq t^j_{(i+2)} - t^j_{(i+1)}$, $\forall i \in \{1, 2, \cdots, m^j - 2\}$.
- *missing observations in different time instants within all time-series* : Time series data of predictor variables are not sampled at the same times across different subjects, i.e., t^j_i *is not necessarily equal to* $t^{j'}_i$, $\forall j, j' \in \{1, \cdots, N\}, i \in \{1, \cdots, m^j\}$

Suppose, we are interested in predictions for a *target subject* (indexed with '+') based on the measurements of the predictor variables $\mathbf{X}^+$ and the measurements of the response variable available up to some time t^+_d:

$$\mathbf{y}^+ = [y^+(t^+_1)\, y^+(t^+_2) \cdots y^+(t^+_d)],$$

where we assume that $t^+_d << t^j_{m^j}$, $\forall j \in \{1, \ldots, N\}$ i.e., the available temporal information for a target subject is limited compared to the number of time instances that are available for training for other subjects primarily due to the need for early prediction. The objective is to try to learn a function f, such that, the future response value at h^{th} time-step can be predicted as,

$$y^+_{(d+h)} = f(\underbrace{\mathbf{X}^j, \mathbf{y}^j}_{\text{training data}}, \underbrace{\mathbf{X}^+, \mathbf{y}^+}_{\text{target data}}) + \epsilon_h, \tag{3.1}$$

where

$$\epsilon_h \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

is independent and identically distributed (i.i.d) gaussian.

There are two multi-step forecasting strategies, direct vs iterative. Note that we use a direct multi-step prediction strategy where the responses at $t^+_{d+1}, \cdots, t^+_{d+h}$ time steps are predicted using only the available data until time t^+_d. However, an iterative multi-step forecasting technique predicts only the next time occurrence at t^+_{d+1} at a time. Multi-step predictions then can be made by including the previously predicted value $(y^+(t^+_{d+1}))$ of the response variable in the training data to predict the response at the next time instance and so on until h^{th} time-step is predicted [22].

3.4 State-of-the-art

In this section, we provide a brief overview of the existing techniques.

3.4.1 Subset Selection

Despite its simplicity, the k-nearest neighbours technique remains the benchmark for the classification of univariate time series [23]. In the case of multivariate time series, we employ k-means based clustering to create k profiles among the given dataset of time series grouping them by similar patterns. We further discuss the implementation details in section 3.6.2.

3.4.2 Time series forecasting

Maximum Likelihood Estimation (MLE): A p^{th}-order polynomial can be estimated with coefficients $\beta = [\beta_0\, \beta_1\, \cdots\, \beta_p]^T$ such that $y^+(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p$. The training can be done by maximizing the likelihood over the available responses $\mathbf{y}^+ = [y^+(t_1^+)\, y^+(t_2^+)\, \cdots, y^+(t_d^+)]$, $\ell(\mathbf{w}) = P(\mathbf{y}^+|\beta)$,

$$\hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}}\, P(\mathbf{y}^+|\beta) = \prod_{i=1}^{d} p(y^+(t_i^+)|t_i^+;\beta). \tag{3.2}$$

Eq. (4.6) is the model created using only a few observations from the target data up to the time t_d^+ days. This method results in personalised models and predictions, but the limited availability of data can hamper inference. This article will show how to properly use data from other subjects to address this issue.

Maximum-a-posteriori estimation (MAP) [2]: The maximum likelihood estimate of $\hat{\beta}$ may be found using the available training data (of other subjects). As an a-priori estimate, the distribution of these coefficient estimates, $p(\beta)$, obtained from the N participants in the training data may be used. The maximum-a-posteriori estimate of the coefficients, $p(\beta|\mathbf{y}^+)$ is calculated by combining the likelihood learned

from the target data with the *prior* distribution learned from the training data using Bayes theorem:

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}}\, p(\beta|\mathbf{y}^+) = \frac{P(\mathbf{y}^+|\beta)p(\beta)}{P(\mathbf{y}^+)}. \tag{3.3}$$

At time t_m^+, the prediction is given by $\hat{\beta}_{MAP}[t_m^+\, t_m^{+2} \cdots, t_m^{+p}]^T$. In both MLE and MAP, the parameter p is selected based on the application of interest, which should be known in advance.

ARIMA: is a method for forecasting time series data based on correlations in historical data [14]. Time series samples must be consistently spaced when utilising ARIMA algorithms for forecasting. Personal training data can be made uniform using linear interpolation between samples. For a uniformly sampled target time series response variable, an ARIMA model of order (p, d, q) capable of modelling $\mathbf{y}_+ = [y^+(t_1^+)\, y^+(t_2^+) \cdots y^+(t_d^+)]$ is defined by the equation:

$$\phi(B)(1-B)^d y^+(t) = \theta(B)w(t), \tag{3.4}$$

where $y^+(t)$ and $w(t)$ represent time series and random error at time t respectively. B is a backward shift operator defined by $By^+(t) = y^+(t-1)$, d is the order of differencing. $\phi(B)$and $\theta(B)$ are autoregressive (AR) and moving averages (MA) operators of orders p and q, respectively, and are defined as,

$$\begin{aligned} \phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^q, \end{aligned} \tag{3.5}$$

where $\phi_1, \phi_2, ..., \phi_p$ are the autoregressive coefficients and $\theta_1, \theta_2, ...\theta_q$ are the moving average coefficients.

LSTM. Long Short-Term Memory (LSTM) networks are a particular case of Recurrent Neural Networks (RNN) with the ability to model temporal dependencies from the past and have shown outstanding prediction performance [6]. This is done by using *forget, memory* and *output gate* that control the flow of the data during learning. This makes it easier to decide whether the data in each LSTM cell should be discarded, filtered, or added to the next cell [6].

Gaussian Processes. The Gaussian Processes (GP) are non-parametric models appropriate for sparsely available data. GP is a collection of

random variables, such that the joint distribution of every finite set of them is Gaussian (multivariate) [11]. We are given a training data $\mathbf{Xs}$ for N subjects:[1]

$$\mathbf{Xs} = \begin{bmatrix} \mathbf{X}^1 \\ \mathbf{X}^2 \\ \vdots \\ \mathbf{X}^N \end{bmatrix} = \begin{bmatrix} x_1^1(t_1^1) & x_2^1(t_1^1) & \cdots & x_K^1(t_1^1) \\ x_1^1(t_2^1) & x_2^1(t_2^1) & \cdots & x_K^1(t_2^1) \\ \vdots & \vdots & \ddots & \\ x_1^1(t_{m^1}^1) & x_2^1(t_{m^1}^1) & \cdots & x_K^1(t_{m^1}^1) \\ x_1^2(t_1^2) & x_2^2(t_1^2) & \cdots & x_K^2(t_1^2) \\ \vdots & \vdots & \ddots & \\ x_1^2(t_{m^2}^2) & x_2^2(t_{m^2}^2) & \cdots & x_K^2(t_{m^2}^2) \\ \vdots & \vdots & \ddots & \\ x_1^N(t_{m^N}^N) & x_2^N(t_{m^N}^N) & \cdots & x_K^N(t_{m^N}^N) \end{bmatrix}, \tag{3.6}$$

and $\mathbf{ys} = [\mathbf{y}^1\, \mathbf{y}^2 \cdots \mathbf{y}^N]^\top$. f is defined from eq. (6.1) as $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with mean and covariance functions $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ respectively. The covariance function encodes all the assumptions of the data such that two independent observations closer to each other have similar outputs. This nearness is used to model the structure of the multivariate time series, given that the covariance remains positive semi-definite [11]. We chose a squared exponential covariance function based on the assumption that the data have independent and identically distributed gaussian noise with variance σ_n^2,

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2\, exp\left(-\frac{1}{2l^2}|\mathbf{x} - \mathbf{x}'|^2\right) \tag{3.7}$$

Given $\mathbf{ys} = [y^1(t_1^1) \cdots, y^1(t_{m^1}^1) \cdots y^N(t_1^N) \cdots y^N(t_{m^N}^N)]^\top$ and $\mathbf{K}$ as a matrix $K_{ab} = k(\mathbf{x}_a, \mathbf{x}_b)$, $\forall \mathbf{x}_a, \mathbf{x}_b \in \mathbf{Xs}$ using eqn. (6.5), and following the optimisation procedure from [11], the hyperparameters $\{\sigma_f,\, l,\, \sigma_n\}$ are estimated by maximising the marginal likelihood $p(\mathbf{ys}|\mathbf{Xs}; \{\sigma_f,\, l,\, \sigma_n\})$. The prediction at time $t_{m^+}^+$ for the observation $\mathbf{x}_{m^+} = [x_k^+(t_{m^+}^+)]_{m^+k}$ is given by the mean function, μ and variance function, σ^2,

$$\begin{aligned} \mu_{\mathbf{x}_{m^+}} &= \mathbf{k}_+^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{ys} \\ \sigma_{\mathbf{x}_{m^+}} &= k(\mathbf{x}_{m^+}, \mathbf{x}_{m^+}) - \mathbf{k}_+^\top (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_+ \end{aligned} \tag{3.8}$$

[1] The superscript represents the j^{th} subject and not the exponent.

where $\mathbf{k}(\mathbf{x}_{m^+})$ is denoted as $\mathbf{k}_+$, and $\mathbf{k}(\mathbf{x}_{m^+}) = [k(\mathbf{x}_{m^+}\, \mathbf{x}_1^1) \cdots, k(\mathbf{x}_{m^+}\, \mathbf{x}_{m^N}^N)]^\top$.

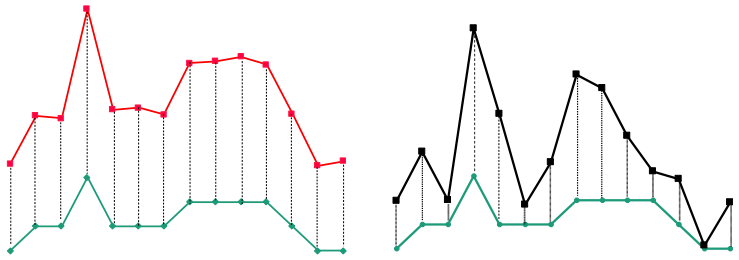
Autoregressive Gaussian Processes (AR-GP) [22] Peterson et al. [22] employ auto-regressive Gaussian processes (AR-GP) to predict the cognitive decline of Alzheimer's disease patients over the next four time steps. This is further discussed in section 3.6.2. They start by building a population-level forecast model using data from training subjects. They use domain-adaptive GPs to sequentially adapt the GP posterior for the test subject using the available data from the test subject. In contrast to our direct technique for multi-step prediction, this is accomplished via an iterative strategy by utilising the data up until time $t-1$ to predict the response at time t. The predictions made are then used again with training data to predict time instant $t+1$ and so on.

3.5 Methodology

In this section, the subset selection (SS) based gaussian process (GP) approach (SS-GP) is introduced. First, a novel approach for SS is described. Second, we develop an algorithm to align the time series in the subset with a target time series.

3.5.1 Dynamic subset selection

Given a discrete time series $\mathbf{y}^{ref}$, and a collection of N time series $\mathbf{y}^j (1 \leq j \leq N)$ we want to find a time series $\mathbf{y}^{sim} \in \mathbf{y}^j$ that is closest to $\mathbf{y}^{ref}$, i.e. $dist(\mathbf{y}^{sim}, \mathbf{y}^{ref}) < dist(\mathbf{y}^j, \mathbf{y}^{ref}) \forall j \in \{1, N\}$[24]. The closeness is calculated by matching time points in two time series based on a distance metric *dist*. For example, to calculate the Euclidean distance between two *equal-length* time series $\mathbf{y}^p = [y_1^p, y_2^p, \cdots, y_m^p]$ and $\mathbf{y}^q = [y_1^q, y_2^q, \cdots, y_m^q]$ a *one-to-one* matching is performed to calculate the distance as $dist(\mathbf{y}^p, \mathbf{y}^q) = \sqrt{\sum_{t=1}^{m} (y_t^p - y_t^q)^2}$. Fig. 3.2 shows examples of *one-to-one* time-point matching with Euclidean distance (dotted line) with Fig. 3.2a exhibiting more similarity with a Euclidean distance of 0.1 as compared to Fig. 3.2b that has a Euclidean distance of 5.1 compared to a reference time series.



(a) (b)

Figure 3.2: Normalised Euclidean distances between (a) similar time series and (b) dissimilar time series. The reference time series that is considered is shown in dark green.

Distance Measurement

Remember that our goal is to make predictions of the response variable $y^+(t)$ for $t > t_d^+$. Our aim in this section is to find a subset of response variables $y^j(t)$ $(1 \leq j \leq M)$ that show similar temporal characteristics with $y^+(t)$ for $t < t_d^+$. This will lead to a subset $\hat{\mathcal{X}} = \{(\mathbf{X}^1, \mathbf{y}^1), \cdots, (\mathbf{X}^M, \mathbf{y}^M)\}$ with $M << N$ of the training dataset $\{Xs, ys\}$ that is used in a non-parametric GP approach for predicting $y^+(t)$. For this purpose, we start by calculating the distances between target response time series data $\mathbf{y}^+ = [y^+(t_1^+)\, y^+(t_2^+) \cdots y^+(t_d^+)]$ and training data's response variable (nearest to the allowed time point,i.e. '$\leq t_d^+$'). Let's denote this distance vector as $\mathbf{\Omega}_+ = [\omega_{1+}\, \omega_{2+} \cdots \omega_{N+}]^T$, where $\omega_{j+} = dist([y^j(t_1^j) \cdots y^j(t_d^j)], [y^+(t_1^+) \cdots y^+(t_d^+)])$. In contrast to equal-length time series in Fig. 3.2, it is difficult to determine the Euclidean distance (dissimilarity) between two time series with unequal lengths. Therefore, we use Dynamic time warping (DTW) [25] as a distance metric *dist* in our study that allows *one-to-many* matching and thus subsumes Euclidean distance. DTW distance has an ability to match time series of different lengths and is robust to shifting and scaling along the time axis [26]. It matches two time series by (i) calculating a local cost matrix between each pair of elements between these time series, and then the goal of minimising the overall cost (distance) is achieved by (ii) finding an optimal alignment that runs along a low cost "valley" within the cost matrix [27].

Fig. 3.3 illustrates that DTW first aligns the time series. Points of the time series that are matched are connected by a dotted line. The final distance is computed by taking the sum of the Euclidean distances between the matched points. Clearly, the reference time series (in green) is more similar in trend to the time series shown in Fig. 3.3b compared to the one shown in Fig. 3.3a.

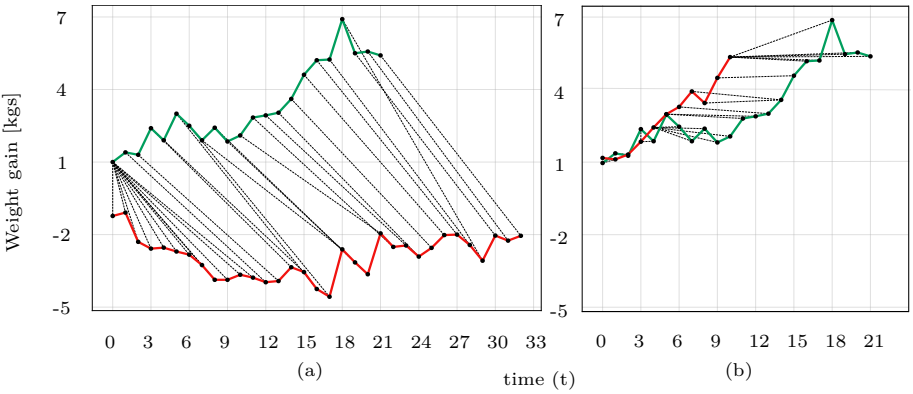


Figure 3.3: DTW distances between time series with different lengths. The matched points are indicated by a dotted line. The reference time series is shown in dark green. In (a) the DTW distance is 170 and the time series are more dissimilar than in (b) where the DTW distance is 6.9.

Since we are calculating the DTW distances in the output space, i.e., between the response time series' ($\mathbf{y^j}$), the distance measurement is applicable in settings where the input time series is multivariate. As long as the output time series is univariate the DTW distance can be calculated as proposed, which is the case in many healthcare applications. For a multidimensional DTW treatment, the reader is referred to [28].

Subset selection

After calculating the distance vector $\mathbf{\Omega}_+$ of length N between a target time series $\mathbf{y}^+$ and other time series' ($\mathbf{y}^j$), the nearest subjects are determined by dynamically calculating a cut-off point for the target time series in the following way :

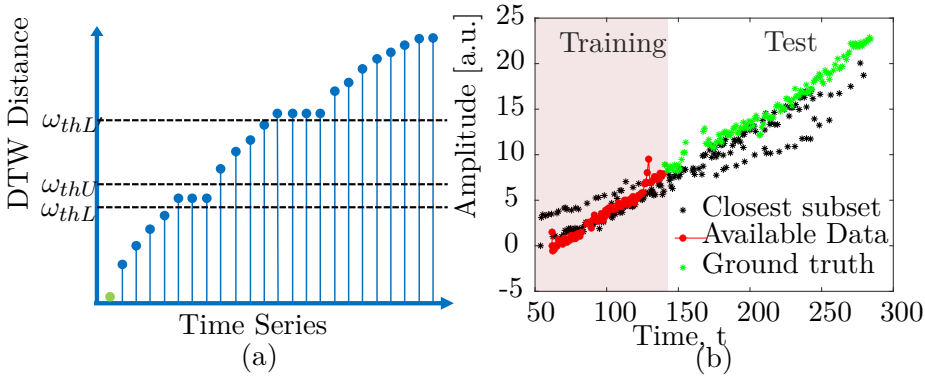


Figure 3.4: (a) DTW distances, dissimilarity measures between time series, plotted in ascending order with some possible choices of threshold values. (b) Proposed heuristic is used to calculate the closest subset on the training part (in light purple, $t < t_d^+$) and the test part for subject 1.This illustration is from the gestational weight gain prediction use-case explained in section 3.6.2

1. *Arrange elements by their closeness to the target time series* : sort the distance vector $\mathbf{\Omega}_+$ in increasing order as $\hat{\mathbf{\Omega}}_+ = [\hat{\omega}_{1+}\, \hat{\omega}_{2+}\, \cdots\, \hat{\omega}_{N+}]^\top$, such that $\hat{\omega}_{k+} \leq \hat{\omega}_{(k+1)+} \forall k \in \{1, 2, \cdots, N\}$.

2. *Select cut-off for subset selection when the rate of change of DTW distance is high*: calculate 'turning points' at index 'k' such that the absolute rate of change of DTW distance is highest in the local neighbourhood (± 1 index), $\left(\hat{\omega}_{(k-1)+} - \hat{\omega}_{(k-2)+}\right) \leq \left(\hat{\omega}_{k+} - \hat{\omega}_{(k-1)+}\right) \geq \left(\hat{\omega}_{(k+1)+} - \hat{\omega}_{k+}\right)$.

3. *Choose a turning point for subset selection* : choose the value at the first turning point '$\hat{\omega}_k$' as our threshold ω_{th} for finding the closest time series set $\hat{\mathcal{X}}$. The closest selected subset consists of all time series whose DTW distance is less than this threshold compared to the target time series.

Note that more turning points can be calculated by choosing the next minimum as described further.

The intuition for *turning* points is represented in Fig. 3.4a, which shows the DTW distances measures between response variables of target and the training time series in ascending order. The possible choices of thresholds calculated as defined by *turning* points occur at locations $\omega_{th_L}, \omega_{th_U}, \omega^{'}_{th_L}, \cdots \omega^{'''}_{th_L}$. Note that ω_{th_L} represents the point where the first minimum occurs in the rate of change in DTW distances. Similarly, multiple such turning points exist that can be used as thresholds represented with the prime ($'$) symbol. Intuitively, ω_{th_U} can be considered as another appropriate choice for threshold. However, the first value of turning point, ω_{th_L} is chosen as the preferred threshold. It selects the "smallest" most informative subset from the training data to capture the trend while keeping the variability among the selected subset to a minimum as compared to other thresholds. For the sake of simplicity, Fig. 3.4b shows a univariate time series of subject 1 from a dataset (explained in section 3.6.2) and the selected closest subset according to the proposed heuristics. Using the proposed heuristics, the subjects that are closer in the training phase (coloured in red) show a similar trend in the forecasting phase.

Using the SS approach proposed above, we can find a subset $\hat{\mathcal{X}}$ from $\{\mathbf{Xs}, \mathbf{ys}\}$. The subset $\hat{\mathcal{X}}$ contains time series that are similar to the target time series and are therefore expected to contain the most essential information for forecasting the target time series data. The subset $\hat{\mathcal{X}}$ will be used to train a non-parametric GP in the proposed SS-GP approach. The computational complexity of a GPs depends on the number of training points n according to $O(n^3)$. Restricting the training of the GPs to the subset $\hat{\mathcal{X}}$ will considerably reduce the computational complexity (as compared to a training on the complete data set $\mathbf{Xs}$) because $n(\hat{\mathcal{X}}) << n(\mathbf{Xs})$. Moreover, we will show through our case studies that an increase in prediction performance can be obtained.

Additionally, such a localised non-parametric distance-based approach allows for the selection of neighbours based on the temporal nature of the data. This makes our approach generally applicable with other learning methods where priors are formed based on the available closest time series data.

3.5.2 Collective temporal realignment

Typically, it is assumed that the time series in the training data are available from some fixed time $t = t_0$. However, in practical scenarios, the time series in the dataset may have different onsets and rates of progression.

Dynamic time warping (DTW) accounts for the similarity in amplitude among time series by calculating the distance between them. It realigns the two time series non-linearly, onto a common set of instants such that the sum of the Euclidean distances between the corresponding points, is smallest. We propose a time series alignment based on the shape of the response variable. We try to find a time instant $\tau_{optimal}$ with respect to the target response series such that when the response time series in the training dataset are lagged/led by $\tau_{optimal}$, their shape most resembles that of the target's response time-series. For a given target response variable ($\mathbf{y}^+$), we realign the time series in $\mathbf{Xs}$ in time. We hypothesise

Algorithm 1 Temporal realignment for target data

procedure TEMPORAL REALIGNMENT
  $Input : \mathbf{y}^+ = [y^+(t_1^+)\, y^+(t_2^+)\, y^+(t_3^+) \cdots y^+(t_d^+)]$
  $lags = [-\tau_d, \cdots, -\tau_1, 0, \tau_1, \tau_2, \cdots, \tau_d]$
  $Output : \tau_{optimal\, N\times 1}$
  for $i = 1$ to N **do**
    $\mathbf{y}^i = [y^i(t_1^i)\, y^i(t_2^i) \cdots y^i(t_d^i)]$
    $minDist = Inf$
    for $iter = 1$ to $2d + 1$ **do**
      $\tau = lags(iter)$
      $curDist = dist(\mathbf{y}^+, \mathbf{y}^i_{(t+\tau)})$
      if $curDist < minDist$ **then**
        $\tau_{optimal}(i) = \tau$

that readjusting the training data with respect to the target data will result in better subset selection. The approach is as follows,

1. Given target data observations of the response variable until time t_d, calculate distance from lagged/led versions of N time series in the training data using the metric $\sqrt{(\sum_{n=1}^{d} y^+(t_n^+) - y^i(t_{n+\tau}^i))^2}$

2. For the j^{th} time series in the training data, the value of τ_{k+} that minimises the above metric is $\tau_{optimal}(j)$

3. We then create lagged/led versions of predictor and response variables in the j^{th} training data using $\tau_{optimal}(j)$ for the given target time series. This gives us the temporal fitted lagged/led version of $\mathbf{Xs}^{\text{aligned}}$.

Before proceeding with DTW-based subset selection, the collective temporal realignment is performed as a pre-processing step. The goal of collective temporal realignment is to adapt the time series' in training data with respect to the time series' in test data because they may have different times of initiation. The training data's input and output time series are then realigned based on the computed delays.

DTW-based dynamic subset selection is then applied to get $\hat{\mathcal{X}}$. After temporal realignment and dynamic subset selection based on the available target data ($\mathbf{y}$) in the training and test dataset we apply Gaussian processes based prediction on $\{\hat{\mathcal{X}}, \mathbf{X}^{+}\}$ as it is most resilient to the missing data in time series. We use the selected $M << N$ time series that are in the closest subset of a given time series along with eqn. (3.8).

Mean Absolute Error (MAE) is used as the performance metric to evaluate the regression performance.

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |y(t^{j}_{m_j}) - y_{pred}(t^{j}_{m_j})|$$

.

3.6 Experiments

We start by describing the setup of our experiments and the methods that we use to benchmark the proposed SS-GP approach. Furthermore, we give a detailed description of the use cases we will treat.

3.6.1 Baseline

Parametric. We fit a 3rd order polynomial on the response variable varying with time. First, an MLE estimate is made on all the subjects in training data. These model estimates are used as prior distribution to calculate a maximum-a-posteriori estimate (explained in section 3.4.2) to learn a final model. The response variable for a given test subject is then predicted using this final model at a given time instant. This is done in a leave-one-subject-out fashion so that each subject's data is estimated once. A 3rd order polynomial is used as it provides the least mean absolute error among other orders (1 to 5) of polynomials for both the data sets.

ARIMA. ARIMA has a limitation that it only works well with uniformly sampled data. This is difficult when data are missing. We fit an ARIMA*(p,d,q)* model on the response variable of the target data as follows i) linearly interpolating the data to make the data evenly-spaced in time, ii) tuning the hyperparameters [29] to find an optimal autoregressive order, degree of differencing, and moving average order by performing a grid search, iii) using the optimised hyperparameters over the training part to forecast at a time instant (given data until day t_d). This is done for each test subject.

LSTM. We evaluate an LSTM-based regression network with 200 hidden units. The training is done using Adam's optimisation to minimise the mean absolute error [30]. The hyperparameters search space was set as follows, epochs $= 50, 100, 150, 200, 250$, learning rate $= 0.0001, 0.0005, 0.001, 0.005$, batch size $= 16, 32, 64, 128$

AR-GP. AR-GP are trained to forecast the response variable using the input and response features until time t. For each subject, the missing observations are filled using the forward filling approach, where data from a previous observation are carried over to the following observation. When the training matrix is completed, the parameters of AR-GPs are learned by minimising the negative log-likelihood [22].

AR-GP + MICE. Multivariate imputation by chained equations (MICE) is an imputation strategy for matrix completion [31]. It works by iteratively building predictive models to fill each specified variable in

the matrix. Each variable is imputed using other variables in the dataset and the iterations are run until convergence is met. AR-GP works by first forward filling the data to complete the matrix for training. We also use the state-of-the-art MICE approach to impute the data and then apply AR-GP to compare the performance.

We evaluate these methods on different univariate and multivariate real-life datasets in a leave-one-subject-out cross-validation scenario. A detailed explanation of how the proposed model is compared with baselines in different datasets is described as follows.

3.6.2 Datasets

Health progression modelling requires longitudinal data from a person that can provide long-term predictions for disease status of an individual. Often, this data exists in the form of electronic health records or sequence readings collected over time. Current state-of the-art methods such as deep learning methods provide accurate models of individuals' health status in case of big data sets where both the number of individuals and the number of individual measurements through time are large [32]. However, in the presence of limited training data (small N and $t_d^+ \approx 0$), such as when early disease discovery is of utmost importance, such approaches produce sub-optimal results. Our framework for time series-prediction in the absence of missing or limited data can enhance health prediction capabilities. Hence, we select two datasets from real life presented as follows:

Gestational Weight Gain

One health demographic is managing gestational weight gain among women. Approximately 70% of pregnant women gain either too little or too much weight at the end of their pregnancy in accordance with the Institute of Medicine recommended guidelines [33]. Inappropriate weight gain during pregnancy has been associated with short- and long-term health complications to the mother and baby. Thus, early recognition of signs of weight gain during pregnancy is essential [2]. In this study, data were collected from diverse subjects in Europe where 80 women

Attribute	**Mean ± Std (80 subjects)**
Age (years)	31 ± 3.5
Height (meters)	1.69 ± 0.07
Pre-pregnancy weight (kgs)	69 ± 15
Pre-pregnancy BMI (kgs/m^2)	24 ± 4
Delivery (days)	277 ± 10
Weight Gained (kgs)	13.7 ± 4.7
Number of weight gain samples	59.83 ± 41.02

Table 3.1: Dataset description for univariate gestational weight gain data

in their fifth week of pregnancy or later were recruited from midwife practices in Eindhoven, The Netherlands. The weight data were collected by a WiFi-connected scale, Withings WS30[2]. The dataset is described in Table 4.1. Note that this is a case of univariate time series data where only one variable (weight gain) is measured with respect to time. A mobile application allowed participants to log their weights weekly, and the weight data was sent to the cloud.

The participants provided an informed consent pre-data collection and the study was approved by the Internal Ethics Committee for Biomedical Experiments of the involved organisations (ICBE Reference number 2015-0079 respectively).

We model the weight (gain) $\mathbf{y}$ as a function of time, $(t_1^j, t_2^j, \cdots, t_m^j)$ using the proposed approach. We achieve this by first normalising measured weight with pre-pregnancy weight to obtain weight gain data and then fitting various forecasting approaches. For the parametric approaches, we utilise the complete data from $N-1$ subjects to generate a prior to estimate a MAP model. We experiment with first, second and third-order polynomial based parametric approaches to fit our time-series data. In cross-validation, we obtain the polynomial order ($= 3$) empirically for the parametric approach, which has the lowest prediction error among all other orders.

[2]https://www.withings.com/

For the proposed non-parametric approach, we use the data from $N-1$ subjects as training data in addition to the available target data of the remaining subject to train the Gaussian processes in the baseline setting. Dynamic subset selection is performed on the training data with respect to the available target data.

Alzheimer's disease prediction

Another health complication is Alzheimer's disease (AD). AD is a neurodegenerative disorder and the most common form of dementia. Prediction of this progressive disorder's symptom onset at early stages is urgent and complex [34]. The design of clinical trials and developing therapeutic interventions depends on accurately detecting patients at the early stages of the disease where treatments are most likely to be effective. The clinical status of an Alzheimer's patient is based on commonly used cognitive scores namely, the mini mental state examination (MMSE) [35], the Washington University Clinical Dementia Rating Sum of Boxes score (CDRSB) [36], and the AD Assessment Scale-Cognitive subtest (ADAS-Cog13) [37].

To this end, we use the data collected as part of the TADPOLE challenge [38] by the Alzheimer's Disease Neuroimaging Initiative (ADNI) consortium[3] [39]. The data from 1737 patients taken every six months over the course of 120 months consists of different modalities such as (1) various features extracted from imaging modalities like magnetic resource imaging (MRI), positron emission tomography (PET) and diffusion tensor imaging (DTI), (2) cerebro-spinal fluid (CSF) markers of amyloid beta and tau-deposition; (3) cognitive assessments measured in the presence of a clinical expert; (4) genetic information such as alipoprotein E4 (APOE4) status from DNA samples and (5) general demographic information [38]. Around 266 features were extracted based on these modalities and merged together over time to form a coherent numerical multivariate time series feature set. Since the complete dataset has a lot of missing visits, we follow the state-of-the-art approach for Alzheimer's disease marker forecast [22] and selected a smaller dataset of 95 subjects such that data from at least ten visits is present and missing data is no

[3] http://adni.loni.usc.edu/

more than 82.5% of the feature set. This helps in benchmarking our proposed approach with the AR-GP approach [22].

In the case of this multivariate time series data, our experimentation to predict a cognitive score (MMSE, ADAS or CDRSB) using 266 features that vary with time is as follows:

1. Collective temporal realignment: The Alzheimer's study [39] recruited patients that were already going through some stage of cognitive decline. Since the disease progression in every individual differs in their onset, the target time series ($\mathbf{y}$) for each of the patients had a different t_0. Therefore, we calculate the value of $\tau_{optimal}(j)$ using the response variable of the target data and the response variable of the j^{th} subject in the training data. This lag is calculated for all the subjects in the training data with respect to a given target subject. Based on the calculated $\tau_{optimal}(j)$, lagged/led versions of the predictor ($\mathbf{X}^j$) and the response ($\mathbf{y}^j$) are created to be used for further training.

2. Subset selection based on the response variable $\mathbf{y}$: Based on the available target data ($\mathbf{X}$,$\mathbf{y}$) from a given test subject until month t_d, we find the subjects in the training dataset with a similar cognitive decline. This is done by applying the subset selection approach explained in section 3.5.1 on $\mathbf{y}^j$. Note that we apply subset selection on the response variable instead of $\mathbf{X}$ since it gives us similar subjects in output space.

   We also compared the performance of our subset selection approach with a k-means clustering approach. Clusters were obtained using the input features of the multivariate time series. For a given test subject, the subjects in the closest cluster are considered as training data. These training subjects along with the available test subject's data are used to train a non-parametric GP as follows: (1) the missing values are forward filled (2) K clusters with centroids $\{\mathbf{c}_k\}$ are created using k-means clustering with training data matrix until month t_d (3) calculate distance $\mathbf{d}_k = dist(\mathbf{X}^+, \mathbf{c}_k)$ of the test subject's predictor variables data $\mathbf{X}^+$ from each centroid $\{\mathbf{c}_k\}$ and the optimal profile (subset) is selected as cluster $\mathbf{c}_{opt} = \mathbf{c}_i$ such

that, $d_i < d_j, \forall i \neq j \in [1, K]$ In what follows, we will refer to this method as the "K-means + GP" approach.

Disease progression Estimation : We perform non-parametric regression using Gaussian processes on the input feature set $\mathbf{X}$. First, given a test subject's response variable $\mathbf{y}^+$, a time aligned training data is made that consists of a lagged version of $\mathbf{Xs}$ and $\mathbf{ys}$. Subsequently, subset selection is performed by finding subjects in the training data whose response variable $(\mathbf{y}^j)$ is close to the test subject's response variable $(\mathbf{y}^+)$. Once a subset is selected, GP based regression is performed on $[\mathcal{X}, \mathbf{X}^+]$ using eqn. (6.5) and (3.8). We perform leave-one-subject out cross-validation on the dataset.

In both cases, developing models to automatically predict Alzheimer disease-related metrics or gestational weight gain is of utmost importance to intervene appropriately and in time. This makes the availability of the target data another challenge. To test how well these methods can perform with limited target data, we experiment by varying the amount of available target data with respect to time, i.e, $0 \leq t_d^+ \leq t_m^+$.

Remark that, in the case of gestational weight gain prediction, the objective is to predict a single observation in time, i.e., the end-of-pregnancy weight gain. The performance is measured by predicting the end-of-pregnancy (≈ 270 day) weight gain for a test subject when data was available until $120, 130, 140, \cdots, 260$ days.

In the case of Alzheimer's disease, however, we are also interested in the disease's trajectory, not merely an ultimate endpoint prediction. Two subsequent visits are spaced an average of 6 months apart, and we will predict the disease progression for each month despite having little data (i.e., only observations from the first 30 months are used in the training phase to predict progression up to month 120).

We evaluate and present the results related to the performance of different approaches across time with different availabilities of the target data. We also benchmark our proposed approach with the K-means + GP approach and the AR-GP approach [22], the latter of which is considered as a state-of-the-art approach for predicting cognitive decline of Alzheimer's patients.

3.7 Results & Discussion

3.7.1 Gestational weight gain prediction

We study the performance of various forecasting algorithms when predicting the weight gain of a target subject for the end of the pregnancy while the time series data of weight gain of the target subject are only available up to time t_d^+. The most crucial aspect of gestational weight gain prediction is whether the weight at the end of the pregnancy is in the range recommended by the IOM guidelines [33]. Therefore, we investigate the weight gain prediction ability of forecasting algorithms that are trained with changing availability of a target individual's data (i.e, varying t_d^+). The results can be found in Fig. 3.5, which shows the prediction error averaged over all the subjects as a function of the moment t_d^+. The prediction error reduces when more training data is available. Also, it can be observed in Fig. 3.5 that the GP approach performs worse than the SS-GP approach. Based on a paired t-test, which assumes equal variances, we found that all differences between performances of the SS-GP model and the other models are statistically significant at a significance level of 5%. Only for the SS-GP and the MAP model performances, no statistically significant difference was found. This is not unexpected because of the simplicity of the dataset. For $t_d^+ > 220$, the performance of ARIMA significantly outperforms the performance of all other approaches. However $t_d^+ = 220$ is too close to the horizon to result in effective intervention. Note that the average delivery day is around day 277. The benefit of using our SS-GP approach is further illustrated in Fig. 3.6.

In Fig. 3.6a, we show the performance of GPs when all training data is used to make predictions for a target subject. Since the training data consists of subjects with various rates of weight gain, the predicted trend in the target subject is influenced by all the measurements in the training data at a given time. The variability in the prediction is reduced by selecting a subset of time series from the training data that share similar patterns with the target data. These subjects are then used to forecast the target data, as shown in Fig. 3.6b.

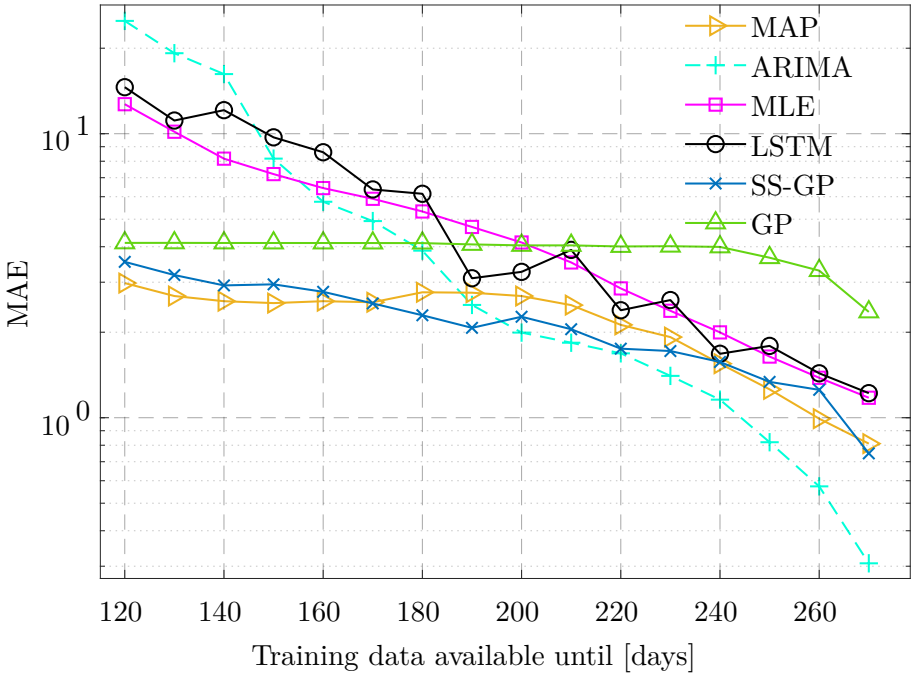


Figure 3.5: MAE of predicted weight on delivery day (multiple steps ahead in time) with respect to different approaches. MAE reduces as more training data becomes available.

3.7.2 Alzheimer's disease prediction

Unlike the gestational weight gain use case, where the final objective was to predict the end-of-pregnancy weight gain because the data was recorded daily, we aim to predict the progression of Alzheimer's disease at each visit, since these visits are separated by six months or more. For this purpose, we will study the performance of several methods for predicting three metrics for cognitive decline that are commonly used by clinicians and that were introduced in section 3.6.2: MMSE, ADAS13, and CDRSB.

Following our realignment approach, we first calculate the optimal τ for each response time series (cognitive score) in the training dataset with respect to the available response data from the test subject. We

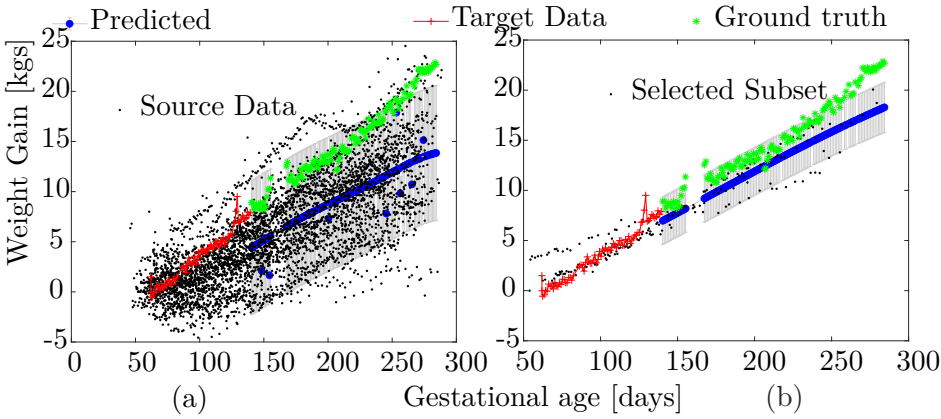


Figure 3.6: Prediction error ($i = 1^{th}$ subject) is (a) high (low confidence) when the complete training dataset is considered due to inter-subject differences but (b) reduces using close subset selection based on heuristics. The prediction confidence (grey) also increases using the SS approach.

compute the standard deviation at a particular time instant for all response time series in the training data that are aligned with respect to the target subject. This standard deviation should be smaller than when no alignment is performed. We experimented with all the subjects in a leave one out fashion. In Fig. 3.7 each line depicts the standard deviation of the ADAS13 matrix created using aligned versions of the time series for a given test subject. By computing the standard deviation without alignment, a baseline was established. We observed that $> 80\%$ of the subjects have a standard deviation less (more desirable) than the baseline when adjusted for the alignment using our temporal realignment approach. This shows that most of the subjects are adjusted in time with respect to disease progression after realignment.

To predict the Alzheimer's disease progression, we varied the availability of target data from month 30 until month 108. Fig. 3.8 shows the cross-validation results, averaged over all subjects, for the prediction of ADAS13 using our SS-GP approach. Each line in Fig. 3.8 corresponds to a different number of available measurements for the test subject. Given the available training data until a specific month, each point on this line represents the prediction error in forecasting cognitive score for the month depicted on the x-axis, averaged over all subjects. One can

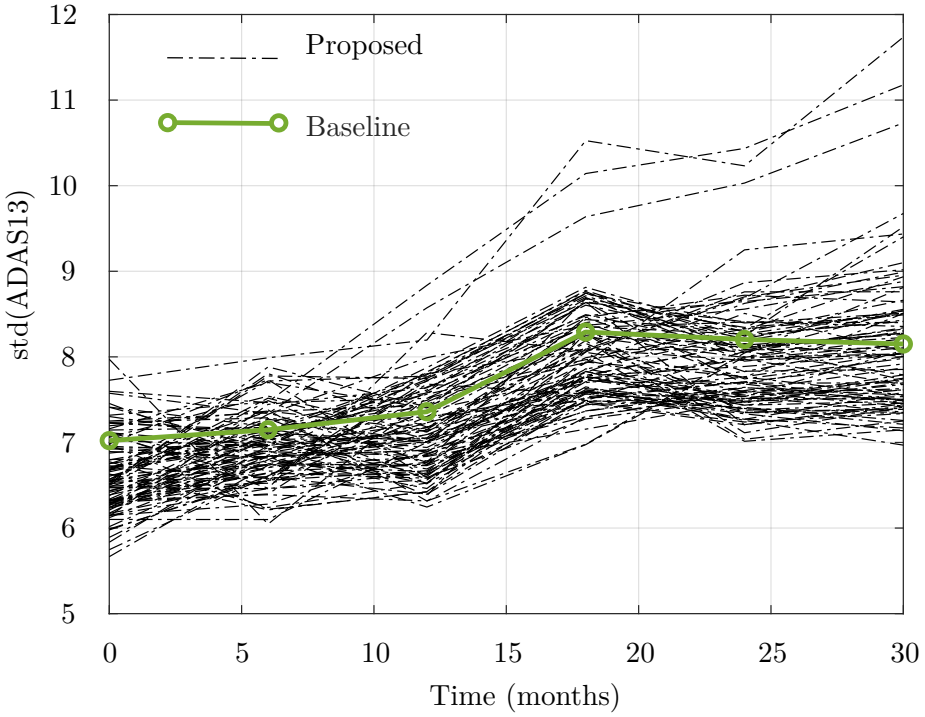


Figure 3.7: Standard deviation (std) of the cognitive decline (ADAS13) after the proposed alignment for each subject (in black). *The closer the std is to the x-axis, the more similar the subjects' time series are.*

observe from Fig. 3.8 that there is an increasing trend in prediction error when the forecast horizon increases. For example, given the training data availability until month 30 (orange line with + marker), the mean absolute error when predicting for month 60 is higher than for month 36. Additionally, Fig. 3.8 shows that the prediction performance improves as more data from the test subject becomes available for training. For instance, for the predictions at month 48, the MAE obtained when training data are available up till month 42 is smaller than the MAE obtained when training data are available up till month 30. The other metrics for cognitive decline (MMSE and CDRSB) were found to show similar patterns in prediction performance.

As seen in Fig. 3.8, the worst result is observed when the available

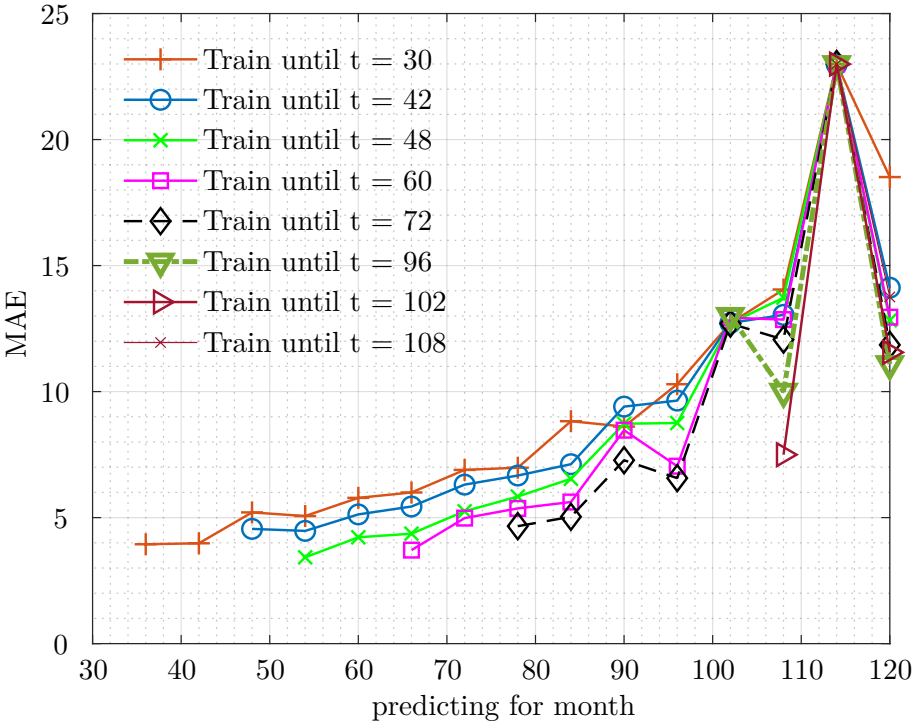


Figure 3.8: Mean absolute error measured with respect to different data available in time for different steps in time prediction for ADAS13. The average MAE for a specific month is lower when the data availability is higher.

training data is highly limited, and the forecast horizon is set far in time. This occurs when training data are only available until month 30 and forecasts are made up till month 120. Since we need to predict as early as possible, we present the results for all further experiments when training data are only available until month 30, and the forecast horizon stretches from month 36 to month 120. Using training data up till month 30 ensures us that during training at least one data point of each subject is included.

Furthermore, at month 114, a peak in MAE is observed in Fig. 3.8. This is due to the fact that, at month 114, no measurements of the target variable are available for a lot of subjects.

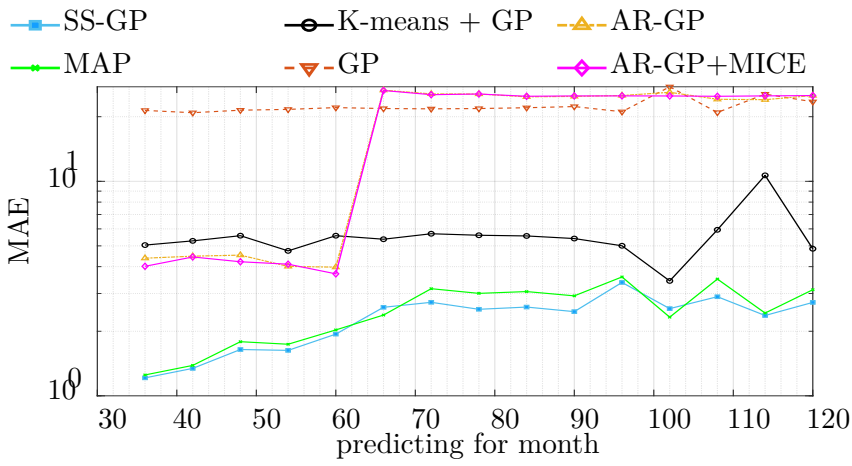
Fig. 3.9 shows the forecast performance for the three cognitive metrics for different subset selection strategies and different regression approaches. For the AR-GP approach a forward filling approach is used to deal with missing data [22]. However, we also studied the performance of the AR-GP approach when a state-of-the-art imputation technique is used instead, i.e. a multivariate imputation by chained equations (MICE) for matrix completion [31]. We refer to this combination as AR-GP + MICE.

Note that on average, a 1-3 point decrease in Mini Mental State Examination [40], a 1-2 point increase in Clinical Dementia Scale sum of boxes [40], and a 3-3.1 point increase in Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog)[41] are indicative of a meaningful decline.

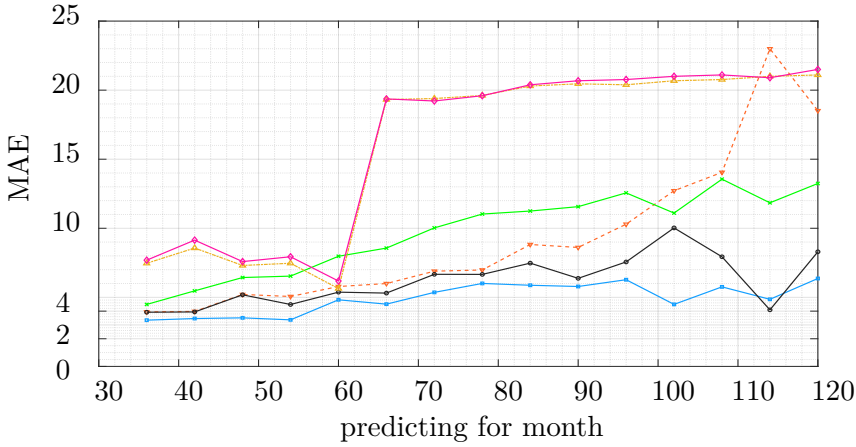
The differences between the proposed and compared models are statistically significant ($p < 0.05$) based on a paired t-test with equal variances. However, compared to the SS-GP approach, no statistical difference is found with the MAP approach when predicting MMSE and with the K-means + GP approach when predicting ADAS13 or CDRSB. Thus, we can conclude that our method performs consistently (equal if not better) across all metrics of cognitive decline when compared with the state-of-the-arts.

Note the methods MLE, ARIMA, and LSTM provided as state-of-the-arts train a model using only the test subject's limited personal data, as these algorithms can handle the time series data from a single participant at a time in a personalised manner as described in section 2.7.

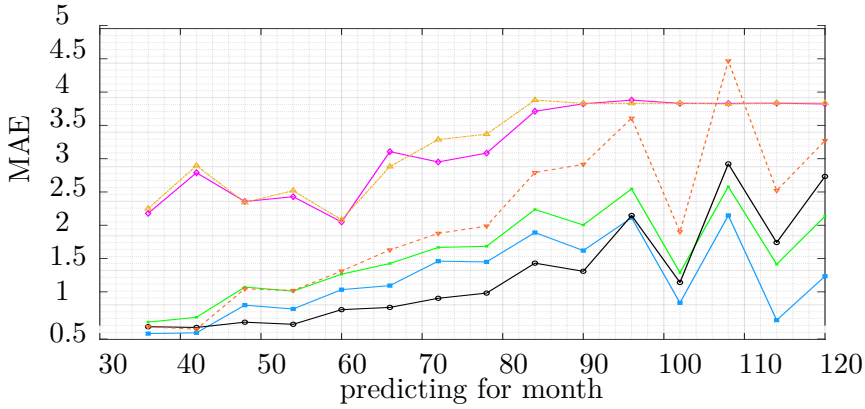
The proposed approach (SS-GP) is also compared to MAP (an extension of the MLE), GP, AR-GP, AR-GP+MICE that utilise other subjects' data along with personal data to improve the model's performance.



(a)



(b)



(c)

Figure 3.9: Proposed approach achieves lowest MAE on the metrics (a) MMSE and (b) ADAS13 and comparable MAE with k-means based clustering on (c) CDRSB.

3.8 Conclusion

In this article, we proposed a novel approach, termed the SS-GP approach, for forecasting time series that are not necessarily uniformly sampled. For this purpose, we combined the non-parametric GP regression with a subset selection procedure that selects a set of time series from the data that closely resembles the test subject's data. Our subset selection procedure is robust as it selects the subset size dynamically based on temporal similarities between the time series in the subset and the test time series. The temporal similarity is measured with a DTW distance that can be computed between time series with a different length. We validated this method on two use cases and compared it with several other approaches.

Firstly, on the univariate gestational weight gain dataset, our approach performs similar to a parametric polynomial fitting which is not unexpected because of the simplicity of the data set. However, the SS-GP is able to reduce the variability in predictions because predictions are only based on time series data that share similar patterns with the data of the test subject.

Secondly, for a more complex data set consisting of multivariate time series data to predict cognitive decline of Alzheimer's patients our SS-GP approach is able to outperform state-of-the-art approaches such as the AR-GP approach [22]. In particular, the SS-GP approach, improves prediction results when the forecast horizon is long and only a limited amount of data is available.

3.9 Limitations & Future Work

Although effective in regression when data is missing, Gaussian Processes (GPs) have a high computational complexity of $\mathcal{O}(n^3)$. Our subset selection is a local approximation technique that decreases complexity by including only the most useful training points ($<< n$) that are close to the test point. However, the collective realignment technique has a high time complexity because it determines the ideal alignment for a specific test time series by comparing it to all the time series in the training

dataset. In future research, we would like to experiment with another scalable sparse approximation of GPs developed in [42] that can further reduce the time complexity.

In addition, the proposed approach is only tested on data sets from healthcare considering the necessity that arises in this domain from data acquisition limitations. In an environment where data acquisition is costly, it would be beneficial to evaluate our method on more data sets and application domains where time series can be sparsely sampled, such as process quality monitoring in industries.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This publication reflects only the authors' view, and the REA is not responsible for any use that may be made of the information it contains.

Bibliography

[1] C. Puri, G. Kooijman, B. Vanrumste, and S. Luca, "Forecasting time series in healthcare with gaussian processes and dynamic time warping based subset selection", *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6126–6137, 2022. DOI: 10.1109/JBHI.2022.3214343.

[2] C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278.

[3] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2017.

[4] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors", *Sensors*, vol. 15, no. 12, pp. 31 314–31 338, 2015.

[5] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel, "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models", in *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, 2012, pp. 389–398.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 95–104.

[8] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber", in *International Conference on Machine Learning*, vol. 34, 2017, pp. 1–5.

[9] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "GRU-ODE-bayes: Continuous modeling of sporadically-observed time series", in *Advances in Neural Information Processing Systems*, 2019, pp. 7379–7390.

[10] M. Liu, A. Zeng, Z. Xu, Q. Lai, and Q. Xu, "Time series is a special sequence: Forecasting with sample convolution and interaction", *arXiv preprint arXiv:2106.09305*, 2021.

[11] C. E. Rasmussen, "Gaussian processes in machine learning", in *Advanced lectures on machine learning*, Springer, 2004, pp. 63–71.

[12] J. G. De Gooijer and R. J. Hyndman, "25 years of time series forecasting", *International journal of forecasting*, vol. 22, no. 3, pp. 443–473, 2006.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling", English (US), in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[14] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[15] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation", in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[16] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series", in *Machine Learning for Healthcare Conference*, 2016, pp. 253–270.

[17] N. Strodthoff and P. e. Wagner, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2020.

[18] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting", *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 414–423, 2019.

[19] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask gaussian process rnn classifier", in *International conference on machine learning*, PMLR, 2017, pp. 1174–1182.

[20] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward", *PloS one*, vol. 13, no. 3, e0194889, 2018.

[21] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian process regression in vital-sign early warning systems", in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, IEEE, 2012, pp. 6161–6164.

[22] K. Peterson, O. Rudovic, R. Guerrero, and R. W. Picard, "Personalized gaussian processes for future prediction of alzheimer's disease progression", *NeurIPS Workshop on Machine Learning for Healthcare.*, 2017.

[23] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, "The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances", *Data mining and knowledge discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[24] E. J. Keogh and M. J. Pazzani, "A simple dimensionality reduction technique for fast similarity search in large time series databases", in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2000, pp. 122–133.

[25] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series.", in *KDD workshop*, Seattle, WA, vol. 10, 1994, pp. 359–370.

[26] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping", *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[27] M. Müller, "Dynamic time warping", *Information retrieval for music and motion*, pp. 69–84, 2007.

[28] M. Shokoohi-Yekta, B. Hu, H. Jin, J. Wang, and E. Keogh, "Generalizing dtw to the multi-dimensional case requires an adaptive approach", *Data mining and knowledge discovery*, vol. 31, no. 1, pp. 1–31, 2017.

[29] R. Shibata, "Selection of the order of an autoregressive model by akaike's information criterion", *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[31] S. v. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R", *Journal of statistical software*, pp. 1–68, 2010.

[32] S. El-Sappagh, T. Abuhmed, S. R. Islam, and K. S. Kwak, "Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data", *Neurocomputing*, vol. 412, pp. 197–215, 2020.

[33] K. M. Rasmussen, P. M. Catalano, and A. L. Yaktine, "New guidelines for weight gain during pregnancy: What obstetrician/gynecologists should know", *Current opinion in obstetrics & gynecology*, vol. 21, no. 6, p. 521, 2009.

[34] J. L. Cummings, "Challenges to demonstrating disease-modifying effects in Alzheimer's disease clinical trials", *Alzheimer's & Dementia*, vol. 2, no. 4, pp. 263–271, 2006.

[35] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""Mini-mental state": A practical method for grading the cognitive state of patients for the clinician", *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[36] C. P. Hughes, L. Berg, W. Danziger, L. A. Coben, and R. L. Martin, "A new clinical scale for the staging of dementia", *The British journal of psychiatry*, vol. 140, no. 6, pp. 566–572, 1982.

[37] W. G. Rosen, R. C. Mohs, and K. L. Davis, "A new rating scale for Alzheimer's disease.", *The American journal of psychiatry*, 1984.

[38] R. V. Marinescu, N. P. Oxtoby, A. L. Young, E. E. Bron, A. W. Toga, M. W. Weiner, F. Barkhof, N. C. Fox, S. Klein, D. C. Alexander, *et al.*, "Tadpole challenge: Prediction of longitudinal evolution in Alzheimer's disease", *arXiv preprint arXiv:1805.03909*, 2018.

[39] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "The Alzheimer's disease neuroimaging initiative", *Neuroimaging Clinics*, vol. 15, no. 4, pp. 869–877, 2005.

[40] J. S. Andrews, U. Desai, N. Y. Kirson, M. L. Zichlin, D. E. Ball, and B. R. Matthews, "Disease severity and minimal clinically important differences in clinical outcome assessments for Alzheimer's disease clinical trials", *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 5, pp. 354–363, 2019.

[41] A. Schrag, J. M. Schott, A. D. N. Initiative, *et al.*, "What is the clinically relevant change on the ADAS-Cog?", *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 83, no. 2, pp. 171–173, 2012.

[42] E. Snelson and Z. Ghahramani, "Local and global sparse gaussian process approximations", in *Artificial Intelligence and Statistics*, PMLR, 2007, pp. 524–531.

Chapter 4

Privacy-Preserving Learning for Gestational Weight Gain Estimation

This chapter was previously published as:
C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Gestational weight gain prediction using privacy preserving federated learning", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 2170–2174

C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 398–399

Abstract

Gestational weight gain prediction in expecting women is associated with multiple risks. Manageable interventions can be devised if the weight gain can be predicted as early as possible. However, training the model to predict such weight gain requires access to centrally stored privacy sensitive weight data. Federated learning can help mitigate this problem by sending local copies of trained models instead of raw data and aggregate them at the central server. In this paper, we present a privacy preserving federated learning approach where the participating users collaboratively learn and update the global model. Furthermore, we show that this model updation can be done incrementally without having the need to store the local updates eternally. Our proposed model achieves a mean absolute error of 4.455 kgs whilst preserving privacy against 2.572 kgs achieved in a centralised approach.
Clinical relevance - Privacy preserving training of machine learning algorithm for early gestational weight gain prediction with minor trade-off to performance.

4.1 Introduction

In pregnancy, inadequate or excessive weight gain remains a key health issue. Global estimates suggest that only around 30% of pregnant women end up being adequately weighed recommended by the Institute of medicine [3], [4]. There are several risks associated with such excessive or inadequate gestational weight gain, for example, excessive weight gain can lead to fetal macrosomia or post-partum maternal obesity putting the mothers at incresased risk of gestational diabetes [5]. Similarly, inadequate weight gain can lead to small-for-gestational-age infants [3].

Early prediction of gestational weight gain can help mitigate this problem by helping neonatal healthcare providers or expecting women in devising better management and interventions. Traditional approaches exist in which raw data from all the subjects is collected and sent to a central location. At this central location, the data is saved, processed and models to estimate gestational weight gain are trained. Even

though, one can achieve high predictive performance in such a centralised approach, there are several privacy concerns associated with such a model building approach, especially in the light of the General Data Protection Regulation (GDPR) imposed by the European Union (EU) and increased awareness about privacy preservation among end-users. The centralized storage creates a large surface area for security and privacy attacks. It leaves the user lacking control of his/her own personal data. Finally, in applications where the data collected is large in size, especially larger than the model, handling sensitive data on the server side becomes cumbersome as well.

Google proposed federated learning [6] where many local devices collaboratively train a model in association with a central server, while keeping raw sensitive data distributed in the users' own devices. This is made possible by the ubiquity and the improved computational capabilities of the edge devices such as smart-phones. In federated learning, user devices only share model updates with the centralized server after training models iteratively on the local data available on-device. Federated learning is particularly applicable to use-cases where the data is collected from user devices and the data is sensitive in nature. In order to achieve this, we have designed and implemented a privacy-preserving federated approach for the prediction of gestational weight gain. The key contributions of this paper are (a) implementation of federated learning approach for prediction of gestational weight gain, (b) studying the effect of varying number of participants in collaborative learning, and (c) updating the global model incrementally such that the local updates are deleted once they are incorporated into a global model.

4.2 Related Works

Parametric methods such as maximum likelihood estimation or ARIMA [7] approaches have been used traditionally for time series prediction that utilise individual training data. Authors in [8] propose an improvement over these state-of-the-art techniques to predict an individual's end-of-pregnancy weight gain as early as day 140 with an average mean absolute error of around 2.572 kgs. The model is trained by learning an a-priori

model based on data from other users stored at a data center and using this information in association with limited data from test individual to predict reliably. Such a centralised data storage implementation needs access to centrally stored data from a variety of users, in this case, pregnant women. This high performance is achieved at the expense of privacy sensitive information of users. Authors in [2] prove that such decentralised learning approach can help predict the gestational weight gain reliably with privacy preserved. However, the model aggregation in which local models are combined to form a single global model required storage of the local models eternally on the central server. This can lead to various forms of attacks on the models stored on the server or intercepted model updates including model inversion [9] and privacy leakage [10], [11]. In this work, we built upon our previous works and propose that such distributed learning can also be achieved by learning the global model incrementally without storing the local updates for infinite amount of time.

4.3 Data

We consider data from 80 women that were in their gestational week 5 or later recruited in Eindhoven, The Netherlands. The weight data was collected using a WiFi-connected weight scale, Withings WS30[1]. The participants were asked to log their weights weekly and the recorded weight data was sent to the cloud via a mobile application. Additional meta-data such as age, height and pre-pregnancy weight were also collected. The participants provided an informed consent pre-data collection and the study was approved by the Internal Ethics Committee for Biomedical Experiments of the involved organizations (ICBE Reference number 2015-0079). This sample dataset's distribution is close to that in [3], which is obtained from a large population of more than a million women, with almost half of the women gaining above the recommended guidelines [8].

[1] https://www.withings.com/

Table 4.1: Dataset description

Dataset Attribute	**Mean ± std**
Age (years)	31 ± 3.5
Height (meters)	1.69 ± 0.07
Pre-pregnancy weight (kgs)	69 ± 15
Pre-pregnancy BMI (kgs/m^2)	24 ± 4
Delivery (days)	277 ± 10
Weight Gained (kgs)	13.7 ± 4.7
Number of recorded weight gain samples	59.83 ± 41.02

4.4 Methods

Given a population of N subjects that acquired N time series of gestational weight gain measurements as $\mathcal{X} = \{(\mathbf{x}^1, \mathbf{y}^1), \cdots, (\mathbf{x}^N, \mathbf{y}^N)\}$, where $\mathbf{x}^i = \{t_1^i, t_2^i, t_3^i, \cdots, t_{m_i}^i\}$ represents the input gestational days upto delivery day $t_{m_i}^i$ and $\mathbf{y}^i = \{y_1^i, y_2^i, y_3^i, \cdots, y_{m_i}^i)\}$ represents the output weight gain for i^{th} subject, where $y_k^i = y(t_k^i)$. It is important to note here that t_k^i *does not necessarily equal* $t_k^j, i, j \in \{1, 2, \cdots, N\}$. This is because the data is *self-reported* such that each subject acquires measurements at different times according to their personal preferences and adherence to data collection.

Furthermore, we are given individual weight measurements from test subject's initial t_d^+ days of pregnancy data, $\mathcal{D} = \{(t_1^+, y_1^+), (t_2^+, y_2^+), \cdots, (t_d^+, y_d^+)\}$.

We try to learn function(s) f from $\mathcal{X}$ and $\mathcal{D}$, such that,

$$y^+ = f(t^+) + \epsilon \tag{4.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent and identically distributed (i.i.d) according to a Gaussian.

4.4.1 Centralised parametric approach

Traditionally used methods include parametric approach like fitting a p^{th}-order polynomial with $f = w_0 + w_1 t + w_2 t^2 + \cdots + w_p t^p$ in eq. (6.1) and estimating the coefficients $\mathbf{w} = [w_0, w_1, \cdots, w_p]^T$ by maximizing the likelihood ($\mathcal{L}$) over an individual's personal-training data $\mathcal{D}$, $\mathcal{L}(\mathbf{w}) = P(\mathcal{D}|\mathbf{w})$,

$$\hat{\mathbf{w}}_{MLE} = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{d} p(y_i^+|t_i^+; \mathbf{w}) \tag{4.2}$$

This can be done on a local device using only the estimates of a single user following eq. (4.2) that refers to the model learnt from the individual's sparse limited observations upto given t_d days. Often, such a prediction is far from reliable as it uses only few points from personal data. Authors in [8] show that such a prediction can be improved by considering the public-training data. The public-training data ($\mathcal{X}$) can be exploited and the maximum likelihood point estimates (MLE) of $\hat{\mathbf{w}}^i$ for each individual time series in the public-training data following eq. (4.2) can be derived. If we assume gaussianity over the distribution of $\mathbf{w}$ such that $\mathbf{w} \sim \mathcal{N}(\mu_{\hat{\mathbf{w}}}, \Sigma_{\hat{\mathbf{w}}})$, we can find a closed-form solution of $\mathbf{w}_{MAP}$ analytically. Here, $\mu_{\hat{\mathbf{w}}}^N = mean([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^N]^T)$, $\Sigma_{\hat{\mathbf{w}}}^N = cov([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^N]^T)$ are mean and covariances of the polynomial coefficients $\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^N$ that are each obtained using the individual gestational weight gain data from each of the N subjects in the public-training data. This distribution over the MLE estimates of the coefficients, $p(\mathbf{w})$ is acquired from the N subjects in the public-training data as an *a-priori* estimate. The likelihood learnt from the individual's personal-training data ($\mathcal{D}$) and the *a-priori* distribution learnt from the population data are then combined using bayes theorem to calculate the maximum-a-posteriori (MAP) estimate of the coefficients $p(\mathbf{w}|\mathcal{D})$.

$$\hat{\mathbf{w}}_{MAP} = \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, \frac{P(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{P(\mathcal{D})} \tag{4.3}$$

The forecast at time t_m^+ is given by $\hat{\mathbf{w}}_{MAP}[t_m^+\, {t_m^+}^2 \cdots {t_m^+}^p]^T$. This approach is called parametric because the choice of order of the polynomial p depends on the application of interest.

4.4.2 Federated approach with eternal updates (F_∞)

Federated learning is the process of storing only the model weights from individual subjects that are pushed to a central server. This preserves the privacy of a subject by only sending the model coefficients instead of complete raw data information as followed in the centralised approach. These small updates of local model coefficients ($\hat{\mathbf{w}}^i$) are sent to the central server where these updates are stored eternally, so that whenever a new model update arrives or a global update is needed all parties can participate and a global model can be aggregated as $\mu_{\hat{\mathbf{w}}} = mean([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^N]^T)$, $\Sigma_{\hat{\mathbf{w}}} = cov([\hat{\mathbf{w}}^1, \hat{\mathbf{w}}^2, \cdots, \hat{\mathbf{w}}^N]^T)$. The federated learning process (Fig. 4.1) that we utilised is as follows:-

(1) the centralized server sends the meta-data, (for example, order p of the polynomial, current global model estimate) to the participating subjects, once all subjects agree upon it,

(2) the local subjects estimate model coefficients $\hat{\mathbf{w}}_{MLE}$ based on maximising the likelihood of the local data,

(3) these local model updates are then shared to the server

(4) the server aggregates the individual models and create an updated global model,

(5) the global model is shared with the participating subjects.

This process is repeated as new subjects participate or the already participating subjects gather more data to push updated local models to the server. The local updates from participating subjects are stored eternally at the central server for secure aggregation to accurately estimate the global update. Hence, this method is also denoted as F_∞ as the updates are stored for infinite time.

4.4.3 Federated approach with ephemeral updates (F_{∞})

Although federated learning with eternal updates gives better privacy guarantees than sharing user data and learning on a central server, it

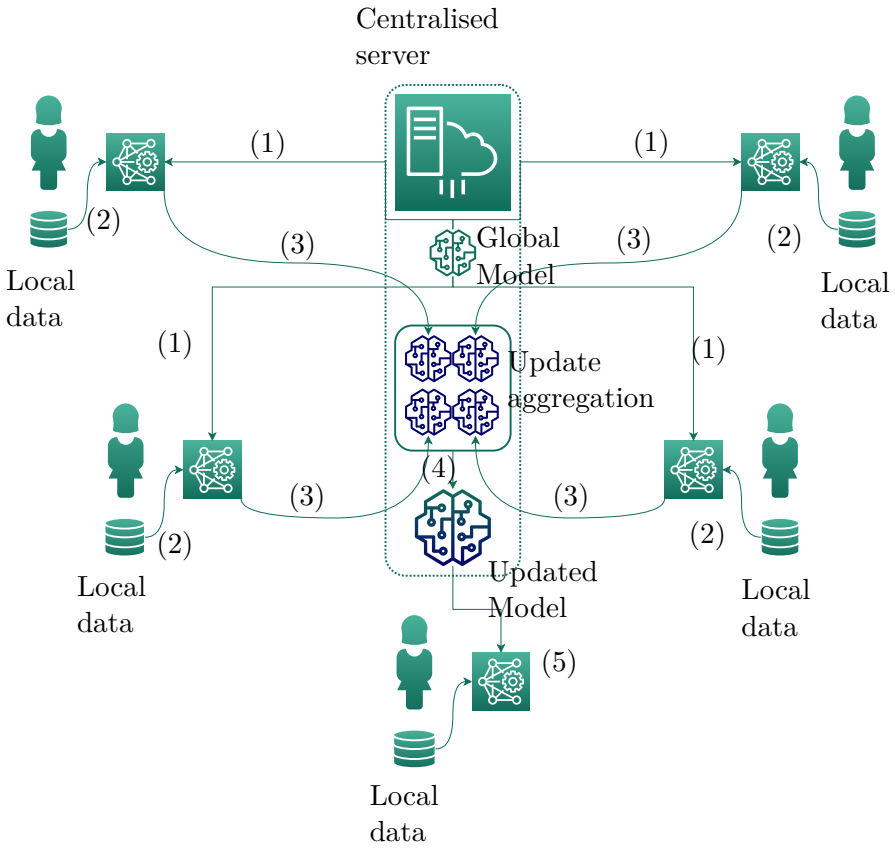


Figure 4.1: Federated learning ensures local data remains on-device and only model weights are shared at the central server.

still leaves the system vulnerable to attacks from older model updates or models themselves. The reason why stored model updates over time can still reveal sensitive information is because they are derived from the sensitive data of the user and is a representation of high level statistical distribution of the data[12]. We propose a scenario where only incremental updates from participating users are shared and are deleted from the central server once the global model is updated. Assuming a multivariate normal distribution, the global model μ^N, $\Sigma^N_{\hat{\mathbf{w}}}$ can be updated using the past global model μ^{N-1}, $\Sigma^{N-1}_{\hat{\mathbf{w}}}$ and the new shared

local model ($\hat{\mathbf{w}}^N$) as follows:-

$$\begin{aligned}\mu_{\hat{\mathbf{w}}}^N &= \frac{(N-1)\mu_{\hat{\mathbf{w}}}^{N-1} + \hat{\mathbf{w}}^N}{N} \\ &= \mu_{\hat{\mathbf{w}}}^{N-1} + \frac{\hat{\mathbf{w}}^N - \mu_{\hat{\mathbf{w}}}^{N-1}}{N}\end{aligned} \tag{4.4}$$

Similarly, covariance for N^{th} update can be estimated as [2],

$$\Sigma_{\hat{\mathbf{w}}}^N = \Sigma_{\hat{\mathbf{w}}}^{N-1} + \frac{\hat{\mathbf{w}}^N \hat{\mathbf{w}}^{N^\top}}{N-1} - \frac{N \cdot \mu_{\hat{\mathbf{w}}}^N \mu_{\hat{\mathbf{w}}}^{N^\top}}{N-1} + \mu_{\hat{\mathbf{w}}}^{N-1} \mu_{\hat{\mathbf{w}}}^{N-1^\top} \tag{4.5}$$

4.5 Experiments

We perform *leave-one-out* cross validation to evaluate and compare performance of our approaches, where training dataset in each iteration consists of weight gain data from $n \leq N$ public-training subjects and self-training data from the test subject. Here, n denotes the number of participants that had already participated in the federated learning pregnancy and an updated global model exists based on these n number of participants. We experiment with different values of n to show the effect of number of initiating users on the regression performance. We subtract the pre-pregnancy weight from the absolute data to get weight-gain data to ensure further local model security. The performance of regression was computed using Mean Absolute Error (MAE), $MAE = \frac{1}{N}\sum_N |y(t_m^i) - y_{ref}(t_m^i)|$. We experiment with first, second, third, fourth and fifth order polynomial based approach to fit our weight-*gain* data. The weight-gain data is normalised to pass through origin, so intercept term can be omitted. We chose third-order polynomial as it obtains minimum prediction error.

4.6 Results

Initially, we assume that $n = 10$ random users have already participated in the model building process and we perform leave-one-out cross validation

[2] See Appendix for proof.

on the rest of 70 subjects by sending a global model learnt based on $n = 10$ subjects as an initial estimate. Fig. 4.2(a) and 4.2(b) show the worst and best performing subjects respectively in terms of estimating end-of-pregnancy weight gain based on such a federated learning scheme. Note that personal weight gain data until day 120 is used which is shown

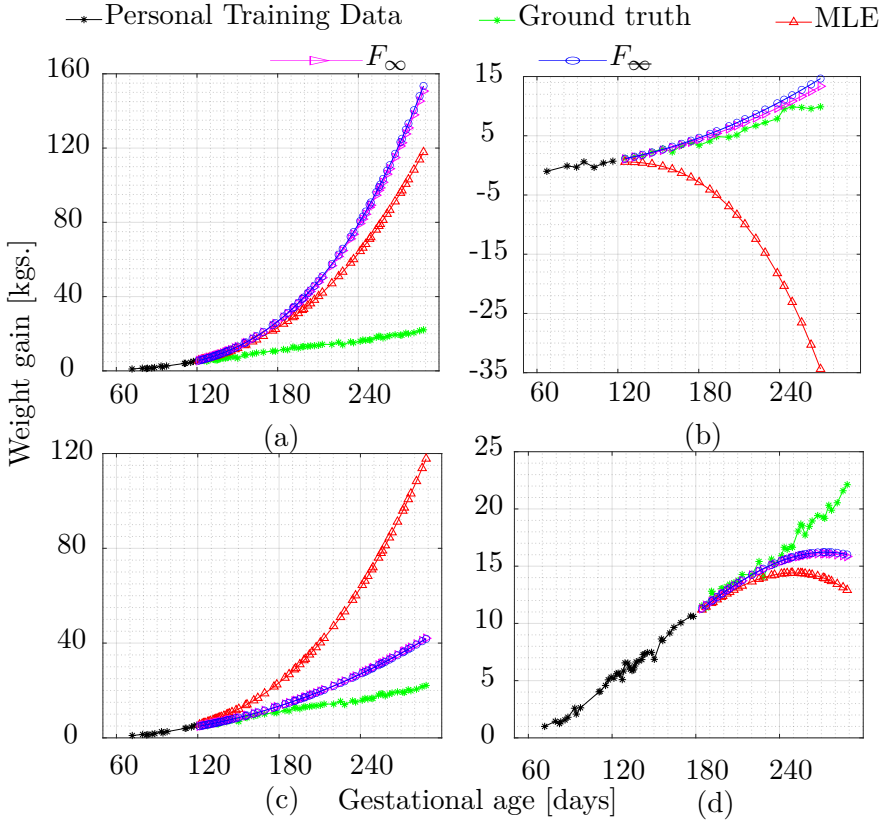


Figure 4.2: Federated learning generates (a) worst (subject id #14) and (b) best result (subject id #47) with limited personal data upto 120 days when only 10 users have participated initially. Performance for the subject id #14 can be seen improving when (c) 70 users participated in federated learning or when (d) the availability of personal-data increased (upto 180 days).

in black in Fig. 4.2 and the further values to be predicted are plotted in green. Fig. 4.2(c) shows that when a global model initiated by $n = 70$

users is distributed, the regression performance improves. The end-of-pregnancy weight prediction also improves with only $n = 10$ participating users if the personal-data availability increases (Fig. 4.2(d)).

Next, we present the prediction results averaged over $N - n$ subjects where n is varied as 10, 40, and 70 and $N = 80$. Fig. 4.3 shows that performance improves (MAE decreases) as self-training data availability increases or when the initial number of users participating in federated learning increase.

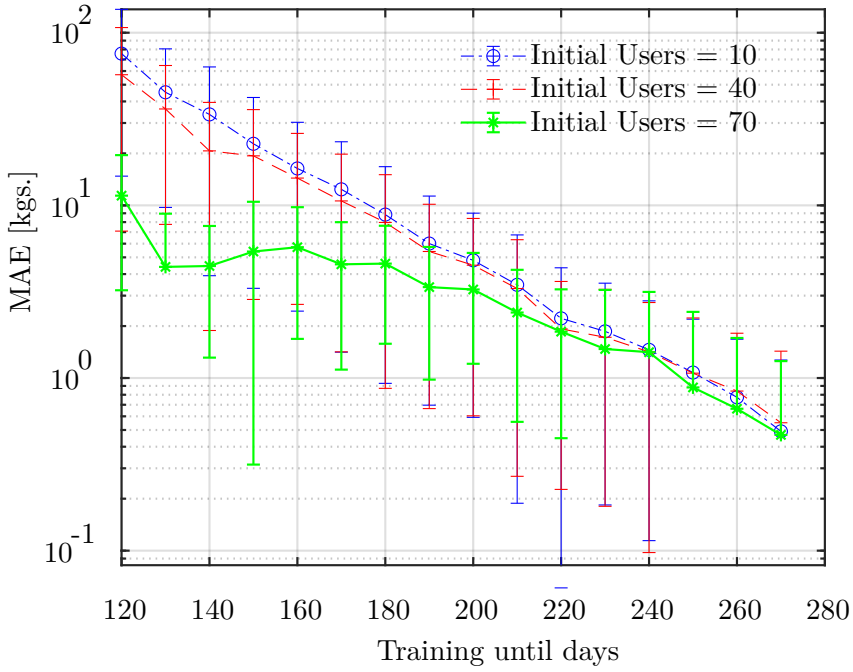


Figure 4.3: Average mean absolute error decreases as personal training data increases or number of initial users increase.

The centralised approach with 80 subjects produces the minimum absolute error in prediction with around 2.57 kgs error in predicting end-of-pregnancy weight gain. The federated approach with ephemeral updates (F_{∞}) performs worse by about 1.89 kgs than centralised MAP

approach with around 4.46 kgs mean absolute error. Fig. 4.4 shows that the federated learning out-performs the rest of the state-of-the-arts in predicting gestational weight gain in the presence of limited personal data (upto 200 days).

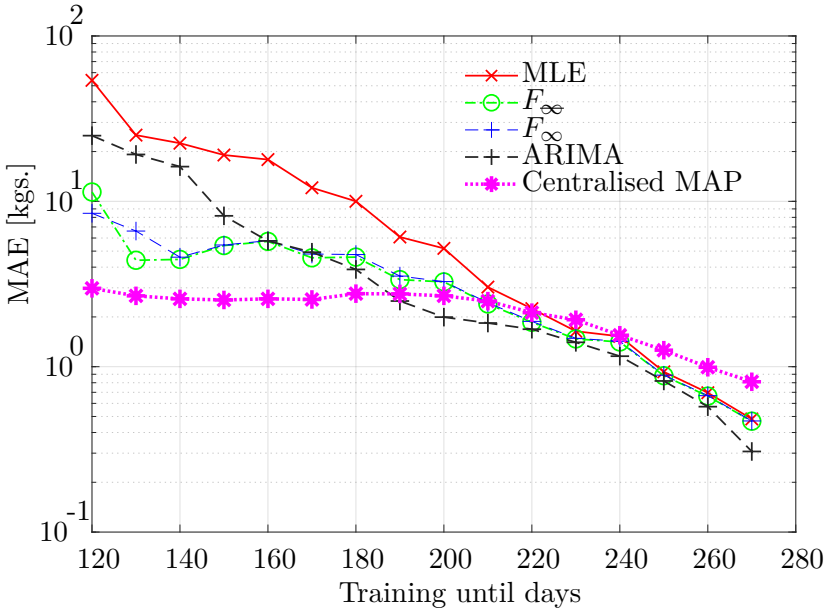


Figure 4.4: Performance of federated learning as compared to state-of-the-art approaches.

4.7 Discussion

In this paper, we propose the implementation of federated learning in time series related to healthcare. Apart from centralised learning where the raw data is shared and stored at the server for model building, we discuss two different federated scenarios that differ in how long the updates are stored at the central server.

Fig. 4.3 shows that the performance of the federated learning approach is different when the initial number of users participating in the training process varies. As the number of users that are involved in initial model

building increases, the performance improves. This can be attributed to the fact that the global model becomes more generalised when the number of users have increased. Similarly, a decreasing trend is observed in mean absolute error from Fig. 4.3 and Fig. 4.4 with respect to the number of training days available. It is intuitive that as more and more training data becomes available, the individual model starts estimating the end-of-the-pregnancy weight more accurately. But, it is desirable to predict the weight gain as early as possible for necessary intervention.

Fig. 4.4 shows that the performance of the two federated learning approaches with different local model storage strategies have identical performance as the availability of the training data increases. It can be observed that the federated approaches (F_{∞} (green) and F_{∞}(blue)) performance in early prediction of the weight gain is much better than the state-of-the-arts and is very close to the centralised approach, thus guaranteeing a good trade-off in performance and privacy preservation. As more and more training data for an individual pregnancy is available the performance of centralised approach, MLE and the federated learning approaches is close to each other as the global a model a-priori has less influence on local model.

4.8 Conclusion

In this paper, we try and propose a federated learning strategy that enables the preservation of privacy of a user while attaining state-of-the-art performance. We try and predict the gestational weight gain at the end of pregnancy as early as possible. The proposed approach achieves around 4.455 kgs of mean absolute error as early as 140 days into the pregnancy. In the future, we would like to improve upon the privacy of the shared model updates by making them differentially private (adding a noise to local weights) and establishing formal privacy guarantees.

APPENDIX

Proof of Federated Covariance estimation with ephemeral updates:

$$
\begin{aligned}
\Sigma_{\hat{\mathbf{w}}}^{N} &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\mathbf{w}}^{i}-\mu_{\hat{\mathbf{w}}}^{N}\right)\left(\hat{\mathbf{w}}^{i}-\mu_{\hat{\mathbf{w}}}^{N}\right)^{\top} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\hat{\mathbf{w}}^{i}\hat{\mathbf{w}}^{i^{\top}}-\hat{\mathbf{w}}^{i}\mu_{\hat{\mathbf{w}}}^{N^{\top}}-\mu_{\hat{\mathbf{w}}}^{N}\hat{\mathbf{w}}^{i^{\top}}+\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{\top}}\right] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\hat{\mathbf{w}}^{i}\hat{\mathbf{w}}^{i^{\top}}-2\left(\sum_{i=1}^{N}\hat{\mathbf{w}}^{i}\right)\mu_{\hat{\mathbf{w}}}^{N^{\top}}+\sum_{i=1}^{N}\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{\top}} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\hat{\mathbf{w}}^{i}\hat{\mathbf{w}}^{i^{\top}}-2N\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{\top}}+N\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{\top}} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\hat{\mathbf{w}}^{i}\hat{\mathbf{w}}^{i^{\top}}-N\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{\top}}
\end{aligned}
\tag{4.6}
$$

In order to calculate the update, we use the $\Delta\Sigma = \Sigma_{\hat{\mathbf{w}}}^{N} - \Sigma_{\hat{\mathbf{w}}}^{N-1}$. Substituting eq. 4.6 to calculate $\Delta\Sigma$, we get

$$
\begin{aligned}
\Delta\Sigma &= \Sigma_{\hat{\mathbf{w}}}^{N} - \Sigma_{\hat{\mathbf{w}}}^{N-1} \\
&= \frac{\hat{\mathbf{w}}^{N}\hat{\mathbf{w}}^{N^{T}}}{N-1} - \frac{N\cdot\mu_{\hat{\mathbf{w}}}^{N}\mu_{\hat{\mathbf{w}}}^{N^{T}}}{N-1} + \mu_{\hat{\mathbf{w}}}^{N-1}\mu_{\hat{\mathbf{w}}}^{N-1^{T}}
\end{aligned}
\tag{4.7}
$$

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

Bibliography

[1] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Gestational weight gain prediction using privacy preserving federated learning", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 2170–2174.

[2] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 398–399.

[3] R. F. Goldstein, S. K. Abell, S. Ranasinha, M. Misso, J. A. Boyle, M. H. Black, N. Li, G. Hu, F. Corrado, L. Rode, *et al.*, "Association of gestational weight gain with maternal and infant outcomes: A systematic review and meta-analysis", *Jama*, vol. 317, no. 21, pp. 2207–2225, 2017.

[4] K. M. Rasmussen, P. M. Catalano, and A. L. Yaktine, "New guidelines for weight gain during pregnancy: What obstetrician/gynecologists should know", *Current opinion in obstetrics & gynecology*, vol. 21, no. 6, p. 521, 2009.

[5] R. Gaillard, B. Durmuş, A. Hofman, J. P. Mackenbach, E. A. Steegers, and V. W. Jaddoe, "Risk factors and outcomes of maternal obesity and excessive weight gain during pregnancy", *Obesity*, vol. 21, no. 5, pp. 1046–1055, 2013.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data", in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[7] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[8] C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278.

[9] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[10] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning", in *2019 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2019, pp. 691–706.

[11] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning", *arXiv preprint arXiv:1812.00984*, 2018.

[12] L. Lyu, "Privacy-preserving machine learning and data aggregation for internet of things", Ph.D. dissertation, 2018.

Chapter 5

Feature Selection for Handling Missing Data

This chapter was previously published as:
C. Puri, G. Kooijman, X. Long, P. Hamelmann, S. Asvadi, B. Vanrumste, and S. Luca, "Feature selection for unbiased imputation of missing values: A case study in healthcare", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 1911–1915

Abstract

Datasets in healthcare are plagued with incomplete information. Imputation is a common method to deal with missing data where the basic idea is to substitute some reasonable guess for each missing value and then continue with the analysis as if there were no missing data. However unbiased predictions based on imputed datasets can only be guaranteed when the missing mechanism is completely independent of the observed or missing data. Often, this promise is broken in healthcare dataset acquisition due to unintentional errors or response bias of the interviewees. We highlight this issue by studying extensively on an

annual health survey dataset on infant mortality prediction and provide a systematic testing for such assumption. We identify such biased features using an empirical approach and show the impact of wrongful inclusion of these features on the predictive performance.
Clinical relevance - We show that blind analysis along with plug and play imputation of healthcare data is a potential pitfall that clinicians and researchers want to avoid in finding important markers of disease.

5.1 Introduction

Missing data is a ubiquitous problem in statistical analysis or data science irrespective of the domain, be it social sciences or health sciences. Most, if not all the machine learning algorithms presume that all the information is present for all the available features. Conventional techniques of handling missing data include performing complete case analysis which is deletion of missing cases but this strategy results in lesser informative subset of the dataset.

Health data are being massively generated due to the advancement of both data acquisition and analysis technologies, examples of which include time-series data from intensive care units (ICU), biomarker data, electronic health records (EHR), or health surveys. The global market for big data in health care has been projected to grow significantly from US$19.6 billion in 2018 to US$ 47.7 billion in 2022 [2]. Undoubtedly, this rise is due to the penetration of data analytics for better predictive clinical outcomes, analyzing disease, and tracking patterns thus increasing overall public health. Modelling such large scale data and predicting the health status for improvement of the patient is challenging. One such challenge is addressing missing values in data, that arise, for example, from unrecorded data from ICU machines due to lead detachment or respondents intentional/unintentional non-responsiveness to health surveys [3].

Datasets (particularly in healthcare) are often preprocessed by various imputation techniques that rely on the assumption of independence between the missing mechanism and the observed data. Statistical tests

to verify this assumption often fail when missing data is abundant and a subset of reasonable size of complete data is absent [4].

In this article, we illustrate the common pitfall of blindly applying imputation techniques that can lead to biased results. To this end, we utilize a publicly available dataset from the annual health survey in India and show how state-of-the art imputation techniques fall short in reliable feature matrix completion for classification purposes. Furthermore, we propose an empirical approach to study the effect of including features that are strongly associated to the occurrence of missing data.

A large part of existing literature on missing data analysis that we discuss later studies one or more methods to impute data. In this article, we highlight the biased effect that imputation might have on the results of a predictive classification model in the presence of imbalanced missingness across different classes and we propose a method that can support in preventing careless imputation of missing data.

The remainder of the paper is structured as follows. In section 5.2, we talk about the imputation techniques and types of missingness. Further, we describe the dataset in section 5.3. Section 5.4 elaborates upon the experiments performed in order to show the impact of the described challenges with unbiased missing values imputation. We conclude by giving final remarks to the reader in section 5.5.

5.2 Related Work

Several approaches exist that handle missing data by (a) *deletion* of the cases that have values missing for a single variable, simply excluding such cases can be used to build complete datasets [5] or (b) estimating a single set of missing values by *single imputation* using statistical moments, k-nearest neighbours or (c) a confidence interval imputation by much more complex *multiple imputation* [6]. A specific implementation of multiple imputation strategy known as the Multivariate Imputation by Chained Equations (MICE) involves multiple steps of imputation in which every variable is imputed conditionally on all other variables [7]. Deletion based imputation can lead to loss of statistical power and can introduce

bias when a smaller complete subset is selected from a non-complete dataset.

Based on the type of missingness, three basic mechanisms are present [5], described as follows, Suppose we have missing data on a variable Y and we have some other variable X, then, one defines:

- Missing completely at random (MCAR) : If the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set, the data is said to be MCAR.
- Missing at random (MAR): If the missingness depends only on the data that are observed but not on the missing components, the data are MAR. i.e., $P(Y\text{missing}|Y, X) = P(Y\text{missing}|X)$
- Not missing at random (NMAR): If the probability that Y is missing depends on the unobserved value of Y itself, then the mechanism is NMAR.

Most of the imputation strategies work under the assumption that the missingness is MCAR [4]. Statistical tests like Little's test [4] exist that can test whether the data is MCAR or not. However, in the absence of a small complete subset (when missing data is abundant), it is difficult to conduct such a test and existing imputation techniques tend to fail in reliably predicting the missing values. Authors in [8] and [9] discuss different imputation methods and compare the performance of imputation techniques with different amount of missingness on different datasets. They advise that different missing data mechanism needs different imputation strategy, however none of the previous works talk about the imbalance in missingness that can be present in different classes when considering a classification problem. Imputation without analysis of such an imbalance can lead to erroneous completion of the feature matrix which we will show later.

In this article, we illustrate the challenges of using imputation methods when the MCAR assumption is not met. For this purpose, we use a case study from healthcare and we propose an algorithm to study the effect of including features that are strongly associated to the occurrence of missing data.

5.3 Data

We chose a publicly available healthcare survey dataset conducted over women that underwent pregnancy in several states in India [10]. Child mortality remains a major challenge in India and is responsible for approximately 39.1 deaths per 1,000 live births in 2017 [11]. Child mortality as a pregnancy outcome is considered a major attribute in building efforts to preventive antenatal care thus reducing infant mortality. Poor pregnancy outcome in India is not just attributed in defining the outcome but is also a consequence of substandard health information systems. The National Institute for Medical Statistics of the Indian Council of Medical Research (ICMR - NIMS) has launched the National Data Quality Forum (NQDF) in collaboration with the Population Council. The purpose of the NQDF is establishing protocols and good practices for betterment of data collection, storage and dissemination[12]. Major barriers to the data quality include (a) lack of comparability, (b) discordance between system and survey level estimates, (c) lengthy questionnaires, (d) questions related to socially restricted conversation topics, (e) age-reporting errors or non-response, (f) intentional skipping of questions, (g) under-reporting due to subjective question interpretation and incompleteness, and (h) paucity of data to generate reliable estimates on mortality [12]. We select data from the open government platform in India where the Indian government has provided open access to datasets, documents, etc. for public use. This dataset is also collected as part of a joint initiative between government of India and US government. Authors in [13] have shown the risks of using such open datasets from non-verified sources such as [14]. They identify that Woman Schedule Section 1 and Section 2 (called WPS dataset) is from a verified source [10]. A number of 355 features in the WPS dataset [10] are present in the form of questionnaire, with fields related to social, economic, health status or demographic indicators as well as the outcome of pregnancy (live or stillbirth).

Since the dataset consists of questions from surveys, some questions are explicitly on the child birth outcome thus making some of the features highly correlated with the fact whether the child birth resulted in a live or stillbirth. Hence, features such as baby weight taken or not,

weight measurement, immunization card details, different vaccines, polio, hepatitis, vit. A, IFA tablet, feeding details, breastfed, animal dairy, solid food month, etc. were removed to maintain causality of the labels with respect to the feature set because these features can only be recorded if the pregnancy outcome is positive. We then selected 233 features out of 355 as the final feature list for further analysis.

5.4 The case study

Given a feature matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x} \in \mathbb{R}^d$ observed for N subjects, the objective is to learn a function $h : \mathbf{X} \rightarrow Y$, where $Y = \{0, 1\}$ corresponds to prediction of still or live birth respectively. The class of stillbirth also includes all cases of induced abortion and spontaneous abortion. The number of cases for live birth are much more than all the stillbirth cases. Hence, we look at the problem of learning a model for binary classification of live and stillbirth.

Imputation of the feature matrix occurs during pre-processing before training the model, as shown in Figure 5.1.

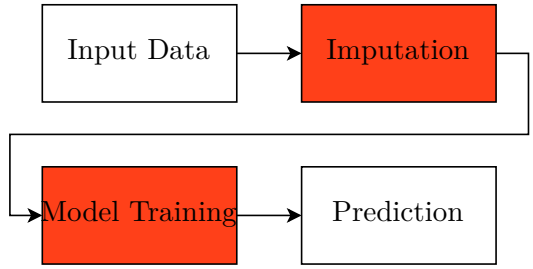


Figure 5.1: Typical processing pipeline for learning with missing data.

We compare the performance of the imputation approach by keeping the processing pipeline fixed i.e the training data and the classifier and its parameters are fixed and only the imputation approach is varied. For our experiments, we perform a 10-fold cross validation with minority class as the positive class (stillbirth) and plot the average receiver operating characteristic. Figure 5.2 shows that a random forest classifier with single

imputation methods like constant based filling for imputation achieves the best performance. This motivated us to look closely into the features and the missingness in relation to the class label.

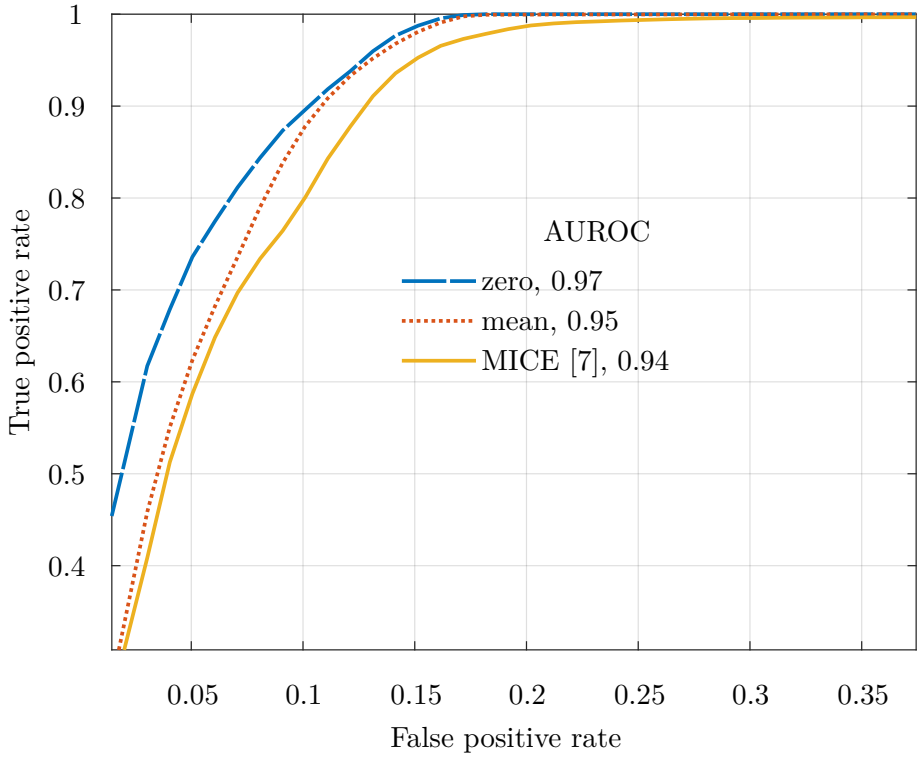


Figure 5.2: A simple zero based constant filling appears to have the most predictive power (keeping classifier and its parameters fixed) when the imputation methods are applied blindly without understanding the type of missingnes.[1]

We take two exemplary features that are discrete-valued categorical features namely "*source_of_anc*" and "*maternity_financial_assistance*". In the annual health survey, "*source_of_anc*" refers to the institution offering antenatal care (ANC). 12 different government or private institutions operating at different governance level are assigned real

[1]Notice the difference in x and y coordinates as this is a zoomed-in snippet of the AUROC curve to improve the visibility of the curves.

numbers. For example, women receiving antenatal care at government operated rural center called *anganwadi* are assigned the real number '1'. Similarly, women receiving ANC from private hospitals are assigned the value '9'. The complete description of the domain space is mentioned in [10] and is mapped to $\mathbb{R}$ in $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 99\}$. For the feature "*maternity_financial_assistance*", women who took financial assistance under the government scheme *Janani Suraksha Yojana (JSY)* are assigned the value '1' for this feature. If they avail any other government scheme, real number '2' is assigned, '3' for any other non-government scheme and '4' in case no financial assistance was availed. The domain space for this feature is mapped in $\mathbb{R}$ to $\{1, 2, 3, 4\}$. Figure. 5.3(a) and (b) represent the feature "*source_of_anc*" and "*maternity_financial_assistance*" respectively. These two features are representative for multiple features which have a lot of missing values or are filled with zero in the questionnaire, possibly due to errors in the interview. For the sake of discrimination, we do not combine the missing values and zero-entries even if they mean the same thing.

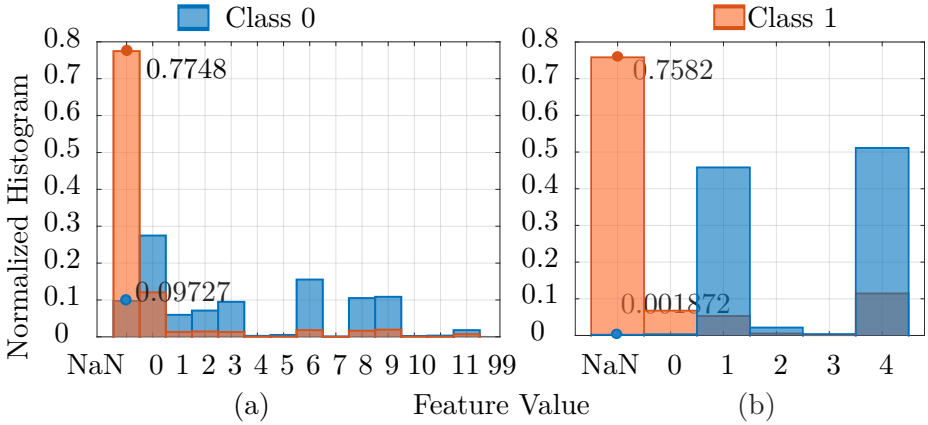


Figure 5.3: Exempelary features (a)"*source_of_anc*" and (b) "*maternity_financial_assistance*" with different classwise imbalance in terms of availability of the data, Class 1 = live birth, Class 0 = stillbirth

As can be observed from Figure. 5.3(a) 9.7% data is missing in class "0" and 77.48% data is missing for class "1" for the feature "*source_of_anc*". Similarly from Figure. 5.3(b), "*maternity_financial_assistance*" feature

has around 0.187% data missing for class "0" and 75.82% data is missing for class "1". This percentage imbalance in missing data will be further irritated if we consider the occurrence of zero in the data as '0' is not in the domain space of most of the features and was recorded maybe as a missing value. Suppose we fill the missing data with a simple single imputation approach, for example, a constant '$c \in \mathbb{R}$' or mean, for feature "*source_of_anc*", then for 77.48% of the data in class '1' the feature value will be c and the remaining 22.52% will take values somewhere in the domain of the feature $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 99\}$. On the other hand, only 9.7% of the data is missing for class '0' and will be assigned the value c. Rest of the class '0' (90.263%) will take values from the domain of the feature. Since class "1" has more missing data than class "0", constant-filling based imputation methods will provide a false sense of discriminatory power to the feature. The institution that provided antenatal care to the mother is indicated by the feature "*source_of_anc*". This feature is less frequently documented in live birth instances (77.48% is missing) than when the baby is stillbirth (22.52% is missing), which causes a prominent peak (in red) in the distributions of the two classes at the value "NaN" (the missing value). When corrected with basic imputation techniques, this mismatch in the feature's recording by the interviewer throughout the data collecting process gives the feature discriminatory power. However, it does not accurately reflect the distribution of the feature between the two classes.

Figure. 5.4 represents a compact view of all the features plotted with respect to availability of data in each class. All the features that exhibit classwise-imbalance in availability of feature data are shown in $*$. The line $y = x$ in Figure 5.4 represents the features that have equal amounts of missing data in each class (marked in $\circ$). The margins along the line $y = x$ represent the tolerance level (e.g. = 10% tolerance) for visualising whether the feature is useful or not in the absence of actual feature value.

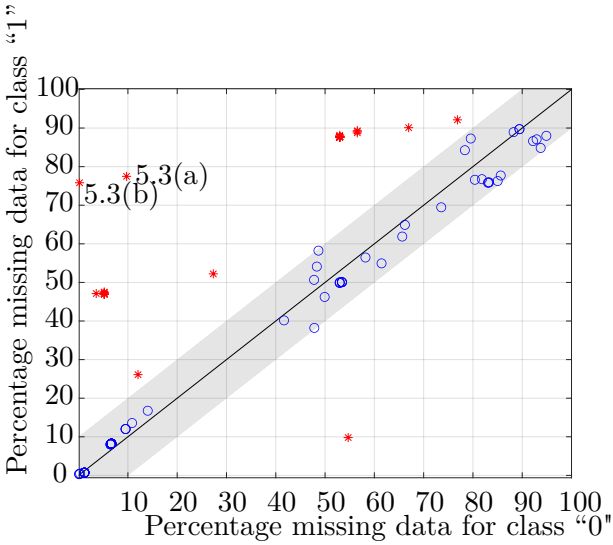


Figure 5.4: Each data point ($*$, $\circ$) represents a feature with x and y coordinates being the missing percentage in class 0 and 1 respectively. Each feature outside the tolerance margins (marked as $*$) have high absolute percentage difference between the available class "0" and class "1". As depicted, features from Figure. 5.3(a and b) are also apparently intolerable features

One way of finding out if the features are missing completely at random is by performing Little's test [15]. We found on performing Little's test that the data is not missing completely at random.

We develop an empirical approach to evaluate the features that exhibit such behaviour and use the algorithm provided in Algorithm 2. The algorithm first, calculates the percentage missing data in each class. If the difference in percentage of the missing data calculated in the previous step differs by a pre-decided tolerable limit, then we say that the feature is a tolerable feature with respect to the imbalance in missing data, otherwise, it is an intolerable feature. For example, as can be observed from Figure 5.3a and b, both the features have an absolute difference of > 60 which is greater than a pre-decided tolerance limit of 10, decided empirically. Hence, both the features are intolerable and have false discriminatory power for model-learning if used with imputation. We test with different values of tolerance thresholds to test the variation of

Algorithm 2 Finding features inside tolerable range

procedure FIND TOLERABLE FEATURES
  $Input : \mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, $\mathbf{x} \in \mathbb{R}^d$, $Y = \{0, 1\}$
  $Parameter : perThresh \in [0, 100]$ ▷ tolerance (in %)
  $mis0 = 0$ ▷ Initiate missing count for class 0
  $mis1 = 0$ ▷ Initiate missing count for class 1
  $tolerableFeatInd = []$
  for $i = 1 : d$ **do**
    $f0 = \mathbf{x}_i(Y == 0)$
    $f1 = \mathbf{x}_i(Y == 1)$
    for $j = 1 : length(f0)$ **do**
      if $isnan(f0(j))$ **then**
        $mis0 = mis0 + 1$
    for $j = 1 : length(f1)$ **do**
      if $isnan(f1(j))$ **then**
        $mis1 = mis1 + 1$
    $misPer0 = 100 * mis0/length(f0)$
    $missPer1 = 100 * mis1/length(f1)$
    $absDiffMiss = abs(missPer0 - missPer1)$
    if $(absDiffMiss < perThresh)$ **then**
      $tolerableFeatInd = [tolerableFeatInd, i]$

performance if such erroneous features are included in model-building blindly. Figure. 5.5 represents the classification of live-stillbirth prediction performance with different imputation strategies and different tolerance thresholds (margin as depicted in Figure. 5.3) as described in algorithm 2.

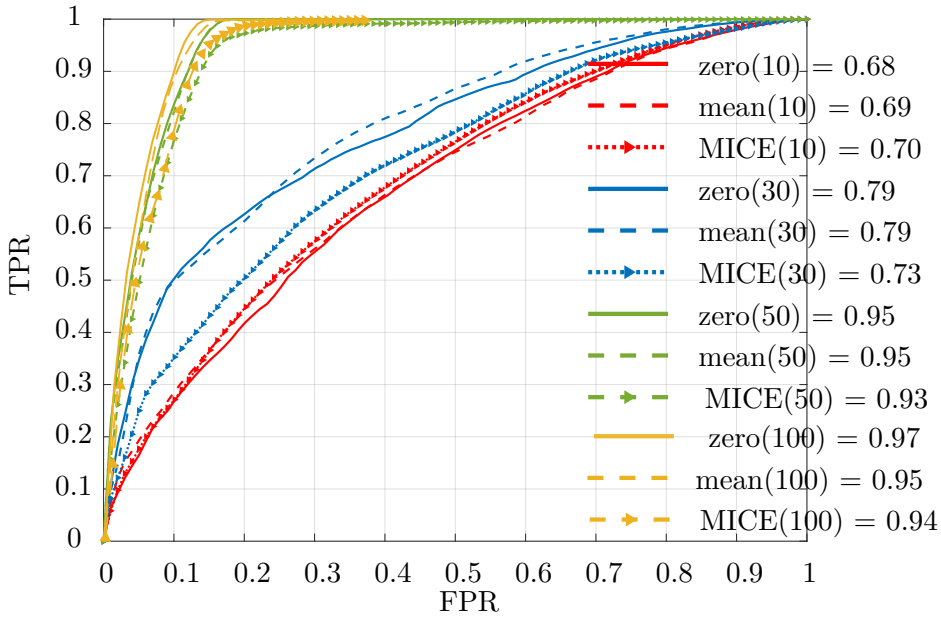


Figure 5.5: Classification performance with zero, mean-filling and MICE based imputation when tolerance threshold varies from [10, 30, 50, 100] and the area under the ROC curve represented upto two decimal places.

A number of 86, 90, 117 and 233 features were selected based on tolerance thresholds 10, 30, 50 and 100 respectively. Figure 5.5 shows that we get much higher performance when the tolerance threshold is set high. This is due to the fact that at high tolerance threshold we include more features that are biased because of the imbalance in missingness in different classes. For example, when tolerance is set to maximum (i.e. = 100), all the 233 features are included in training and the performance is the same as shown in Fig. 5.2. However, when the tolerance threshold is as low as 10, we include less biased features (depicted as ○ in Figure 5.4). Here, the final performance achieved with tolerance level 10 is around 0.68. We also observe that at the threshold of 10, where minimum number of biased features are included, the state-of-the-art MICE approach performs better than the constant-filling approaches.

5.5 Conclusion

This paper reflects on the need for caution when imputing missing values for classification. The assumptions such as MCAR or MAR are not always easy to verify. Most of the state-of-the-art imputation techniques work well when data is MCAR and a subset of complete data is present for guiding the imputation process. We showed the effect of imputation on the performance by studying a case in healthcare. It was evident from our experiments that attention was needed when features were used with missing values that are strongly associated with the class label and including these in a predictive model can lead to a false sense of discriminatory power. In the future, we would like to develop methods to find the tolerance threshold and fill the missing data in an unbiased manner.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

Bibliography

[1] C. Puri, G. Kooijman, X. Long, P. Hamelmann, S. Asvadi, B. Vanrumste, and S. Luca, "Feature selection for unbiased imputation of missing values: A case study in healthcare", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 1911–1915.

[2] R. ltd and Markets, *Healthcare big data analytics market: Global industry trends, share, size, growth, opportunity and forecast 2019-2024*, `https://www.researchandmarkets.com/reports/4856240/healthcare-big-data-analytics-market-global`. (visited on 12/01/2020).

[3] N. R. Council *et al.*, *Nonresponse in social science surveys: A research agenda*. National Academies Press, 2013.

[4] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[5] P. D. Allison, *Missing data*. Sage publications, 2001, vol. 136.

[6] J. S. Murray *et al.*, "Multiple imputation: A review of practical and theoretical findings", *Statistical Science*, vol. 33, no. 2, pp. 142–159, 2018.

[7] S. v. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R", *Journal of statistical software*, pp. 1–68, 2010.

[8] L. A. Hunt, "Missing data imputation and its effect on the accuracy of classification", in *Data Science*, Springer, 2017, pp. 3–14.

[9] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data", *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[10] *Census of india : Annual health survey 2010 - 11 fact sheet*, https://www.censusindia.gov.in/vital_statistics/AHSBulletins/Factsheets.html. (visited on 02/11/2021).

[11] *India demographics profile*, en, https://www.indexmundi.com/india/demographics_profile.html. (visited on 03/08/2021).

[12] InsightsIAS, *National data quality forum(ndqf)*, en-US, Jul. 2019. [Online]. Available: https://www.insightsonindia.com/2019/07/26/national-data-quality-forum-ndqf/ (visited on 03/08/2021).

[13] A. Trivedi, S. Mukherjee, E. Tse, A. Ewing, and J. L. Ferres, "Risks of using non-verified open data: A case study on using machine learning techniques for predicting pregnancy outcomes in india", *Proceedings of NeurIPS 2019 Workshop on Machine Learning for the Developing World: Challenges and Risks of ML4D, arXiv preprint arXiv:1910.02136*, 2020. arXiv: 2001.00249 [cs.CY].

[14] *Predict outcome of pregnancy*, en, https://kaggle.com/rajanand/ahs-woman-1. (visited on 02/10/2021).

[15] R. J. Little, "A test of missing completely at random for multivariate data with missing values", *Journal of the American statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

Chapter 6

Pain Estimation in Workplace

This chapter is accepted as:
C. Puri, S. Keyaerts, M. Szymanski, L. Godderis, K. Verbert, S. Luca, and B. Vanrumste, "Daily pain prediction in workplace using gaussian processes", *HEALTHINF*, 2023 (accepted).

Abstract

Work-related Musculoskeletal disorders (MSDs) account for 60% of sickness-related absences and even permanent inability to work in the Europe. Long term impacts of MSDs include "Pain chronification" which is the transition of temporary pain into persistent pain. Preventive pain management can lower the risk of chronic pain. It is therefore important to appropriately assess pain in advance, which can assist a person in improving their fear of returning to work. In this study, we analysed pain data acquired over time by a smartphone application from a number of participants. We attempt to forecast a person's future pain levels based on his or her prior pain data. Due to the self-reported nature of the data, modelling daily pain is challenging due to the large number of missing values. For pain prediction modelling of a test subject, we employ a subset selection strategy that dynamically selects a closest subset of individuals from the training data. The similarity between the

test subject and the training subjects is determined via dynamic time warping-based dissimilarity measure based on the time limited historical data until a given point in time. The pain trends of these selected subset subjects is more similar to that of the individual of interest. Then, we employ a Gaussian processes regression model for modelling the pain. We empirically test our model using a leave-one-subject-out cross validation to attain 20% improvement over state-of-the-art results in early prediction of pain.

6.1 Introduction

Musculoskeletal disorders (MSD) are presently a widespread type of work-related health problem and a leading cause for absenteeism from work across all sectors and occupations. Around 60% of all the health related problems in Europe (EU) are work-related MSDs that account for 60% of sickness related absences and even permanent inability to work [2]. This creates a financial burden on individuals, businesses, and society [3]. Prevention of MSDs from the outset of a person's career will allow for an extended work life and better job satisfaction [4]. MSD prevention can also address the long-term implications of demographic ageing, as outlined in the objectives of the Europe 2020 strategy for smart, sustainable, and inclusive growth. Consequently, MSDs are not only an occupational burden, but also a public health and societal challenge [3].

Long-term impacts of MSDs include "Pain chronification", which is the transformation of transient pain into permanent pain as a result of recurrent physical strain sustained while doing work-related activities [5]. Other than physical pain experience, there is vast amount of evidence on the importance of pain coping strategies, cognitive appraisals (e.g. catastrophizing, high threat values, and fear-avoidance believes), negative emotions and expectations [6], [7], [8]. These factors influence how sensory information is processed in the spinal cord and the brain. There are several models that integrate different biopsychosocial factors to the perception of pain such as the fear-avoidance model [9], avoidance-endurance model [10], and the common sense model [11]. These models illustrate how various persons experience pain, which results in the

subjectivity of pain assessments. Thus, predicting pain in a personalised manner for early intervention is essential for preventing pain persistence. It is crucial that both patients and medical practitioners have the education and abilities necessary to manage pain correctly [5].

Currently, pain management is done based on the initial patient evaluation (history, physical examination) which is followed by prompt treatment based on the level of the pain [5]. This is especially true for the acute stages of pain. In cases of chronic pain, evidence suggests that therapies should be directed less by current pain levels and more by participation in valued activities despite discomfort [12]. Therefore, appropriately measuring pain in early stages can aid in pain management by evaluating medication efficacy, comprehending the complicated relationship between pain and personal/contextual factors, and preparing patients and healthcare providers for a challenging period with flare-ups. Preventative pain management can also reduce the likelihood of it becoming chronic. Multiple pain management applications exists but are only limited to maintaining logs of the level/intensity of the pain [13]. However, for more successful pain management, it is essential to accurately estimate the pain in advance, preferably several days ahead, which can help a person moderate his or her expectations and anxieties about returning to work. Additionally, pain prediction can provide healthcare practitioners with a better understanding of the required treatment and assist with individualised planning.

This study attempts to forecast the pain experienced by workers from various industries several days ahead based on the daily recorded history of pain. To this end, workers were asked to record their daily levels of satisfaction and pain in a smartphone application on a scale from 0 (no pain) to 100 (worst possible pain). Then, we attempt to forecast future pain levels by modelling individual pain levels recorded until a certain day. This is technically difficult because the data is self-reported, and app users did not always indicate their pain levels on a daily basis, resulting in missing data. Fig. 6.1 shows how an individual's pain data looks like with respect to time. With the study data of over 300 days, more than 70% of the participants lack 66% of the daily data (< 100 entries).

The lack of observations in an individual's time series data restricts

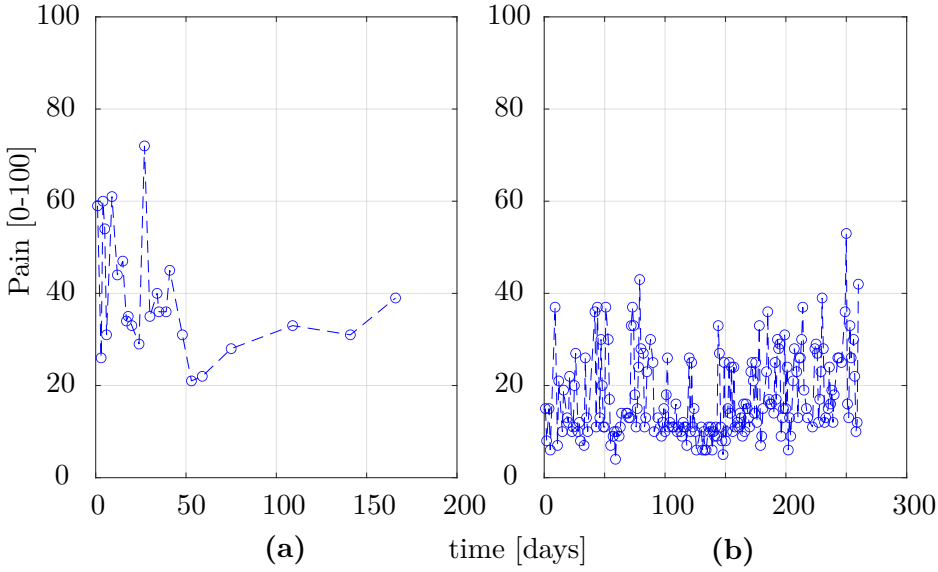


Figure 6.1: Two distinct users documented their pain levels over a 300-day period using (a) as few as 22 samples and (b) as many as 228 samples.

the application of traditional time series modelling techniques such as the autoregressive integrated moving average model (ARIMA), which requires uniformly sampled data [14].

Authors in [15] provide an extensive survey of the application of modern machine learning techniques for the estimation or detection of pain. The majority of the pain experiences discussed in the literature [15] are related to a hospital or post-operative scenario, rather than persistent workplace-related pain. Furthermore, pain forecasting models that use machine learning are built on clinical data (e.g. drugs administered, patient comorbidity data) collected during pain experiences post a surgical operation, enriching the information available for modelling [16], [17].

Deep learning (DL) is a branch of machine learning that, when given massive volumes of data, may automatically learn representations from raw data to achieve a specific objective, such as classification or regression [18], [19], [20], [21]. DL has been used in multiple healthcare related applications that can predict the health of an individual from the time series data. For example, detecting cardiac abnormality [22] or forecasting

glucose levels [23] in individuals. Works such as [24], [25] developed deep learning techniques that can address the non-uniformly sampled time-series data when a large training dataset is available. Traditional machine learning strategies, however, outperform deep learning strategies when training data is insufficient [26].

In this work, we follow the method proposed in [27], where a subset of the training data is selected followed by learning a regression model based on Gaussian processes (GP). Here, we would like to showcase the efficacy of the subset selection approach followed by GP based regression to model an individual's pain measurements until a certain day. The subset selection approach works by first selecting individuals from training data that resemble closely the progression of pain over time to that of the target individual. The number of individuals to be selected is chosen dynamically based on the similarities across individuals. The dynamically chosen subset along with the avaiable data from the target subject is then used to train a regression model for improved prediction performance. However, directly applying the method of [27] doesn't give the best results owing to the subjective nature of the pain measurements. Hence, we add a pre-processing treatment of the data prior to subset selection and learning the GP model. We explain the need to do so as follows,

1. Pain measurements of an individual vary a lot across time. This might be because of the pain persistence over time or by the number of individual days with more stress resulting in more pain. Hence, unlike a general increasing trend in gestational weight gain [27], it is difficult to find a pattern in the pain measurements over time. Thus, there are anomalous instances in the pain measurements that can result in an inaccurate general model.

2. Pain measurements are self-reported and are highly subjective in nature. This means that individuals have certain biases to only rate their pain (scored between [0-100]) around a fixed baseline, e.g., a person with a baseline *reported* pain of 20 will seldom report a pain of 80. Thus, scaling individual pain measurements for modelling is a necessary step.

The objective of this work is to study if:

- It is possible to estimate an individual's pain levels from a small number of non-uniformly collected historical pain measurements.
- Given the subjective nature of pain data, is it possible to use previous pain measurements of other individuals in a training dataset to enhance pain prediction?

The main contributions of this paper are:

1. We develop models of daily pain data to forecast and manage pain level trends over time.
2. We propose a two-step pre-processing strategy to enhance pain prediction modelling. This is accomplished by smoothing the pain time series in training data and self-normalising the target individual's pain data with the few measurements provided.
3. We use a subset-selection strategy to generate the most informative subset of training data for a given target individual. Individuals in this closest subset exhibit similar pain trends to the individual of interest.
4. We devise modelling based on the selected subset using Gaussian processes for *multi-step* forecasting of pain up to n-days ahead in time.

The dataset is described in section 6.2, followed by the proposed methodology in section 6.3. In section 6.4, we describe the experiments conducted to generate the results. Results and their implications are discussed in greater detail in section 6.5, and concluding remarks are presented in section 6.6. Section 6.7 concludes the paper by discussing potential future directions and constraints.

6.2 Data

In this study, 340 participants were recruited from various work sectors. At the start of the study, participants were asked about different

work-related factors, their pain complaints, pain-related perceptions, coping strategies, and other contextual factors (indeed, such as physical activity and time spent sitting). From January 2021 to May 2022, they were required to maintain a daily journal in which they recorded their overall pain levels, mood (not with yes/no questions), stress levels, and satisfaction along with baseline questions such as age, gender, height, weight, and industry of employment. The pain levels were recorded on a scale of 0 (best) to 100 (worst) using an mHealth smartphone application.[1]. Questions related to mood (sad, angry, happy, fatigued, cheerful) were also part of the daily journal.

190 participants were excluded because they did not record daily pain values at all. In addition, 51 more individuals were removed based on the criterion of not having more than 10 daily pain values recorded, with more than 2 values separated by 1 week. The remaining 99 participants' data were used to develop pain prediction models. Fig. 6.2 presents the gender-wise distribution of participants in different industries.

This study was conducted within the context of the Personal Health Empowerment project, which focused on investigating and developing new monitoring and treatment options for employees with MSDs. The PHE project and corresponding studies were approved by the Social Ethics Commission of KU Leuven (G-2019081713) and carried out according to the Belgian and international privacy and ethical legislation. The Belgian occupational service for protection and prevention at work (IDEWE) was responsible for the recruitment. They distributed the information about the project amongst their clients and employees. Interested employees had to provide informed consent to participate.

6.3 Methodology

Let's assume pain levels measured across time are available for N subjects as 'training data' $\mathcal{D} = \{(\mathbf{x}^1, \mathbf{y}^1), ...(\mathbf{x}^N, \mathbf{y}^N)\}$, where $\mathbf{x}^i = [t_1^i\, t_2^i\, t_3^i\, \cdots\, t_m^i]$ represents the input variable 'time' up to a certain day t_m^i and $\mathbf{y}^i = [y_1^i\, y_2^i\, y_3^i\, \cdots\, y_{m^i}^i]$ represents the output variable 'pain' for the i^{th} subject, where $y_k^i = y(t_k^i)$.

[1] https://www.idewe.be/health-empower

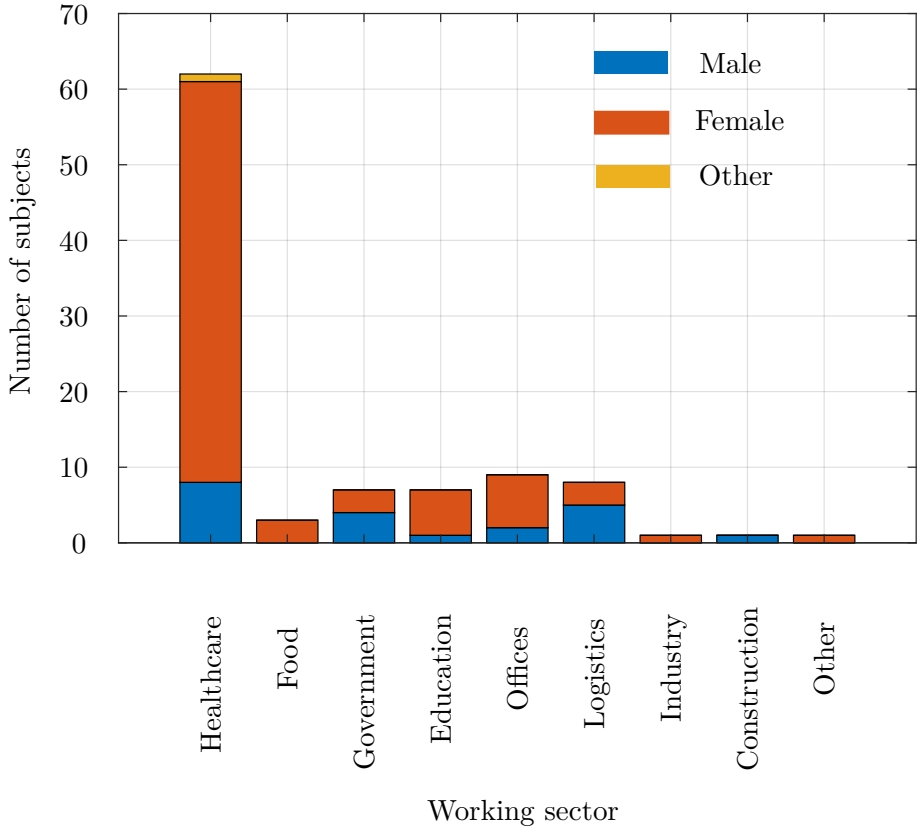


Figure 6.2: 21 males and 77 females participated in the study with majority (62 out of 99) working in the healthcare industry providing care.

In addition, data from a person of interest, henceforth referred to as the target individual, are provided till a certain day t_d^+ as $\mathcal{S} = \{(t_1^+, y_1^+), (t_2^+, y_2^+), \cdots, (t_d^+, y_d^+)\}$.

We try to learn a mapping f from the training and target data, such that,

$$y^+ = f(t^+) + \epsilon. \tag{6.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independent and identically distributed (i.i.d) gaussian.

Using the learnt model f, the target individual's pain measurements are then predicted at time $t^+_{m^i}$ as $y(t^+_{m^i}) = f(t^+_{m^i})$.

6.3.1 Smoothing

Let's begin by discussing the smoothing operation. Given a time series in training data $\mathbf{y}^i = [y^i_1\, y^i_2\, y^i_3\, \cdots\, y^i_{m^i}]$, a moving average (MA) of order w can be used to obtain a smoothed time series $\hat{\mathbf{y}}^i = [\hat{y}^i_1\, \hat{y}^i_2\, \hat{y}^i_3\, \cdots\, \hat{y}^i_{m^i}]$. This w-MA can be written as

$$\hat{y}^i_t = \frac{1}{w} \sum_{j=-\frac{w-1}{2}}^{\frac{w-1}{2}} y^i_{t+j}, \tag{6.2}$$

where w is an odd integer. Moreover, $\lfloor \frac{w}{2} \rfloor$ zeros are padded to the beginning and end of the given time series $\mathbf{y}^i$ to obtain same m number of observations in the derived w-MA time series in eq. 6.2. If the w-length time window contains missing observations for a given non-uniformly sampled time series, just the available points are used to calculate the moving average.

6.3.2 Self-Normalisation

We normalise a given time-series with its available individual information. A time-series $\mathbf{y}^i = [y^i_1\, y^i_2\, y^i_3\, \cdots\, y^i_{m^i}]$ is normalised using mean $\mu_{\mathbf{y}^i}$ and standard deviation $\sigma_{\mathbf{y}^i}$ calculated as follows:

$$\begin{aligned} \mu_{\mathbf{y}^i} &= \frac{1}{m^i} \sum_{j=1}^{m^i} y^i_j \\ \sigma_{\mathbf{y}^i} &= \sqrt{\frac{1}{m^i} \sum_{j=1}^{m^i} (y^i_j - \mu_{\mathbf{y}^i})^2}. \end{aligned} \tag{6.3}$$

The j^{th} observation $(\overline{y}^i_j)$ of normalised time-series $\overline{\mathbf{y}}^i = [\overline{y}^i_1\, \overline{y}^i_2\, \overline{y}^i_3\, \cdots\, \overline{y}^i_{m^i}]$ is obtained from the time-series $\overline{\mathbf{y}}^i$ as

$$\overline{y}^i_j = \frac{y^i_j - \mu_{\mathbf{y}^i}}{\sigma_{\mathbf{y}^i}} \tag{6.4}$$

The normalised data can be rescaled to original scale as $y_j^i = \overline{y}_j^i \times \sigma_{\mathbf{y}^i} + \mu_{\mathbf{y}^i}$.

6.3.3 Regression

We use Gaussian Processes (GP) as they are the state-of-the-art time series modelling methods when dealing with missing data. GP is defined as a set of random variables, such that any finite number of them have a joint Gaussian distribution [28]. 'f' from eq. (6.1) is defined as a GP $f(t) \sim \mathcal{GP}(m(t), k(t, t'))$, with mean function $m(t)$ and covariance function $k(t, t')$. We assume the data is noisy with i.i.d gaussian noise, having noise covariance σ_n^2, and choose a squared exponential kernel as the gaussian covariance function to model the closeness of two observations,

$$k(t, t') = \sigma_f^2 \, exp\left[\frac{-(t - t')^2}{2l^2}\right]. \tag{6.5}$$

As is evident from eqn. 6.5, the similarity between two observations decreases exponentially as t begins to differ from t', i.e the similarity is highest when $t = t'$. Thus, when two observations are far apart in time, the kernel considers them more dissimilar than when they are closer together in time.

Given $\hat{\mathbf{y}} = [y_1^1, \cdots, y_m^1, \cdots, y_N^1, \cdots, y_N^N]^T$ and $\mathbf{K}$ as a matrix of entries $K_{p,q} = k(t_p, t_q), \forall t_p, t_q \in \mathcal{D}$. We optimise the hyper-parameters $\{\sigma_f, l, \sigma_n\}$ by maximising the marginal likelihood $p(\hat{\mathbf{y}}|\mathcal{D}; \{\sigma_f, l, \sigma_n\})$ [28]. The prediction at time $t_{m^i}^+$ is given as a gaussian distribution whose mean, μ and variance, σ^2 are given by

$$\begin{aligned} \mu(t_{m^i}^+) &= {\mathbf{k}_+}^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \hat{\mathbf{y}} \\ \sigma(t_m^+) &= k(t_m^+, t_m^+) - {\mathbf{k}_+}^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_+, \end{aligned} \tag{6.6}$$

where $\mathbf{k}_+ = \mathbf{k}(t_m^+)$, $\mathbf{k}(t_m^+) = [k(t_m^+, t_1^1), \cdots, k(t_m^+, t_m^N)]^T$.

Gaussian process prediction is hampered by the fact that the computing complexity of inference and likelihood evaluation is $\mathcal{O}(n^3)$, where n is the input size, making it impractical for bigger data sets. Next, we will explore subset selection, which can minimise computing complexity while enhancing prediction accuracy.

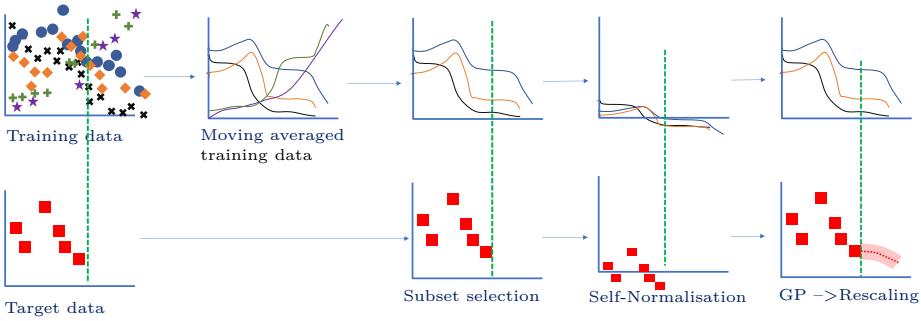


Figure 6.3: An illustration of our methodology. Moving averaging is performed on the training data to smoothen it. Target data is available until a day t_d^+ (dotted green line). Subset selection is performed on moving-averaged training data that shares similar temporal pattern to the target observations. Each time series (target or training) is self-normalised with its available observations before being fed to Gaussian Processes. A prediction on target data is made (red dotted line).

6.3.4 Subset selection

We follow the subset selection approach from [27] to find a smaller but informative subset ($\hat{\mathcal{D}}$) of the training data for a given target individual's data. Particularly, a subset $\hat{\mathcal{D}}$ with $M(<< N)$ individuals' data is found from the given training data $\mathcal{D}$,

$$\begin{aligned}\hat{\mathcal{D}} &= \{(\mathbf{x}^1, \mathbf{y}^1), \cdots, (\mathbf{x}^M, \mathbf{y}^M)\} \\ &= \left\{(t_1^1, y_1^1), .., (t_{m^1}^1, y_{m^1}^1), .., (t_1^M, y_1^M), .., (t_{m^M}^M, y_{m^M}^M)\right\},\end{aligned} \tag{6.7}$$

such that the individuals selected in the subset are similar to target individual's pain trend.

Using a subset $\hat{\mathcal{D}}$ with $M(<< N)$ individuals' data gives a computational advantage over considering N subjects, as the time complexity of GPs training and inference is proportional to the cubic power of the number of observations. Furthermore, if the most informative subset is selected, the prediction capability is improved. This is due to the fact that, during training, observations from M patients with a similar trend in pain are close to each other and have less variability at any given time t. Due to

inter-subject variances, this variability (at time t) is high when all N individuals are considered for training Gaussian processes.

To find the closeness between two time series, we use the Dynamic Time warping (DTW) as the distance metric. The choice of DTW metric as a distance measure is due to its capability to index time series with unequal lengths [29].

The subset selection is a two-step process in which (i) distances between the target time series and time series in training data is calculated, and then (ii) the nearest subset is *dynamically* selected based on the calculated distances.

Distances between the target data $\mathcal{S} = \{(t_1^+, y_1^+), (t_2^+, y_2^+)..., (t_d^+, y_d^+)\}$ and individual time series in training data $\mathcal{D}$ are calculated using the dynamic time warping (DTW) distance metric. Let's denote the DTW distance between target time series (denoted by $+$) and i^{th} time series in training data by λ_{i+}. Remark that that target data is only available until t_d^+ but the time-series in training data are present until $t_m^i (>> t_d^+)$. Therefore, the data for time series in training data are considered only until day t_d^+ to calculate the distance λ_{i+}. If the data at t_d^i is not available, the nearest time point $< t_d^+$ is chosen. The distance vector $\mathbf{\Lambda}_+ = [\lambda_{1+}\, \lambda_{2+}\, \cdots\, \lambda_{N+}]$ is calculated between target time series and all the time series in training data.

Subset selection is dynamically done based on the distance vector $\mathbf{\Lambda}_+$. First, $\mathbf{\Lambda}_+$ is sorted in ascending order. This ensures that the subjects are arranged in order of their closeness to the target subject, $\hat{\mathbf{\Lambda}}_+ = [\hat{\lambda}_{1+}\, \hat{\lambda}_{2+}\, \cdots\, \hat{\lambda}_{N+}]$, such that $\hat{\lambda}_{k+} \leq \hat{\lambda}_{(k+1)+} \forall k = 1, 2, \cdots, N$. Second, *turning* points at index 'k' are calculated, such that,

$$\left(\hat{\lambda}_{(k-1)+} - \hat{\lambda}_{(k-2)+}\right) \leq \left(\hat{\lambda}_{k+} - \hat{\lambda}_{(k-1)+}\right) \geq \left(\hat{\lambda}_{(k+1)+} - \hat{\lambda}_{k+}\right),$$

Multiple such turning points can exist at different indexes in $\hat{\mathbf{\Lambda}}_+$ vector. Third, the value at the distance value at the index k where the first turning point occurs (λ_k) is chosen as the distance threshold to calculate the closest subset. i^{th} time series in $\mathcal{D}$ is selected in the subset if $\lambda_{i+} < \lambda_k$. Choosing the first turning point enables the dynamic selection of the smallest and most informative subset. Fig. 6.3 showcases the processing pipeline where the moving average based smoothing is done on the

training data before subset selection. Since pain levels in each individual series are normalized using self-data, all training time series are scaled to the same level prior to being fed into the Gaussian process model. We will observe that this enhances the reliability of the predictions.

6.4 Experiments

We perform leave-one-subject-out (LOSO) cross-validation to evaluate the performance of our proposed approach. In each iteration, a unique individual's data is treated as target data and rest of the subjects' data are the training data. We first smoothen the training data and target data using a moving averaging of order five ($w = 5$). Then, in each iteration, a closest subset is evaluated dynamically with respect to the target data followed by self-normalising each time series (target data and selected subset) using eq. 6.3 and 6.4. Note that our subset selection approach dynamically selects a threshold in each iteration (i.e for each target data). A GP based regression is performed to forecast the future values for the target subject. The performance of regression was computed using Mean Absolute Error (MAE) averaged over N subjects. MAE for prediction at a time t_h is given as $MAE(t_h) = \frac{1}{N}\sum_{i=1}^{N} |y^{pred}(t_h^i) - y^{orig}(t_h^i)|$.

6.4.1 State-of-the-art

- **Baseline**: A baseline was created to judge the performance of the algorithms. This baseline was created by using the last available value of the target subject as future prediction of the daily pain value.

- **ARIMA**: Auto-Regressive Integrated Moving Average (ARIMA) has remained a state-of-the-art time series forecasting approach with uniformly spaced samples of time series [30]. Through linear interpolation, uniformity was introduced into the sparsely sampled pain time series of the subject of interest. Then, an ARIMA(p, d, q) model was fit on the uniformly sampled target time series. In order to find the optimal autoregressive order (p), degree of differencing (d), and moving average order (q), a grid-search was performed

to find the optimal hyperparameters following [31]. The learned model is then used to make a multi-step-ahead prediction of pain levels using the optimized hyperparameters.

- **LSTM**: Long short-Term Memory networks (LSTM) are deep learning techniques that can produce exceptional prediction performance by implementing gates (forget, memory, and output) that regulate the flow of information during training [32]. We follow a similar approach as with ARIMA approach where the avaiable data from a target subject is uniformly sampled by linear interpolation. We evaluate an LSTM network with 10 hidden units and the training is done using ADAM's optimisation to minimise mean absolute error [33].

- **Maximum-a-Posteriori (MAP) estimation**: A l^{th} order polynomial can be fit using available target data to estimate the polynomial coefficients θ_i, $\forall i \in \{1, 2, \cdots, l\}$ [27]. Moreover, subjects from training data can be used to create priors over the polynomial coefficients to get a better estimate known as *maximum-a-posterior* (MAP) estimate [27]. We test with polynomial of different orders (order 1 to 5) to find that the first order polynomial produces the least mean absolute error in LOSO cross-validation.

6.5 Results & Discussion

In this research, we investigate whether it is possible to estimate a person's pain levels using a small number of non-uniformly collected historical pain measurements. As pain data is subjective and varies amongst individuals, we also intended to determine if we might improve pain prediction by incorporating the prior pain measurements of other individuals into the training dataset. For this reason, we study the performance of various algorithms presented in this paper when predicting pain levels of an individual in future. In Fig. 6.4, we present the Mean Absoute Error (MAE) when predicting the pain 7 days ahead on the y-axis. On the x-axis in Fig.6.4, the availability of target data until a certain day is presented. Subset selection (SS) along with moving averaging (MA) and/or self-Normalisation (SN) were performed and Gaussian processes

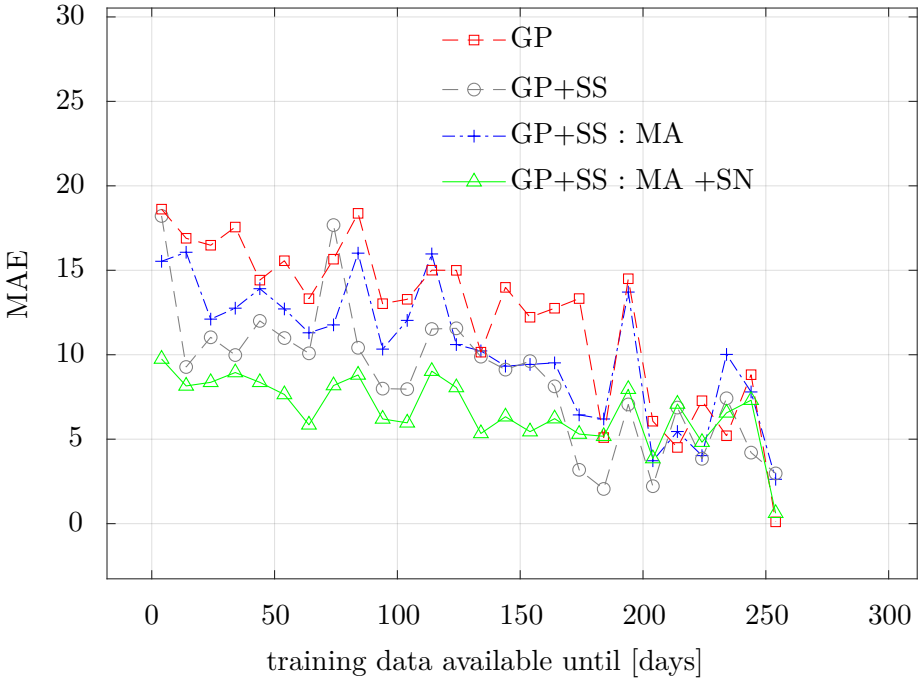


Figure 6.4: Mean absolute error (MAE) is measured with respect to availability of target data. Different combinations of subset selection (SS) followed by Gaussian processes (GP) were performed with proposed pre-processing components such as moving averaging (MA) and/or self-Normalisation (SN).

was used as a regression model. It is evident from Fig. 6.4 that the performance of subset selection (SS) followed by Gaussian processes (GP) is demonstrably superior to that of Gaussian processes alone. This is a result of the inclusion of an informative subset of participants in training who exhibit a comparable trend in pain to the target data. Additionally, subset selection on the moving averaged (MA) time series of the training data, followed by self-normalisation and subsequently the Gaussian process, performed the best, particularly when predicting for less available target data.

We hypothesised that pain data is subjective and that self-reported pain measurements are biased because individuals can only compare

their current pain feelings to their past pain experiences. Therefore, self-normalisation with respect to the historical pain measurements of an individual provides this significant performance improvement. In addition, as the availability of personal pain data increases over time, so does the accuracy of prediction. We believe that as more training data becomes available from an individual, the selected subset will consist of subjects whose patterns resemble that of the target subject more closely than when there are only a few data points. Thus, the variance in the training data available for regression is less and thus the prediction improves. This is evident by the decreasing trend in MAE when more training data becomes available. We also tested with self-normalization prior to subset selection and found no significant performance differences. This may be due to the fact that the DTW distance comparison for subset selection compares the relative difference in distances between two time series and picks more or less similar individuals with or without self-normalisation.

Next, we present the comparison of the proposed approach (GP+SS:MA+SN) with state-of-the-art approaches presented in section 6.4.1 when predicting pain values $[0 - 100]$. Fig. 6.5 shows that the proposed approach's performance is best when it comes to early prediction using only few available data points (until day 100).

The performance becomes comparable (if not better) with the state-of-the-art approaches (MAP) as more data in time becomes available for a given individual. On the basis of a paired t-test with equal variances, the performance differences between the proposed approach and other SOTA methods are statistically significant at 5% level of significance (until day 50). We discovered no statistically significant difference between the proposed method and MAP-based polynomial estimate when training with data for more than 100 days. Given the simplicity of the dataset, it seems intuitive that when more pain data becomes available, simple polynomial-based estimating algorithms will perform better.

We also observed that the state-of-the-art approaches (except LSTM) perform worse than the baseline when the availability of individual training data is limited (at least until day 50). Remark that the baseline is simply the previous observed value of pain carried forward for the prediction of future values. This is due to the difficulty of modelling

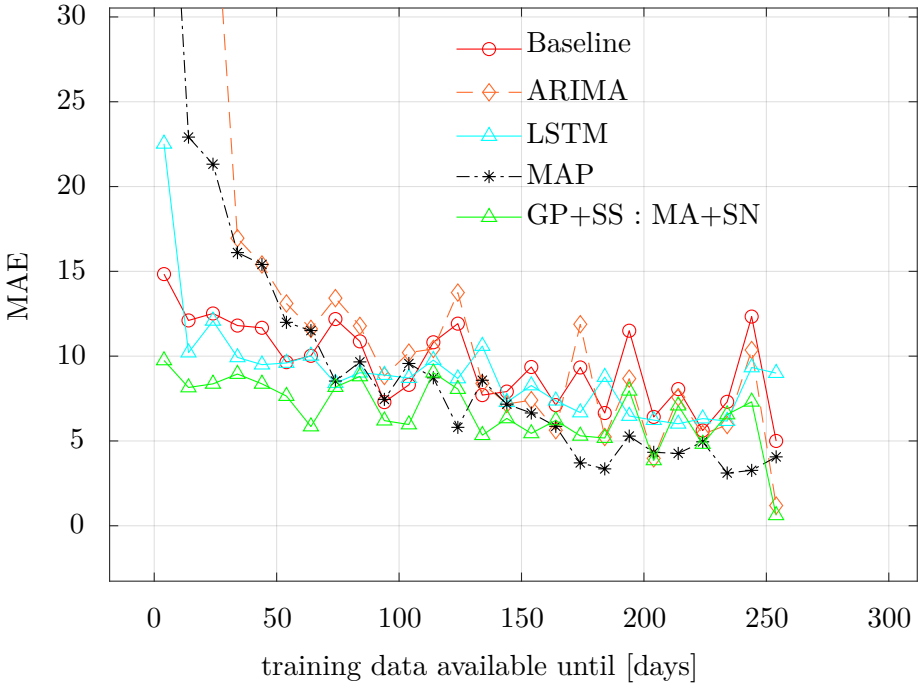


Figure 6.5: Comparison of the proposed approach with state-of-the-art approaches. When little training data is available (until day 100), the proposed method beats SOTA, and when more training data becomes available, it performs comparably or even better.

sparsely sampled time series with few observations. Our proposed method, on the other hand, overcomes this difficulty by incorporating the subjective nature of pain experience and modelling information rich subset selection along with personal data.

6.6 Conclusion

We proposed a novel Gaussian processes estimator and information-rich preprocessing to model an individual's workplace-related pain experiences. When time series data is irregularly sampled, the proposed approach outperforms state-of-the-art time-series forecasting algorithms

for early prediction. This can aid in the development of interventions for managing pain in the workplace, thereby reducing the possibility of 'pain chronification'.

6.7 Limitations & Future work

A limitation of our approach is the scalability of Gaussian processes as we believe that considering a large number of subjects ($N > 10^4$) will result in a larger subset (high value of M) of training data, increasing the computational complexity of our method. Sparse GPs are model approximation techniques that, when applied to a large number of subjects, can further reduce complexity [28].

In the future, we hope to broaden the modality of the input data in order to obtain more objective feedback on pain experiences. Finding an association of pain with physical activity data measured by a wearable, for example, can help as another meaningful feature to improve prediction performance. Similar to the maximum-a-posteriori approach, priors on the normalisation constants can be generated from training data and used to adjust the self-normalisation mean and standard deviation.

ACKNOWLEDGMENT

Chetanya Puri has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 766139. This work is part of the research project Personal Health Empowerment (PHE) with project number HBC.2018.2012, financed by Flanders Innovation & Entrepreneurship. This publication reflects only the authors' view and the REA is not responsible for any use that may be made of the information it contains.

Bibliography

[1] C. Puri, S. Keyaerts, M. Szymanski, L. Godderis, K. Verbert, S. Luca, and B. Vanrumste, "Daily pain prediction in workplace using gaussian processes", *HEALTHINF*, 2023 (accepted).

[2] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Safer and Healthier Work for All — Modernisation of the EU Occupational Safety and Health Legislation and Policy*. COM, 2017, p. 9. [Online]. Available: `https://ec.europa.eu/social/BlobServlet?docId=16874&langId=en`.

[3] J. Kok, P. Vroonhof, J. Snijders, G. Roullis, M. Clarke, K. Peereboom, P. Dorst, I. Isusi, and European Agency for Safety and Health at Work, *Work-related musculoskeletal disorders : prevalence, costs and demographics in the EU*. Publications Office, 2020. DOI: `doi:10.2802/66947`.

[4] D. Kim, "Effect of musculoskeletal pain of care workers on job satisfaction", *Journal of Physical Therapy Science*, vol. 30, pp. 164–168, Jan. 2018, ISSN: 0915-5287.

[5] B. Morlion, F. Coluzzi, D. Aldington, M. Kocot-Kepska, J. Pergolizzi, A. C. Mangas, K. Ahlbeck, and E. Kalso, "Pain chronification: What should a non-pain medicine specialist know?", *Current Medical Research and Opinion*, vol. 34, no. 7, pp. 1169–1178, 2018.

[6] G. L. Moseley and A. Arntz, "The context of a noxious stimulus affects the pain it evokes", *PAIN®*, vol. 133, no. 1-3, pp. 64–71, 2007.

[7] J. Nijs, C. P. Van Wilgen, J. Van Oosterwijck, M. van Ittersum, and M. Meeus, "How to explain central sensitization to patients with 'unexplained'chronic musculoskeletal pain: Practice guidelines", *Manual therapy*, vol. 16, no. 5, pp. 413–418, 2011.

[8] R. R. Edwards, R. H. Dworkin, M. D. Sullivan, D. C. Turk, and A. D. Wasan, "The role of psychosocial processes in the development and maintenance of chronic pain", *The Journal of Pain*, vol. 17, no. 9, T70–T92, 2016.

[9] J. W. Vlaeyen and S. J. Linton, "Fear-avoidance and its consequences in chronic musculoskeletal pain: A state of the art", *Pain*, vol. 85, no. 3, pp. 317–332, 2000.

[10] M. I. Hasenbring, D. Hallner, B. Klasen, I. Streitlein-Böhme, R. Willburger, and H. Rusche, "Pain-related avoidance versus endurance in primary care patients with subacute back pain: Psychological characteristics and outcome at a 6-month follow-up", *Pain*, vol. 153, no. 1, pp. 211–217, 2012.

[11] S. Bunzli, A. Smith, R. Schütze, I. Lin, and P. O'Sullivan, "Making sense of low back pain and pain-related fear", *journal of orthopaedic & sports physical therapy*, vol. 47, no. 9, pp. 628–636, 2017.

[12] L. M. McCracken and C. Eccleston, "A prospective study of acceptance of pain and patient functioning with chronic pain", *Pain*, vol. 118, no. 1-2, pp. 164–169, 2005.

[13] C. Lalloo, L. A. Jibb, J. Rivera, A. Agarwal, and J. N. Stinson, "There's a pain app for that", *The Clinical journal of pain*, vol. 31, no. 6, pp. 557–563, 2015.

[14] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2017.

[15] J. Lötsch and A. Ultsch, "Machine learning in pain research", *Pain*, vol. 159, no. 4, p. 623, 2018.

[16] P. J. Tighe, C. A. Harle, R. W. Hurley, H. Aytug, A. P. Boezaart, and R. B. Fillingim, "Teaching a machine to feel postoperative pain: Combining high-dimensional clinical data with machine learning algorithms to forecast acute postoperative pain", *Pain Medicine*, vol. 16, no. 7, pp. 1386–1401, 2015.

[17] J. Lee, I. Mawla, J. Kim, M. L. Loggia, A. Ortiz, C. Jung, S.-T. Chan, J. Gerber, V. J. Schmithorst, R. R. Edwards, *et al.*, "Machine learning-based prediction of clinical pain using multimodal neuroimaging and autonomic metrics", *Pain*, vol. 160, no. 3, p. 550, 2019.

[18] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long- and short-term temporal patterns with deep neural networks", in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 95–104.

[19] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber", in *International Conference on Machine Learning*, vol. 34, 2017, pp. 1–5.

[20] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "GRU-ODE-bayes: Continuous modeling of sporadically-observed time series", in *Advances in Neural Information Processing Systems*, 2019, pp. 7379–7390.

[21] M. Liu, A. Zeng, Z. Xu, Q. Lai, and Q. Xu, "Time series is a special sequence: Forecasting with sample convolution and interaction", *arXiv preprint arXiv:2106.09305*, 2021.

[22] N. Strodthoff and P. e. Wagner, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, 2020.

[23] K. Li, C. Liu, T. Zhu, P. Herrero, and P. Georgiou, "GluNet: A deep learning framework for accurate glucose forecasting", *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 414–423, 2019.

[24] Z. C. Lipton, D. Kale, and R. Wetzel, "Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series", in *Machine Learning for Healthcare Conference*, 2016, pp. 253–270.

[25] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask gaussian process rnn classifier", in *International conference on machine learning*, PMLR, 2017, pp. 1174–1182.

[26] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward", *PloS one*, vol. 13, no. 3, e0194889, 2018.

[27] C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278.

[28] C. E. Rasmussen, "Gaussian processes in machine learning", in *Advanced lectures on machine learning*, Springer, 2004, pp. 63–71.

[29] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping", *Knowledge and information systems*, vol. 7, no. 3, pp. 358–386, 2005.

[30] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[31] R. Shibata, "Selection of the order of an autoregressive model by akaike's information criterion", *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

Chapter 7

Conclusion

The primary aim of this dissertation was to determine *if machine learning models can attain reliable performance when given with a range of data-related challenges, notably in the context of time series data in healthcare.* This aim was further sub-categorised into different research questions and their application specific sub-objectives, as detailed in section 1.3. Fig. 1.4 outlines the specific applications that address one or more of these research questions. These research questions are:

RQ1: Can we predict a patient's health state with limited patient-specific data?

RQ2: Can we detect infant mortality using structured tabular data with a very high percentage of missing data?

RQ3: Can we create personalized machine learning models that can adapt over time to generate accurate predictions using few data points?

RQ4: Can we build machine learning models that can train in a secure manner while dealing with sensitive raw data without losing prediction performance?

In this concluding chapter, we will revisit the various research questions and discuss the conclusions and lessons learned throughout the research study. The next section is devoted to limitations, which pave the way

for further research. The final section addresses the valorisation of the study's findings.

7.1 Revisiting the research questions

RQ1 : Can we predict a patient's health state with limited patient-specific data?

By constructing automated models that can learn patterns from a subject's historical data and anticipate future behavior, it is currently possible to monitor a person whether they are in a hospital or at home. Healthcare applications, such as those discussed in this dissertation suffer from challenges such as limited availability of data. These limitations were described in great detail in the chapter 1.1.1.

We collected and analysed weight gain data from pregnant women (Chapters 2), multi-modal data related to cognitive decline in alzheimer's patients (Chapter 3) and pain levels in an individual in a workplace (Chapter 6). Chapter 2, 3 and 6 focus on prediction of time-series when in a small data corpus, the individual data is only available until a certain day for model creation. In these chapters, the sub-objectives of RQ1 were addressed by conducting experiments in which models were learned with varying amounts of training data availability until a certain time, for example, in the pregnancy use case, end-of-pregnancy-weight gain predictions were made when personal training data was only available until day $50, 60, \cdots, 260$. Similar tests were conducted on other use-cases. We proposed learning of informative priors from the training data followed by maximum-a-posteriori estimation to get high performance in forecasting. This prompted us to consider the possibility that the morphology of the time series might be indicative of similar behaviour across two or more subjects. Typically, clustering algorithms can be used to create multiple groups of individuals (subsets), with each group containing data from subjects with a similar trajectory in their health state over time, such as similar weight gain trends or cognitive decline. Therefore, whenever data from a new subject becomes available, it can be mapped to a subset including subjects with which the new subject shares similarities. However, it is difficult to create these groups when

the lengths of two time series are unequal. Additionally, the number of groups must be determined beforehand. This thesis introduced a novel method (Chapter 3, 6) that can find a subset of individuals without requiring a parameter for number of subsets and can also handle time series with different lengths. The proposed subset selection method not only aided in dynamically selecting informative training data points, but it also resulted in better computational complexity of the Gaussian Processes-based regression technique. A basic schematic of the proposed approach is shown in Fig. 3.1. This helped us in building models that can tackle limited time-series data from an individual while still being able to produce reliable performance.

We were able to accurately model time series data when only a limited amount of personal data was available by exploiting an informative subset of the training data to create accurate predictions about the future. Predicting gestational weight gain, estimating Alzheimer's patient's cognitive decline, and estimating workplace pain were three applications with similar challenges on which the proposed method was evaluated. These applications are discussed in Chapters 2, 3, and 6. We tested the performance of the proposed approach by varying the data availability in time for each use case which is discussed in the relevant chapters. The results demonstrate that, when enhanced by an informative subset, personal data with only a few measurements across time can be modeled to make accurate predictions, even when the prediction horizon is large.

RQ2 : Can we detect infant mortality using structured tabular data with a very high percentage of missing data?

Chapter 5 examined the case of tabular healthcare data with a lot of missing obsevations. We learned the importance of understanding the underlying mechanisms in which data is missing, as opposed to naively imputing it prior to feeding it to a machine learning model. A case study was provided in which blind imputation's flaws led to findings that created a misleading impression of a good model. Consequently, it is essential to identify variables that may be biased or informative in the manner missing data is observed (or not).

We developed a method to eliminate biased features caused by missing data.

RQ3: Can we create personalized machine learning models that can adapt over time to generate accurate predictions using few data points?

In Chapters 2, 3 and 4, we discovered that models that incorporated a global model and were fine-tuned with personal data outperformed those that were simply learned from personal data or only a general global model. There was an improvement of around 25% and 31% in mean absolute error over the best of state-of-the-art for the gestational weight gain prediction datasets in the Netherlands and China respectively as stated in chapter 2.6. Similarly, cognitive decline over a period of eight years at 6-month intervals was predicted by a training model using personal data spanning thirty months with an improvement of over 20% over state-of-the-art methods as described in section 3.7.2. Our models, MAP or subset selection with Gaussian processes outperformed the state-of-the art in respective use-cases with the ability to personalise with less training data from an individual. Due to the selection of informative subsets, intra-subject variability was reduced, resulting in improved prediction performance.

Moreover, an interesting finding was made while investigating the Alzheimer's use case. The individuals enrolled for this investigation were at various stages of cognitive decline. This necessitated matching the time series in the training data with respect to their decline rather than their recruitment time as the initial time point. We proposed a collective time-series realignment approach that further improved time series forecasting performance.

Early prediction is also an important aspect that is closely connected to the challenge of personalising a model in the presence of limited individual data. Personalisation performance in our techniques improves as more data becomes available, but essential interventions can be effective if the model can detect a condition as early as feasible, resulting in fewer personal data. Therefore, the performance of several algorithms (proposed or existing) was evaluated based on the availability of target subject's data till a specific time.

This thesis also offered a learning scenario in which the federation of edge devices leads to the formation of a global model via the aggregation

of personalised local models rather than the transmission of raw data to a centralised server to create a global model., which is discussed in detail in Chapter 4.

Model personalisation was achieved by beginning with a general model and then fine-tuning it using target individual's data. This allowed us to develop high-performance time-series forecasting techniques that could learn an individual's characteristics from a small number of personal observations. We also demonstrated that a model's performance improves as more personal data becomes available.

RQ4: Can we build machine learning models that can train in a secure manner while dealing with sensitive raw data without losing prediction performance?

With the expansion of digital rights protection activities, the need for AI applications to guarantee privacy while processing personal data has become critical. For instance, the cross-border transmission of sensitive data is now subject to multiple reviews to determine if particular privacy safeguards are implemented to the data, but transmission can also be sometimes prohibited due to the level of sensitivity of the data. Additionally, several data minimisation and other data anonymisation principles are employed to the "data" in order to maintain anonymity of the users. We presented an additional "model-based" privacy preserving technique in Chapter 4 that is in line with the Google's federated learning. By sharing individual models trained on local data that are aggregated at the server, we demonstrate that a federation of edge-devices may perform substantially more effective in personalised time-series forecasting. This differs from the conventional centralised approach, in which raw data is transmitted to a central server where the model is trained.

Additionally, we experimented with the cross-border transfer of ML models from The Netherlands to China rather than raw data, as the transfer of raw data across borders was prohibited. The results indicate that the model trained on Dutch data predicted weight gain far more accurately than the model trained and predicted on solely Chinese data. This was due to the greater diversity present in the dataset from Netherlands.

With the proposed method, we were able to train machine learning models

at the edge devices, where sensitive data pertaining to pregnant women was not shared, and a global model was generated by the aggregation of smaller model updates that individuals learned on their private data. Due to the limitations of cross-border data transfer, we also employed a transfer learning technique in which trained models rather than raw data were transferred from one geographical location to another.

7.2 Limitations and Future Work

We would like to address some of the limitations of our proposed approach and describe how we envision future study in this field and various use-cases. We examined a number of time series forecasting algorithms for datasets with limited and/or missing data. We conducted an experiment in chapter 2 where the underlying principle is to first construct a general model that is trained using the provided training data. Individual data can then be used to fine-tune the general model for enhanced performance in prediction. This was the maximum-a-posteriori approach for polynomial modelling. We developed subset selection (SS) followed by gaussian processes regression (GP) based on a similar principle. SS identifies a subset of individuals whose data is most informative in relation to the target data. The proposed SS can work with unequal length time series. The SS approach is non-parametric that determines the size of the subset automatically based on the relative similarities for a given target subject. Powerful GP approaches that can model input data with sparse observations are used to model this selected subset of data and the target data. The addition of an informative subset improves the computational cost of the GPs while maintaining a particular level of prediction performance.

Finding similarities between a target subject and individuals from training during subset selection proved challenging when there were limited observations of the target individual. This is due to the fact that the dissimilarity calculation employing dynamic time warping had only a few data points sporadically positioned in time to compare two individuals. This could result in a subset where subjects are similar to the target individual until the small window of available target data, but the long-

term behavior of a particular individual might differ significantly from the selected subset. If the time series is sampled at a high sampling rate and samples are separated consistently, this inaccuracy in temporal similarity can be reduced. However, it was difficult to obtain such a time series in the use-cases that we discuss in this dissertation.

Given huge amounts of data, deep learning models can be a good choice for learning these tasks. Most of the time-series literature that deals with time series classification using deep learning is usually supplied with large databases of time series with very high sampling rate. The *M*-competition is a forecasting competiton being organised for the past 40 years that aims at to find strategies to increase forecasting accuracy by empirically analyzing various forecasting systems and finding the most accurate one [1]. Multiple recent works [2]–[5] perform multi-step forecasting using deep neural forecasting models that achieve promising results on the M5-challenge [1]. Unlike the data considered in the thesis, this gold standard dataset for time series forecasting algorithms also consisted of large datasets with uniformly sampled time series data. Learning on a small dataset with non-uniformly sampled time-series data is a potential future work that needs to be addressed via deep learning.

Next, we'd like to examine the various use-case-specific limitations listed below.

7.2.1 Gestational Weight Gain prediction

The current study population is a sample of women from Eindhoven and Shanghai in the Netherlands and China, respectively. We would still like to explore if the model can be generalised to different geographies/ethinicities. For example, physiological, ethnic and cultural differences in populations may influence the progression of pregnancy, its accompanying weight gain and the risk severity [6]–[8]. Also, at the time of recruitment, it is difficult to predict if the individual will end up being underweight or overweight at the end of their pregnancy. Inadvertently, only a few women recruited for the study ended up being underweight at the end of the pregnancy; hence, the dataset is not completely representative of the population. This issue can be overcome by gathering more data and recruiting new pregnant women to the

study. To reliably forecast the weight gain at the end of pregnancy, the prediction of weight gain still requires data up to the middle of pregnancy. This is evident from the results shown in Fig. 2.7. This can be improved by taking more measurements in the early stages of pregnancy, which could not be done in our study. This is because the enrolled subjects were at least 10 weeks pregnant and had to have at least one measurement recorded prior to day 120. However, it is a first step toward pregnancy weight gain prediction.

In addition, the computational complexity of regression based on Gaussian processes is significant ($\mathcal{O}(n^3)$), which can cause scaling issues when training for more examples as the dynamic subset selection might select a large number of subjects. Identifying effective methods for lowering the computational complexity of Gaussian processes could be another future objective. For example, incorporating sparse GPs as model approximation techniques can further reduce complexity [9].

A different strategy that takes into account the similarities between various curves is to first find K similar curves compared to the test subjects' data. Then, create a histogram of weights at the end of pregnancy from these K similar subjects and identify the weight with the highest probability within it. In comparison to gaussian processes that account for the previous history in time of several individuals, this approach would likely be computationally less expensive, but this remains to be further researched.

Addition of different features such as pre-pregnancy weight and pre-pregnancy BMI for pre-processing in chapter 2 enhanced the prediction performance. Because we incorporate pre-pregnancy information into the weight gain time-series data, we did not use these characteristics to condition the priors in the chapter 2. However, the use of these features in combination with weight gain data to choose a more refined subset using either clustering methodologies or the proposed subset selection strategy in chapter 3 needs to be investigated in the future. Various modalities can be added to the dataset such as the amount of physical activity using a wearable or body glucose monitoring. Adding these modalities in addition to the typical self-reported or questionnaire-based modalities might provide new insights into the gestational health of the women.

We discussed the association of inadequate weight gain and its effect on pregnancy health in section 2.1. For example, authors in [10] show the effect of gestational weight gain on stress. The consequences of early weight gain prediction on health outcome may be a future study subject.

In Chapter 4, we proposed a privacy-preserving approach for predicting gestational weight gain, in the sense, that at no point the individual raw data is shared at the central server. Individual information is modeled into the parameters of the local model and might conceal the user's sensitive information. However, it does not offer complete privacy protection. For example, authors in [11] have shown that model inversion or membership inference can allow a malicious person to derive individual information from the ML models, if cross-referenced with public databases.

Homomorphic encryption is the process of encoding raw data into a form that enables users to perform computations on the encoded data. By adding homomorphic encryption [12] or differential privacy (withholding individual information and sharing collective patterns) to the learnt individual updates, more privacy can be added to the system [13]. In addition, we would like to point out that we created a modest proof-of-concept to demonstrate the performance with the addition of users, but we did not examine the complete design considerations from a functional standpoint to be able to deploy in a real-world scenario. For instance, a complete functional system design would take into account the elements like authorization checks, client dependability, system monitoring, client dropouts at various times, formal privacy guarantees, costs to communicate parameters between the server and local devices, as well as the energy consumption necessary for local devices [14].

7.2.2 Alzheimer's Disease Prediction

The subjects recruited in the study were at different stages of cognitive decline when they were recruited. We tackle this by performing collective realignment as described in section 3.5.2. It finds an optimal time-shift for the time series in training data by comparing (aligning) each one of them with the target time series. This process of finding an optimal lag is computationally expensive ($\mathcal{O}(n^2)$). This is one limitation of the proposed approach that can be addressed in the future. The data in this study

consisted of features extracted from different modalities as described in Fig. 1.2. These features were part of the publicly accessible database. Using modern deep learning techniques such as convolutional neural networks directly on the raw imaging data (MRI, PET) as opposed to the derived features could result in a performance increase. Such imaging techniques and their individual effects in predicting cognitive impairment in Alzheimer's patients are investigated in [15]–[17]. An ensemble model that incorporates multiple models learnt from the raw data from distinct modalities is yet to be investigated.

7.2.3 Infant Mortality Prediction

Healthcare survey data, such as that presented in the case study in chapter 5 suffers from missing values. We investigated exhaustively the identification of such features that contribute in a deceptive manner, primarily because the missing data is dependent on the class labels. This can also be determined by comprehending the MAR, NMAR, and MCAR assumptions underlying the missing data. Even though we investigate the identification of features that are biased due to these underlying assumptions, we have not proposed imputation techniques that can fill in these data gaps for improved machine learning model learning. Recently, authors in [18] have shown that whatever the missing assumption, jointly optimising the imputer and the classifier can provide the best performances.

7.2.4 Pain Management at Workplace

In chapter 6, we analyzed participants from several workplaces, with the majority of workers in the healthcare sector (nurses, caregivers, etc.). The acquired data's subjectivity is a significant constraint of pain data analysis. Pain is a subjective experience that varies from person to person based on cultural influences, situational perception, and other psychological factors [19]. There are other approaches available, including verbal rating scales and numerical rating scales, such as the one utilised in this study. This complexity of pain data may not always demonstrate great concordance, but it is the gold standard for delivering the most

reliable assessment of pain experience to date [20], [21]. Addition of more objective sources such as affective state detection from face scan, wearable based bio-physical signals such as Electrocardiogram (ECG), or galvanic skin response (GSR) could serve as potential future work [22].

Furthermore, at the time of data analysis, the acquisition of physical activity data utilizing FitBit data was still in progress. Using Fitbit-based physical activity data to determine the link between increased physical activity and pain sensation, or vice versa, could be a future research possibility.

7.3 ML in Healthcare : a Multidisciplinary view

The implementation of ML, particularly in healthcare, is constrained by a number of legal, economic, ethical, and societal challenges in addition to the technical difficulties discussed in this thesis. In order to address these multidisciplinary difficulties, we would like to briefly discuss them and propose a call for collaborative approaches.

The machine learning approach proposed in this thesis that are enabled by IoT devices such as smartphone and weighing scale based pregnancy monitoring pose a threat to users' security and privacy, especially when the data of users is shared across multiple applications. This is partially addressed in chapter 4 where raw private data of a user is not shared at a central server. However, the proposed approach is still a long way from being applied in a real-world situation. Similar to this, other use-cases where the data is accessible to doctors and healthcare professionals, such as the Alzheimer's cognitive decline measurement, should be maintained securely and processed in accordance with ethical and legal requirements at every stage of data processing. For instance, the General Data Protection Regulation (GDPR) in Europe upholds the right to the processing of personal data. This implies that the relevant authorizations for use must be obtained for the data processing, storage, and training of ML algorithms [23].

Fairness must be a priority in machine learning development if users are to use it over the long run and if different stakeholders are to trust it. This

is true for datasets used to train ML algorithms, where it is important to guarantee model training and validation to employ prediction algorithms ethically [24]. For instance, if the datasets used to train ML systems do not fully reflect the population, they could widen the gap between health disparities. In chapter 5, we address how bias affects the predictive power of machine learning models, but the proposed use-case is a case study on a publically available dataset, hence the findings in this PhD are more reactive than preventive since most biases emerge during the data collection process. These biases must be eliminated during the data collection process by design.

In chapter 2, we looked at a dataset of pregnant women from developed countries where the level of education and socioeconomic characteristics were not incorporated in the modelling process. This was as a result of the study's restriction to just the women registered with the participating midwife clinics. The importance of education in predicting food intake and BMI in pregnant women should not be underestimated, though. The addition of employment and household income to education improves the description of socioeconomic inequalities in food and health-related parameters [25]. Additionally, there was a high correlation between pre-pregnancy obesity and socioeconomic status, as well as negative behavioural patterns that can impair the efficacy of the required therapies [26]. As a result, in order to scale the solution, it will be crucial to thoroughly research the socioeconomic aspects and the influence of the proposed models in datasets from emerging nations in the future.

Scientific societies and regulatory agencies must work together to create best practises in order to prevent such problems. A number of institutional review bodies, as well as an ethics committee, must examine whether these requirements are being met [24].

A threat and risk assessment should be performed to ensure that the complete ML solutions adhere to numerous ethical, security, and privacy norms. Developers and legal professionals should work together to create these solutions.

7.4 Valorisation

This dissertation was developed as part of HEART project[1]. The project had multiple PhD students working towards interdisciplinary research to create innovative solutions for applications in healthcare.

In order to solve the technical, legal, and economical issues created by the digital transformation of the healthcare industry, an action research approach was developed for the construction and operation of a prototype of a health-related activity detection platform based on IoT, termed HEART [23].

The platform is the result of a collaboration founded by a European project – Horizon 2020 involving partner from healthcare industry, Data science department & personal health solutions at Philips Research, Eindhoven and two European universities: faculty of law at the University of Macerata (UniMC) in Italy, with expertise in privacy and business aspects of data analytics and market trends; and eMedia lab at the department of electrical engineering and department of computer sciences at KU Leuven (KUL), in Belgium, with expertise in developing activity recognition algorithms for Internet of Things (IoT). The platform stimulated research at multiple levels, which includes the acquisition of raw data and the implementation of several security and privacy measures in accordance with legal policies and application-specific requirements, followed by data analytics. The data analytics component is at the core of the HEART platform, around which this dissertation was developed. It entails extracting relevant information from data using cutting-edge and novel data analytics and machine learning techniques. These proof-of-concepts have the potential to become applications for end-users, such as healthcare professionals or individuals, who will be supported by the HEART platform. [23]. Thus, most of the challenges and applications addressed in this thesis are at the intersection of real-life implementation problems.

This thesis focuses on several use-cases and automating the predictive analytics component through the application of machine learning. We envision several applications based on these use-cases as follows,

[1] http://heart-itn.eu/

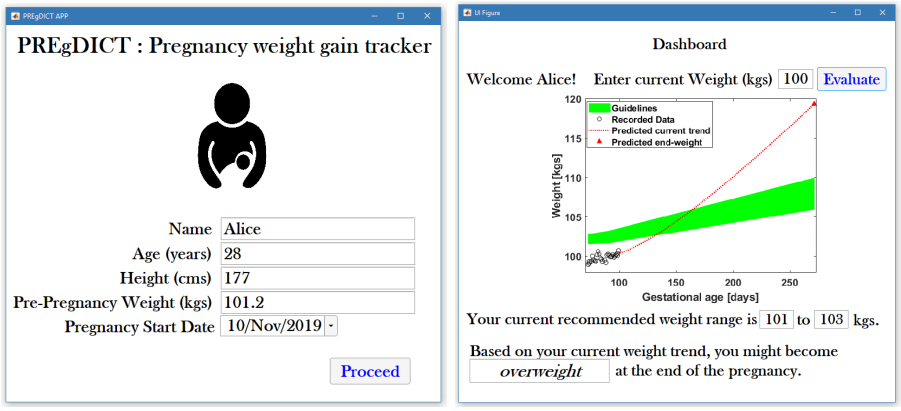
7.4.1 Pregnancy Health Application

Current healthcare practices, such as midwives, require manual data collection and storage in the patient's record. These records are frequently overwritten with the most recent value and do not provide a comprehensive perspective of a person's health over time. Pregnant women at risk must routinely see midwife practices for further health information. These midwife practices that may provide care for low to moderate risk pregnant women are one of the target groups for forming a business. If recognized early, the hazards can be easily managed by gathering further data/recommending tests. Additionally, pregnant women may feel comforted by this frequent monitoring of their pregnancy status. As a potential remedy, a health monitoring service might be developed to alleviate this industry gap, that can allow healthcare providers to monitor data over time and not just store it. This study's research permits unobtrusive remote monitoring of a pregnant woman. We anticipate an application that can be used securely by a pregnant woman and their healthcare provider, including features such as interpretation and visualization of health data in the home setting. This research also developed methods to analyse data that does not intrude the user's privacy.

It is anticipated that the global market for women health applications would rise at a compound annual growth rate (CAGR) of greater than 15 percent over the course of the following ten years, from its current value of US$ 2.3 billion in 2020 [27]–[29]. COVID19 has had a significant effect on the market landscape. Pregnant women and those undergoing reproductive treatments had fewer hospital and doctor visits due to the pandemic. This increased the utilisation of virtual care and self-care via women's health applications. Pregnancy apps segment have a major market share in this global industry with most women in high-income countries frequently using them. However, developing countries still lack the relevant penetration due to several cultural and linguistic reasons [30]. Furthermore, because young women and first-time mothers actively seek information and are inexperienced, they are more vulnerable to less reputable sources [30]. A coaching/lifestyle-intervention system based on projected pregnancy health can be incorporated to current apps as a feature. This system can track the weight gain in pregnant women and

can also serve as a guidance to healthcare professionals to keep track of the individual's gestational health.

To realise this, we also created a proof-of-concept (based on MATLAB software) that simulated different edge devices that collaborated to learn a global model to show the potential of privacy-preserving learning applied to the health-data. We demonstrated this at the ACM SenSyS conference in New York, United States, and received great response [31]. Fig. 7.1 showcases the snapshot of the application created as a proof-of-concept.



(a) Home screen (b) Weight Gain visualisation

Figure 7.1: Snapshot of the proof-of-concept created in MATLAB to enable privacy-preserving weight gain management

However, there are several considerations that should be taken into account,

- The clinical validity of the proposed approach must be established.
- Even though the application has the ability to monitor weight gain, its business potential needs to be researched.
- Additional capabilities such as lifestyle counseling could be implemented to provide customers with a comprehensive experience. However, a pilot study must be done to see whether people will adopt (especially in developing countries like India) and pay for such a service.

We feel that the proposed solution has enormous potential to serve a large demographic, namely pregnant women, and various stakeholders involved such as the healthcare professionals.

7.4.2 Alzheimer's Clinical Trial Design

There is an urgent requirement of disease modifying therapies that can delay the onset or slow down the progression of Alzheimer's Disease that burdens the current healthcare system and society. Despite significant breakthroughs in our understanding of the biology of Alzheimer's disease, no new molecular entity for the prevention or treatment of Alzheimer's disease has been approved since the 2003 memantine [32]. One of the major factors for these failures associated with the drug development can be late intervention, or poor selection of participants.

The majority of patients with cognitive decline first interact with their primary caregiver. Due to the absence of available diagnostic tools and the clinician's inability to identify a patient at risk for cognitive decline, patients in need could not enroll in clinical trials [33]. The proposed method can aid the caregiver in determining whether the individual is a possible early-onset Alzheimer's disease patient.

This work sets the path for future research in this area, which is not yet ready for commercialisation. We anticipate a collaborative research endeavor between a university with experience in statistical analysis and a medical institution with expertise in Alzheimer's disease. A pilot study must be conducted in which patients are screened based on the recommendations generated by machine learning. This technique would differ from a random screening. Randomly selecting participants in a clinical trial from heterogeneous populations may result in the following two outcomes:

1. More people with rapid cognitive decline are allocated to the intervention group than there are in the control group.
2. More people with slow cognitive decline are allocated to the intervention group than there are in the control group.

If the former is true, then the described treatment would appear useless, despite having a considerable influence on individuals with slower cognitive deterioration. Similarly, if the random selection results in the second scenario, the claimed treatment efficacy will be overestimated. Thus, a randomisation based on machine learning predictions can help in evenly allocating the subjects for an unbiased treatment assignment.

7.4.3 Pain Management Application

This study's data was obtained using questionaire-based data made accessible to participants via a mobile application. This study and application design was done by other PhD students at KU Leuven in association with IDEWE[2], an external service for Prevention and Protection at Work in Belgium. The mHealth application is available for iOS and Android users[3]. Participants are able to monitor their health, pain, and physical activity with the aid of novel advice and visualization techniques.

The application functions by gathering questionnaire responses as user input in 'MyDaily', a short daily questionnaire of physical activity and pain experience and daily satisfaction. We anticipate our proposed method of daily pain prediction to be incorporated into this application. This can be achieved by

- Monitoring employees with pain and its underlying cause.
- Providing recommendations to lessen the burden of pain on their daily life.

The pain prediction service could be implemented as (a) a visualization of personal pain history and anticipated future pain levels, and (b) a group-based insight, where a user is shown how similar they are to other users in a similar work-related context and with similar histororical pain levels. To the best our knowledge such an application does not exist.

[2]https://www.idewe.be/

[3]https://www.idewe.be/health-empower

7.4.4 Societal Impact

The study presented in this thesis focuses on several aspects of predictive modelling in healthcare. This is demonstrated by studying various use-cases spanning from in-hospital care to preventive care. We believe that if the research presented in this thesis is ever implemented in a practical context, it will have a broad societal impact in the future.

For instance, screening tools for estimating early cognitive deterioration can be created using the suggested approach. A computerised platform that offers clinical decision support and care planning can incorporate the patient's cognitive status and possible trajectory. This may lower healthcare expenses in situations where tests are suggested based on the progression of the disease. Furthermore, Alzheimer's clinical trial selection can help reduce the time required to perform **efficient clinical trials**. This is enabled by patient group segregation based on detected severity as described in section 7.4.2.

In addition, by identifying at-risk women and educating them on **healthier lifestyle** measures, as well as lowering the burden on healthcare providers, pregnancy health management can be improved. The proposed approach allows for real-time weight gain monitoring and can relieve physicians of regular tasks while **increasing engagement** with pregnant moms and their families in a more active and wellness-driven lifestyle.

The most common cause of chronic incapacity in industrialised nations is currently musculoskeletal problems, as detailed in chapter 6. They account for the significant expenditures of repeated treatment, extended absences from employment, and early retirement. Therefore, the necessity for developing effective therapies is urgent. Predictive modelling can improve public health management, such as the prevention of chronic pain in the workforce and the management of absenteeism at work. Early identification and intervention can enable people to monitor their health through digital coaching and adopt a healthier lifestyle. The occupational health services can create prevention programs for occupations that are physically challenging.

We believe that employing machine learning in healthcare has enormous

promise, but it must be human-centered. The multidisciplinary difficulties covered in the section 7.3, including as legislation, ethics, privacy, and security concerns, are critical considerations for any developer working with such powerful and sensitive data as healthcare.

Bibliography

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions", en, *International Journal of Forecasting*, Jan. 2022, ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2021.11.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021001874 (visited on 06/21/2022).

[2] B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "Meta-learning framework with applications to zero-shot time-series forecasting", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 9242–9250.

[3] D. C. Boris N. Oreshkin, N. Chapados, and Y. Bengio, "N-BEATS: neural basis expansion analysis for interpretable time series forecasting", in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=r1ecqn4YwB.

[4] K. G. O. Cristian Challu, B. N. Oreshkin, F. Garza, M. Mergenthaler, and A. Dubrawski, "N-hits: Neural hierarchical interpolation for time series forecasting", *CoRR*, vol. abs/2201.12886, 2022. arXiv: 2201.12886. [Online]. Available: https://arxiv.org/abs/2201.12886.

[5] T. Iwata and A. Kumagai, "Few-shot learning for time-series forecasting", *arXiv preprint arXiv:2009.14379*, 2020.

[6] M. C. Lu, M. Kotelchuck, V. Hogan, L. Jones, K. Wright, and N. Halfon, "Closing the black-white gap in birth outcomes: A

life-course approach", *Ethnicity & disease*, vol. 20, no. 1 0 2, S2, 2010.

[7] A. A. Creanga, C. J. Berg, J. Y. Ko, S. L. Farr, V. T. Tong, F. C. Bruce, and W. M. Callaghan, "Maternal mortality and morbidity in the united states: Where are we now?", *Journal of women's health*, vol. 23, no. 1, pp. 3–9, 2014.

[8] E. A. Howell, N. Egorova, A. Balbierz, J. Zeitlin, and P. L. Hebert, "Black-white differences in severe maternal morbidity and site of care", *American journal of obstetrics and gynecology*, vol. 214, no. 1, 122–e1, 2016.

[9] C. E. Rasmussen, "Gaussian processes in machine learning", in *Advanced lectures on machine learning*, Springer, 2004, pp. 63–71.

[10] J. Eichler, R. Schmidt, A. Hiemisch, W. Kiess, and A. Hilbert, "Gestational weight gain, physical activity, sleep problems, substance use, and food intake as proximal risk factors of stress and depressive symptoms during pregnancy", *BMC pregnancy and childbirth*, vol. 19, no. 1, pp. 1–14, 2019.

[11] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", New York, NY, USA: Association for Computing Machinery, 2015, ISBN: 9781450338325. DOI: 10.1145/2810103.2813677. [Online]. Available: https://doi.org/10.1145/2810103.2813677.

[12] C. Gentry, *A fully homomorphic encryption scheme*. Stanford university, 2009.

[13] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis", *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[14] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[15] A. Mehmood, M. Maqsood, M. Bashir, and Y. Shuyuan, "A deep siamese convolution neural network for multi-class classification of alzheimer disease", *Brain sciences*, vol. 10, no. 2, p. 84, 2020.

[16] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici, *et al.*, "A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain", *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[17] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data", *Frontiers in aging neuroscience*, vol. 11, p. 220, 2019.

[18] M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux, "What's a good imputation to predict with missing values?", *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 530–11 540, 2021.

[19] R. Melzack and J. Katz, "The mcgill pain questionnaire: Appraisal and current status.", 2001.

[20] M. Haanpää, N. Attal, M. Backonja, R. Baron, M. Bennett, D. Bouhassira, G. Cruccu, P. Hansson, J. A. Haythornthwaite, G. D. Iannetti, *et al.*, "Neupsig guidelines on neuropathic pain assessment", *PAIN®*, vol. 152, no. 1, pp. 14–27, 2011.

[21] G. Cruccu, C. Sommer, P. Anand, N. Attal, R. Baron, L. Garcia-Larrea, M. Haanpaa, T. Jensen, J. Serra, and R.-D. Treede, "Efns guidelines on neuropathic pain assessment: Revised 2009", *European journal of neurology*, vol. 17, no. 8, pp. 1010–1018, 2010.

[22] Y. Chu, X. Zhao, J. Han, and Y. Su, "Physiological signal-based method for measurement of pain intensity", *Frontiers in neuroscience*, vol. 11, p. 279, 2017.

[23] D. Lepore, K. Dolui, O. Tomashchuk, H. Shim, C. Puri, Y. Li, N. Chen, and F. Spigarelli, "Interdisciplinary research unlocking innovative solutions in healthcare", *Technovation*, p. 102 511, 2022, ISSN: 0166-4972. DOI: https://doi.org/10.1016/j.technovation.2022.102511. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016649722200058X.

[24] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: Addressing ethical challenges", *PLoS medicine*, vol. 15, no. 11, e1002689, 2018.

[25] H. Freisling, I. Elmadfa, and I. Gall, "The effect of socioeconomic status on dietary intake, physical activity and body mass index in austrian pregnant women", *Journal of human nutrition and dietetics*, vol. 19, no. 6, pp. 437–445, 2006.

[26] S.-K. Ng, C. M. Cameron, A. P. Hills, R. J. McClure, and P. A. Scuffham, "Socioeconomic disparities in prepregnancy BMI and impact on maternal and neonatal outcomes and postpartum weight retention: The EFHL longitudinal birth cohort study", *BMC pregnancy and childbirth*, vol. 14, no. 1, pp. 1–15, 2014.

[27] *Global Women's Health Apps Market - Global Industry Trends and Demand 2021 to 2030*, en-US. [Online]. Available: `https://marketresearch.biz/report/womens-health-apps-market/` (visited on 06/22/2022).

[28] *Women's Health App Market Size Analysis Report 2021-2028*, en. [Online]. Available: `https://www.grandviewresearch.com/industry-analysis/womens-health-app-market` (visited on 06/22/2022).

[29] P. M. Research, *Women's Health App Market Size Is Projected to Reach $9.42 Billion By 2028 | CAGR: 19.1% : Polaris Market Research*, en. [Online]. Available: `https://www.prnewswire.com/news-releases/womens-health-app-market-size-is-projected-to-reach-9-42-billion-by-2028--cagr-19-1--polaris-market-research-301437803.html` (visited on 06/22/2022).

[30] J.-a. P. Hughson, J. O. Daly, R. Woodward-Kron, J. Hajek, and D. Story, "The rise of pregnancy apps and the implications for culturally and linguistically diverse women: Narrative review", *JMIR mHealth and uHealth*, vol. 6, no. 11, e9119, 2018.

[31] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, ACM, 2019, pp. 398–399.

[32] F. Mangialasche, A. Solomon, B. Winblad, P. Mecocci, and M. Kivipelto, "Alzheimer's disease: Clinical trials and drug development", *The Lancet Neurology*, vol. 9, no. 7, pp. 702–716, 2010.

[33] J. Chodosh, D. B. Petitti, M. Elliott, R. D. Hays, V. C. Crooks, D. B. Reuben, J. Galen Buckwalter, and N. Wenger, "Physician recognition of cognitive impairment: Evaluating the need for improvement", *Journal of the American Geriatrics Society*, vol. 52, no. 7, pp. 1051–1059, 2004.

[34] C. Puri, G. Kooijman, F. Masculo, S. V. Sambeek, S. D. Boer, J. Hua, N. Huang, H. Ma, Y. Jin, F. Ling, G. Li, D. Zhang, X. Wang, S. Luca, and B. Vanrumste, "A personalized bayesian approach for early intervention in gestational weight gain management toward pregnancy care", *IEEE Access*, vol. 9, pp. 160 946–160 957, 2021.

[35] C. Puri, G. Kooijman, B. Vanrumste, and S. Luca, "Forecasting time series in healthcare with gaussian processes and dynamic time warping based subset selection", *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6126–6137, 2022. DOI: 10.1109/JBHI.2022.3214343.

[36] C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278.

[37] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 398–399.

[38] C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Gestational weight gain prediction using privacy preserving federated learning", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 2170–2174.

[39] C. Puri, G. Kooijman, X. Long, P. Hamelmann, S. Asvadi, B. Vanrumste, and S. Luca, "Feature selection for unbiased imputation of missing values: A case study in healthcare", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 1911–1915.

[40] C. Puri, S. Keyaerts, M. Szymanski, L. Godderis, K. Verbert, S. Luca, and B. Vanrumste, “Daily pain prediction in workplace using gaussian processes”, *HEALTHINF*, 2023 (accepted).

List of Publications

Articles in Internationally Reviewed Academic Journals

C. Puri, G. Kooijman, F. Masculo, S. V. Sambeek, S. D. Boer, J. Hua, N. Huang, H. Ma, Y. Jin, F. Ling, G. Li, D. Zhang, X. Wang, S. Luca, and B. Vanrumste, "A personalized bayesian approach for early intervention in gestational weight gain management toward pregnancy care", *IEEE Access*, vol. 9, pp. 160 946–160 957, 2021

D. Lepore, K. Dolui, O. Tomashchuk, H. Shim, C. Puri, Y. Li, N. Chen, and F. Spigarelli, "Interdisciplinary research unlocking innovative solutions in healthcare", *Technovation*, p. 102 511, 2022, ISSN: 0166-4972. DOI: https://doi.org/10.1016/j.technovation.2022.102511. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016649722200058X

C. Puri, G. Kooijman, B. Vanrumste, and S. Luca, "Forecasting time series in healthcare with gaussian processes and dynamic time warping based subset selection", *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6126–6137, 2022. DOI: 10.1109/JBHI.2022.3214343

Articles in International Conferences

C. Puri, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Luca, and B. Vanrumste, "Pregdict: Early prediction of gestational weight gain for pregnancy care", in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4274–4278

C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. D. Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Privacy preserving pregnancy weight gain management: Demo abstract", in *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*, 2019, pp. 398–399

C. Puri, K. Dolui, G. Kooijman, F. Masculo, S. Van Sambeek, S. Den Boer, S. Michiels, H. Hallez, S. Luca, and B. Vanrumste, "Gestational weight gain prediction using privacy preserving federated learning", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 2170–2174

C. Puri, G. Kooijman, X. Long, P. Hamelmann, S. Asvadi, B. Vanrumste, and S. Luca, "Feature selection for unbiased imputation of missing values: A case study in healthcare", in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 1911–1915

C. Puri, S. Keyaerts, M. Szymanski, L. Godderis, K. Verbert, S. Luca, and B. Vanrumste, "Daily pain prediction in workplace using gaussian processes", *HEALTHINF*, 2023 (accepted).

FACULTY OF ENGINEERING TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
ESAT-STADIUS
Andreas Vesaliusstraat 13
3000 Leuven
chetanya.puri@kuleuven.be
http://www.stadius.esat.kuleuven.be/stadius