TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

CLINICAL RISK MODELLING WITH MACHINE LEARNING: ADVERSE OUTCOMES OF PREGNANCY

Aki Koivu

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 10 | TECHNICA – INFORMATICA | TURKU 2022

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

CLINICAL RISK MODELLING WITH MACHINE LEARNING: ADVERSE OUTCOMES OF PREGNANCY

Aki Koivu

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 10 | TECHNICA – INFORMATICA | TURKU 2022

University of Turku

Faculty of Technology
Department of Computing
Computer Science
Doctoral Programme in Technology (DPT)

Supervised by

Associate Professor, Tapio Pahikkala
University of Turku
Turku, Finland

Assistant Professor, Antti Airola
University of Turku
Turku, Finland

Professor, Timo Knuutila
University of Turku
Turku, Finland

Reviewed by

Assistant Professor, Smisha Agarwal
Johns Hopkins Bloomberg School of Public Health
Baltimore, Maryland, United States

Assistant Professor, Pekka Marttinen
Aalto University
Espoo, Finland

Opponent

Professor, Mark van Gils
Tampere University
Tampere, Finland

The originality of this publication has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

ISBN 978-951-29-8771-9 (PRINT)
ISBN 978-951-29-8772-6 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)
Painosalama, Turku, Finland 2022

To my lovely wife who supported me. Forever is our today.
To my family, thank you for believing in me.

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Computer Science
AKI KOIVU: Clinical Risk Modelling with Machine Learning: Adverse Outcomes of Pregnancy
Doctoral Dissertation, 118 pp.
Doctoral Programme in Technology
April 2022

ABSTRACT

As a complex biological process, there are various health issues that are related to pregnancy. Prenatal care, a type of preventative healthcare at different points in gestation is comprised of management, treatment, and mitigation of such issues. This also includes risk prediction for adverse pregnancy outcomes, where probabilistic modelling is used to calculate individual's risk at the early stages of pregnancy. This type of modelling can have a definite clinical scope such as in prenatal screening, and an educational aim where awareness of a healthy lifestyle is promoted, such as in health education. Currently, the most used models are based on traditional statistical approaches, as they provide sufficient predictive power and are easily interpreted by clinicians.

Machine learning, a subfield of data science, contains methods for building probabilistic models with multidimensional data. Compared to existing prediction models related to prenatal care, machine learning models can provide better results by fitting more intricate nonlinear decision boundary areas, improve data-driven model fitting by generating synthetic data, and by providing more automation for routine model adjustment processes.

This thesis presents the evaluation of machine learning methods to prenatal screening and health education prediction problems, along with novel methods for generating synthetic rare disorder data to be used for modelling, and an adaptive system for continuously adjusting a prediction model to the changing patient population. This way the thesis addresses all the four main entities related to predicting adverse outcomes of pregnancy: the mother or patient, the clinician, the screening laboratory and the developer or manufacturer of screening materials and systems.

KEYWORDS: Machine learning, Pregnancy outcomes, Risk calculation

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietojenkäsittelytiede
AKI KOIVU: Kliinisen riskin mallinnus koneoppimismenetelmin: raskaudelle haitalliset lopputulemat
Väitöskirja, 118 s.
Teknologian tohtoriohjelma
Huhtikuu 2022

TIIVISTELMÄ

Raskaus on kompleksinen biologinen prosessi, jonka etenemiseen liittyy useita terveysongelmia. Äitiyshoito voidaan kuvata ennalta ehkäiseväksi terveydenhuolloksi, jossa pyritään käsittelemään, hoitamaan ja lievittämään kyseisiä ongelmia. Tähän hoitoon sisältyy myös raskauden haitallisten lopputulemien riskilaskenta, missä probabilistista mallinnusta hyödynnetään määrittämään yksilön riski raskauden varhaisissa vaiheissa. Tällä mallinnuksella voi olla selkeä kliininen tarkoitus kuten prenataaliseulonta, tai terveyssivistyksellinen tarkoitus missä odottavalle äidille esitellään raskauden kannalta terveellisiä elämäntapoja. Tällä hetkellä eniten käytössä olevat ennustemallit perustuvat perinteiseen tilastolliseen mallinnukseen, sille ne tarjoavat riittävän ennustetehokkuuden ja ovat helposti tulkittavissa.

Koneoppiminen on datatieteen osa-alue, joka pitää sisällään menetelmiä millä voidaan mallintaa moniulotteista dataa ennustekäyttöön. Verrattuna olemassa oleviin äitiyshoidon ennustemalleihin, koneoppiminen mahdollistaa parempien ennustetulosten tuottamisen sovittamalla hienojakoisempia epälineaarisia päätösalueita, tehostamalla data-keskeisten mallien sovitusta luomalla synteettisiä havaintoja ja tarjoamalla enemmän automaatiota rutiininomaiseen mallien hienosäätöön.

Tämä väitös esittelee koneoppimismenetelmien evaluaation prenataaliseulonta- ja terveyssivistysongelmiin, ja uusia menetelmiä harvinaisten sairauksien datan luomiseen mallinnustarkoituksiin ja jatkuvan ennustemallin hienosäätämisen järjestelmän muuttuvia potilaspopulaatiota varten. Näin väitös käy läpi kaikki neljä asianomaista jotka liittyvät haitallisten lopputulemien ennustamiseen: odottava äiti eli potilas, kliinikko, seulontalaboratorio ja seulonnassa käytettävien materiaalien ja järjestelmien kehittäjä tai valmistaja.

ASIASANAT: Koneoppiminen, Raskauden lopputulema, Riskilaskenta.

Table of Contents

Symbols and Notation

A	Set of average ensemble weights
α	Learning rate constant of gradient descent
β	Regression coefficient
b	Base of a logarithm
γ	Weighted sum of ensemble probabilities
δ	Fixed point of a normalization mapping for standard deviation of 1, used as a parameter of SELU
ε	Fixed point of a normalization mapping for mean of 0, used as a parameter of SELU
z	Stochastic noise from a given distribution
φ	Activation function of an artificial neural network
Y	Set of response values
X	Set of predictor variables
ln	Natural logarithm
G	Generator network of GAN
D	Discriminator network of GAN

Abbreviations

ACOG	The American College of Obstetricians and Gynecologists
actGAN	Activation-specific generative adversarial network
AE	Average ensemble
AFP	α-fetoprotein
ANN	Artificial neural network
ANOVA	Analysis of variance
ARPS	Adaptive risk prediction system
ART	Assisted reproductive technology
AUC	Area under the curve (ROC)
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
BMI	Body mass index
CDC	Centers for Disease Control and Prevention
CVS	Chorionic villus sampling
DNA	Deoxyribonucleic acid
DNN	Deep neural network
DOHMH	Department of Health and Mental Hygiene
DT	Decision tree
fHCGβ	serum free β-human chorionic gonadotrophin

FPR	False positive rate
GAN	Generative adversarial network
GBDT	Gradient boosted decision tree
GD	Gradient descent
GDM	Gestational diabetes
GpVC	Global p value cutoff
HIV	Human immunodeficiency virus
ICD-10	International Classification of Diseases, 10th revision
IL	Incremental learning
IQR	Interquartile range
IVF	In vitro fertilization
LGBM	Light gradient boosted machine
LR	Logistic regression
ML	Machine learning
MoM	Multiple of the median
NIPT	Noninvasive prenatal screening
NT	Nuchal translucency
NYC	New York City
OR	Odds ratio
PAPP-A	Pregnancy-associated plasma protein
pAUC	Partial area under the curve
PlGF	Placental growth factor
PTB	Preterm birth
ReLU	Rectified linear unit activation function
RF	Random forest
ROC	Receiver operating characteristic
SHAP	Shapley additive explanations
SELU	Scaled exponential linear unit
sFlt-1	Soluble fms-like tyrosine kinase-1
SGD	Stochastic gradient descent
SMOTE	Synthetic Minority Oversampling Technique
SURUSS	Serum, Urine and Ultrasound Screening Study
SVM	Support vector machine
TL	Transfer learning
TPR	True positive rate
T21	Trisomy 21, Down Syndrome
WA	Weighted average ensemble
WGAN	Wasserstein GAN
WGAN-GP	Wasserstein GAN with gradient penalty
WHO	World Health Organization

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Koivu A, Korpimäki T, Kivelä P, Pahikkala T, Sairanen M. Evaluation of machine learning algorithms for improved risk assessment for Down's syndrome. *Computers in Biology and Medicine*, 2018; 98: 1-7.

II Koivu A, Sairanen M. Predicting risk of stillbirth and preterm pregnancies with machine learning. *Health Information Science and Systems*, 2020; 8(1): 14.

III Koivu A, Sairanen M, Airola A, Pahikkala T. Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association*, 2020; 27(11): 1667-1674.

IV Koivu A, Sairanen M, Airola A, Pahikkala T, Leung W C, Lo T K, Sahota D. Adaptive Risk Prediction System with Incremental and Transfer Learning. *Computers in Biology and Medicine*, 2021; 138: 104886.

The original publications have been reproduced with the permission of the copyright holders.

1 Introduction

Pregnancy is an event that contains various complex biological processes aimed at developing and delivering a healthy newborn. Because of this complexity, along with environmental factors affecting the mother, various health issues can rise that relate to the mother or the fetus, at different time points of gestation. Management, treatment and mitigation of such issues during the gestational period is commonly called prenatal care (Fiscella, 1995). Prenatal care is considered a type of preventative healthcare, as it contains periodical visits to a medical professional that can assess the progress and the current state of a pregnancy. These visits can also include the use of risk prediction -based screening procedures, that can be used to manage adverse outcomes.

There are commonly four entities relating to risk prediction of adverse outcomes of pregnancy: the mother or patient, the clinician, the screening laboratory and the developer or manufacturer of screening materials and systems. The role of the manufacturer is to provide software, instrumentation, or biochemical reagent kits to be used by the other entities. The centralized screening laboratory specializes in running the related biochemical tests, and commonly provides results to a clinician with high patient sample throughput. The clinician's role is to communicate and conduct management, treatment, or mitigation with his or her patients. In addition to processing screening results, these clinicians have a significant role in maternal health education, as they can recommend lifestyle choices to the mother that have a beneficial effect on the mother's and child's health during pregnancy.

In the domain of prenatal care, probabilistic modelling is used to assess the individual's risk of adverse pregnancy outcomes (Rose, et al., 2020). Depending on the clinical gravitas of the associated outcome and its treatment and management, the risk assessment can be used by the patient's clinician to determine the most opportune treatment strategy, or by the patient herself for educational purposes. In this chapter, applications of clinical prenatal screening and health education of risk modelling for adverse pregnancy outcomes are described, along with the aims of the study.

1.1 Prenatal Screening

Prenatal screening is an aspect of prenatal care that consists of detecting affected events of the ongoing pregnancy (Rose, et al., 2020). This can concern the health of the mother or the fetus. For the fetus, in the most desirable case this enables timely treatment of the event or the underlying condition before or after birth. However, if treatment is not available, the screening gives the parents a chance to prepare for a baby with a health problem or a disability. In cases of lethal or severely disabling disorder, the pregnancy is commonly decided to be terminated. For the mother, conditions relating to the pregnancy can be detected, such as pre-eclampsia which can be life-threatening for the mother and child if left untreated (Steegers, et al., 2010).

Most common birth defects are related to hereditary genetic disorders (Yoon, et al., 2001). Therefore, prenatal screening commonly refers to aneuploidy or chromosomal abnormality screening. When prenatal screening is deployed, it is commonly done with a large patient population (Rose, et al., 2020). Centralized screening laboratories use affordable and highly sensitive biochemical and biophysical screening tests for conducting prenatal screening, usually to patients within their region. During certain time frames of the pregnancy, patient’s sample is collected and analysed in the lab. The results are then transferred to the patient’s clinician who provides diagnosis, and the appropriate treatment is then arranged for the patient. The screening tests that are commonly done are related to the conditions that the laboratory has selected for screening, and this is guided by the regional prenatal screening programs and guidelines they comply.

The severe implications of a wrong prediction or diagnosis are the reason why prenatal screening is highly regulated. The American College of Obstetricians and Gynecologists or ACOG is an example of an organization that provides guidelines for screening tests conducted in the first and second trimester (Rose, et al., 2020). The regulation for In Vitro Diagnostics or IVD also relates to this domain (Food and Drug Administration, 2011), as they apply to the used screening instruments and reagents.

Screening and diagnosis methods can be categorized as invasive or non-invasive tests (Rose, et al., 2020). Traditionally, invasive diagnosis methods in this context involve probes or needles that are inserted into the uterus, such as chorionic villus sampling or CVS and amniocentesis (Alfirevic, et al., 2017). These methods have a risk of affecting the fetus, and in extremely rare cases can result to pregnancy loss. Non-invasive screening tests on the other hand oppose minimal risk to the success of a pregnancy. These methods are usually based on ultrasonography or maternal serum screens from a blood sample. Invasive methods are usually deployed as a second tier diagnostic test when the initial first tier non-invasive method has given a positive result that cannot provide definitive diagnosis.

Different screening tests can also be time-dependent, meaning that they are feasible in certain time frame of the pregnancy due to development level of the fetus, or time restriction of the applicable treatment. For example, nuchal translucency scan or NT is usually offered around 11 to 13 weeks of gestational age because at that time the fetus has developed enough for the measurement to take place. For prenatal screening, first and second trimester are the main stages of pregnancy when testing is conducted due to time limitations of some management options. The general timeline of a pregnancy is depicted in Figure 1.

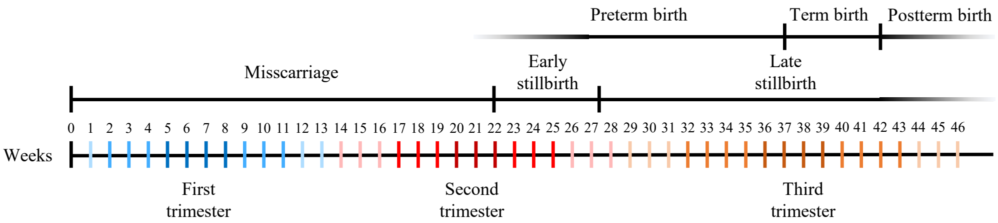

Figure 1. Timeline of a pregnancy. First trimester is defined as the gestation period of weeks 0 to 13, while the second trimester is defined as the period of weeks 13 to 28 (The American College of Obstetricians and Gynecologists, 2020). Pregnancies with live birth that end before the week 37 are considered as preterm births, while term births are considered to occur during weeks 37 to 42 (World Health Organization, 2014). Pregnancies that end in loss of the child before week 22 are considered as miscarriages, during weeks 22 to 27 they are classified as early stillbirth, and beyond week 28 as late stillbirth (World Health Organization, 2014).

Because of the severity and low incidence of rare genetic disorders such as Down syndrome, prenatal screening has adapted the process of detecting pregnancies belonging to the high-risk pregnancies group (James, et al., 2010). Conceptionally, if a screening test result shows elevated risk for a condition, there are treatments and increased monitoring that can be applied before the condition fully takes an effect. The detection of a high-risk pregnancy is commonly an assessment of the prior risk of the patient, which consists of clinically significant risk factors (Parritz & Troy, 2018). These factors consider the maternal demographics and history of the mother. Applicable biochemical and biophysical measurements can then be added to improve the screening performance.

1.1.1 Down syndrome screening

Down syndrome is a genetic disorder caused by the presence of all or part of a third copy of chromosome 21 (Patterson, 2009). Therefore, the condition is also known as Trisomy 21 or T21. The affected individual has three copies of the genes on chromosome 21 instead of the usually two, this is caused by the nondisjunction of the 21st chromosome during egg or sperm development (Reisner, 2013). It is one of the most common chromosome abnormalities due to aneuploidy in humans, the incidence is estimated to be 1 in 1000 (Weijerman & de Winter, 2010). There have been multiple studies that have investigated the association of T21 to maternal history and demographics without concrete success, excluding maternal age (Sherman, et al., 2007). There is no known treatment for T21, and its severity can vary. Also, affected babies have an increased risk of developing other disorders such as heart problems, diabetes and Alzheimer's (Abbag, 2006; Menéndez, 2005). Since its prenatal screening has been introduced, most pregnancies are decided to be terminated by the parents (Orthmann Bless & Hofmann, 2020).

The development of prenatal screening of T21 is heavily intertwined with the development of prenatal screening in general. The study of α-fetoprotein or AFP levels collected from a maternal serum sample at second trimester to detecting neural tube defects conducted in the mid-1970s can be considered as the first significant step (Wald, et al., 1977). In that study, it was found that AFP concentration tended to be lower when the fetus had T21. After this, improvements on the quality of obstetric ultrasound, multiple additional biomarkers measured from the mother's blood sample along with risk modelling and maternal age-specific stratification of risk enabled feasible first trimester screening for T21 (Davis, et al., 2014).

The most widely used first trimester screening test for T21 is also named the combined test, because it incorporates the blood serum sample analysis after 14 weeks of gestation with ultrasound at 10 weeks of gestation (Davis, et al., 2014). Pregnancy-associated plasma protein or PAPP-A (Breathnach & Malone, 2007) and serum free β-human chorionic gonadotrophin or fHCGβ (Ong, et al., 2000) were measured from the blood sample, while NT (Souka, et al., 2005) could be measured from an ultrasound image. This method had a false positive rate or FPR of approximately 5% with the fixed true positive rate or TPR of 85% (Cuckle & Benn, 2010; Cuckle & Maymon, 2016; Padula, et al., 2014). The combined test represents one feasible protocol, as there are various other biomarkers to consider (Cuckle & Maymon, 2016). The widely agreed upon protocol has not emerged however as the scientific community, regulatory bodies and region representatives continue this assessment to this day. Advancements in technology have also been a topic of discussion, as DNA screening or NIPT becomes more and more commercially viable with highly accurate performance as time goes on (Gil, et al., 2015). The high prediction performance of T21 screening with the combined test is the result of highly predictive biomarkers applied to a clear and distinct outcome, which is not the case for health education.

1.2 Health Education

A significant part of prenatal care is also to educate the mother and promote health and safety of the fetus (Alexander & Kotelchuck, 2001). Clinical risk algorithms can be used for this purpose; however, the performance and clinical relevance requirements of a clinical risk algorithm are strict, and usually require extensive validation. This does not however determine the clinical usefulness of a model, as there are widely used prenatal risk models that are not routinely used for clinical risk determination (Olivia Kim, et al., 2019). Patient risk produced by these types of models can be useful in terms of health education.

Detection of a patient belonging to a high-risk group for adverse pregnancy outcomes such as preterm birth and stillbirth in the early stages of pregnancy is essential for their management and ultimately prevention. In such cases, non-invasive monitoring and healthy lifestyle changes can improve the patient's odds for the adverse outcome not to manifest during the pregnancy. Increasing the awareness and educating the mother of these possibilities is crucial. These preventive measures have no apparent drawback to the patient's or the fetus's health, so the threshold for starting management and monitoring is low. This thesis covers two pregnancy outcomes that currently can be considered as health education -related, stillbirth and preterm birth.

1.2.1 Stillbirth

World Health Organization or WHO defines stillbirth as an infant born without signs of life after the gestational age of 22 weeks (World Health Organization, 2014). The week threshold for stillbirth defined by ACOG is 20 weeks (The American College of Obstetricians and Gynecologists, 2009). WHO's ICD-10 classification also defines early and late stillbirth, the former having the gestational age window of 22 to 27 weeks, while the latter is defined by gestational age window of 28 weeks and beyond (World Health Organization, 2014). The epidemiology of stillbirth differs from other well-known fetal conditions, because the outcome can be a combination of multiple causing effects, and they are rarely fully understood (Aminu, et al., 2014). These can include other pregnancy-related conditions such as genetic disorders, infections and structural malformations. The incidence of stillbirth correlates with the income level of a country, because most frequently the reported cause is maternal factors such as malaria, diabetes mellitus, syphilis or HIV positive status (McClure, et al., 2009; Turnbull, et al., 2011), and these conditions are more frequent in low-income countries. In high-income countries the incidence is estimated to be 1-9 in 1000 individuals, while in low- and middle-income countries it is estimated to be 3-30 in 1000 (McClure, et al., 2009; Stanton, et al., 2006). However, globally the rate of stillbirth has slowly declined mostly due to progress in developed regions, while the highest rates and slowest declines are reported in Southern Asia and sub-Saharan Africa (McClure, et al., 2011).

Despite the uncertainty relating to diagnosing stillbirth, several studies have shown that there are risk factors associated with it (Aminu, et al., 2014). Maternal age, ethnicity, body-mass index or BMI, smoking, various substance abuses, low level of education, low socioeconomic class, diabetes mellitus, multiple gestation and previous stillbirth are some of the more significant risk factors. Compared to T21 which has a substantial amount of scientific literature that associates certain biomarkers with the condition, a wide range of potential biomarkers for detecting stillbirth have been proposed (Smith, 2017), but feasible clinical verification is currently lacking for a majority of them, so the scientific community has not come to consensus regarding the subject. Some of the biomarkers with more clinical validity include PAPP-A (Smith, et al., 2002; Smith, et al., 2004), AFP (Smith, et al., 2007), placental growth factor or PlGF and soluble fms-like tyrosine kinase-1 or sFlt-1 (Chaiworapongsa, et al., 2013), along with unconjugated estriol or uE3 (Yaron, et al., 1999). Because of the research activities for finding feasible biomarkers for stillbirth is still ongoing, demographic risk factors are the current standard method for determining risk of stillbirth for educational purposes (Trudell, et al., 2017). The performance of these methods is modest and ranges from 0.64 to 0.67 area under the ROC curve or ROC AUC (Trudell, et al., 2017; Yerlikaya, et al., 2016).

Given the detection of a high-risk pregnancy for stillbirth, there are multiple management and treatment methods for preventing it due to the nature of the multiple underlying causes. One of the most potent treatments proposed in terms of late stillbirth is induced labour around 39 weeks of gestation, because the prevalence of stillbirth is thought to be constant after 24 weeks until term when it starts to increase again (Smith, 2001). This method is however being criticised for resulting in too many interventions compared to the number of prevented deaths. Low doses of aspirin during multiple weeks of gestation has been shown to reduce the risk of stillbirth by 14% (Duley, et al., 2007). Low-molecular-weight heparin has also been suggested as a preventative treatment (Dodd, et al., 2013), however feasible clinical validation of this is currently lacking. Nitric oxide is also thought to have a key role in the control of placental development and its deficit could be one of the causes leading to stillbirth (Abdel Razik, et al., 2016), however clinical validation of this treatment is also lacking. In addition to these, experimental treatments such as supplemental oxygen (Say, et al., 2003) and gene therapy (Spencer, et al., 2014) could become clinically viable in the future.

1.2.2 Preterm birth

Spontaneous preterm birth or PTB is defined as infants born alive before 37 weeks of gestation (World Health Organization, 2014). Babies that are born before term are susceptible to various lethal or disabling outcomes, because commonly the baby is born before their organs are mature enough to sustain life on their own. This problem is amplified in low-income settings where a lack of proper newborn care results in loss of life (Chang, et al., 2013). The incidence of PTB in such countries is estimated to be 1 in 7 births on average, while in high-income countries it is estimated to be 1 in 10 births on average (Purisch & Gyamfi-Bannerman, 2017). Early PTB, before 34 weeks of gestation, is frequently associated with morbidity and mortality such as respiratory distress syndrome, necrotizing enterocolitis, intraventricular haemorrhage and neurological deficits (Martin, et al., 2009).

Maternal history -related risk factors for pregnancies ending in PTB have been proposed in the past, such as multiple previous gestations, history of PTB and prior cervical surgery (Werner, et al., 2011). However, their causality is hard to prove, because PTB can occur to women without elevated risk results (Iams, et al., 2001). Proposed educational risk models for PTB have therefore modest performance, resulting in AUC values of 0.51 to 0.67 with evidence to overfitting (Meertens, et al., 2018). One of the most promising screening tests is measuring the length of the cervix by trans-vaginal ultrasound, the only concrete concern being the subjectivity of the measurement (Werner, et al., 2011). Multiple biomarkers for detecting PTB have been proposed in the past (Considine, et al., 2019; Souza, et al., 2019; Waller, et al., 1996), however widely accepted testing protocol remains to be adopted. This also highlights the fact that similar to stillbirth, the occurrence of PTB can have multiple underlying causes. As for diagnostic tests, fetal fibronectin swab test has been proposed to be used in combination with cervix-length screening for symptomatic patients as a rule-out method (Son & Miller, 2017). Measuring c-reactive protein from amniotic fluid has also been proposed as a functioning, however invasive option (Ghezzi, et al., 2002).

Women that are identified as having high risk for PTB can be targeted for more thorough antenatal surveillance and preventive healthcare (Medley, et al., 2018). These treatments can be highly specified because the underlying cause for PTB can be one of many. In the systematic review by Honest et al., antibiotic treatment for high-risk women due to bacterial vaginosis was found to significantly reduce the occurrence of PTB (Honest, et al., 2009). Also, progesterone supplements (Norwitz & Caughey, 2011), periodontal therapy (Radnai, et al., 2009), fish oil (Harper, et al., 2010) smoking cessation programs (Soneji & Beltrán-Sánchez, 2019) cervical cerclage (Alfirevic, et al., 2017) and pessaries (Arabin & Alfirevic, 2013) showed promising results. In addition to them, non-steroidal anti-inflammatory agents in tocolytic therapies where premature pregnancies are medically suppressed showed most promise, however good-quality evidence was deemed insufficient for tocolytic maintenance therapy (Abramovici, et al., 2012). Lastly, bed rest in a hospital or at home is also a common procedure, as the hard physical activity which is proposed to be associated with preterm delivery is minimized (Sosa, et al., 2004).

1.3 Aims of the study

The goal of the thesis was to find novel methods for advancing the various clinical analysis processes associated with prenatal risk assessment. Different entities relating to risk prediction of adverse outcomes of pregnancy have differing use cases and goals relating to them. Manufacturers want better development tools to provide more accurate and robust prediction models to their customers. Centralized screening labs want more automation without sacrificing performance, and reduce overall costs related to screening. Clinicians want better tools for characterising their patient before deciding what management path to suggest. Patients want more information about their ongoing pregnancy. All these needs were considered by approaching the problem in multiple angles, as each study aims to investigate new approaches to the established clinical analysis workflows. These would include assessing machine learning or ML applicability to different risk assessment tasks and developing novel methods for this particular domain.

In the retrospective T21 study reported in Publication **I**, the focus was to evaluate several ML algorithms that were applied to predicting the risk of T21 from first trimester data. Improving prediction performance with existing predictor variables when compared to a "golden standard" IVD risk prediction software was the main objective. Generating more efficient models would mean less unnecessary testing conducted by the centralized screening lab.

In the retrospective population study for stillbirth and PTB reported in Publication **II**, discovering clinically viable prediction models using only patient's demographics and maternal history was the goal. ML models and ensemble learning were investigated for improving prediction performance. Producing a robust prior-risk model for stillbirth and PTB could improve their detection in the natural world and provide clinicians new tools for creating treatment decisions.

In the retrospective method development study reported in Publication **III**, minority oversampling methods were investigated for their applicability to prenatal risk data. Development of a specialized method for this domain was also considered. Generation of synthetic data that could be added to highly class-imbalanced training data has the potential of creating more robust decision boundaries within the associated models. Manufacturers and researchers of the domain can benefit from the novel method presented in the paper.

In the retrospective study reported in Publication **IV**, incremental -and transfer learning techniques were investigated for prenatal risk assessment. The goal was to propose a novel automated modelling system where given new T21 screening data, the system would adapt to it if it deemed it necessary. This would eliminate the need for the screening lab staff to update their models over time in a manual manner, which requires deep domain expertise. The articles and their themes are summarized in Table 1.

Table 1. Summarization of the themes considered by the publications. The prediction of T21 is considered in **I** and **IV**, PTB by **II** and Stillbirth by **II** and **III**. The applicability of ML methods to clinical prediction was investigated in **I** and **II**, while novel ML methods for the prenatal risk assessment domain was investigated in **III** and **IV**. The related entities of the research are also listed.

Article	Related entity	T21	PTB	Stillbirth	ML applicability	Novel ML methods
I	Screening lab	•			•	
II	Clinician		•	•	•	
III	Manufacturer			•		•
IV	Screening lab	•				•

2 Literature Review

Prenatal risk assessment consists of utilizing predictors with suboptimal outcome discrimination and measurement variance, and data commonly having a significant amount of class-imbalance. The currently deployed domain-dependent methods and data sampling strategies are described in this chapter, along with their machine learning-based alternatives. In addition to this, novel methods applicable to the prenatal risk assessment domain which are enabled by machine learning are also described.

2.1 Domain-Dependent Methods

Data in the domain of prenatal risk assessment commonly refers to tabular datasets and databases, where information is structured in a matrix form that is comprised of columns and rows, or variables and patient records (Howard, 1987). These variables can have multiple data types depending on their origin. Biochemical and biophysical measurements are usually continuous, and some of the mother's demographic information such as weight. Mother's age can be processed as continuous or discrete values. Other demographics such as ethnicity and education are nominal, while maternal history and possible infection status variables are commonly logical.

Statistical analysis in the domain of prenatal risk assessment has specialized over time to a point where "gold standard" methods (Greenhalgh, 1997) have been defined by the research community. Some are a result of research efforts directed at transforming biochemical measurements into a more comparable form over multiple laboratories and patient populations, while some are used due to necessity imposed by the highly imbalanced population data. In this section the most dominant domain-dependent methods of prenatal risk assessment are described.

In screening tests, the ideal design of the measurement is to fully discriminate the affected and unaffected outcomes (Coste & Pouchot, 2003). In this case, determining a classification cutoff value is the only necessary definition for using the test for its purpose. Example of a classification cutoff where elevated measurement would indicate the presence of a condition would be

$$Screening\ Result\ = \begin{cases} Affected & if\ measurement > cutoff \\ Unaffected & if\ measurement \leq cutoff. \end{cases} \tag{1}$$

However, perfect discrimination is rarely achieved in real-world applications. In these cases, the “gold standard” method can be though as the best available test (Versi, 1992). Sources of known and not known variation can affect the result in a way that measurement value distributions of the affected and unaffected populations overlap to some degree. Using a measurement cutoff in this scenario will produce a fixed amount of true positive and false positive results, this fact is visualized in Figure 2.

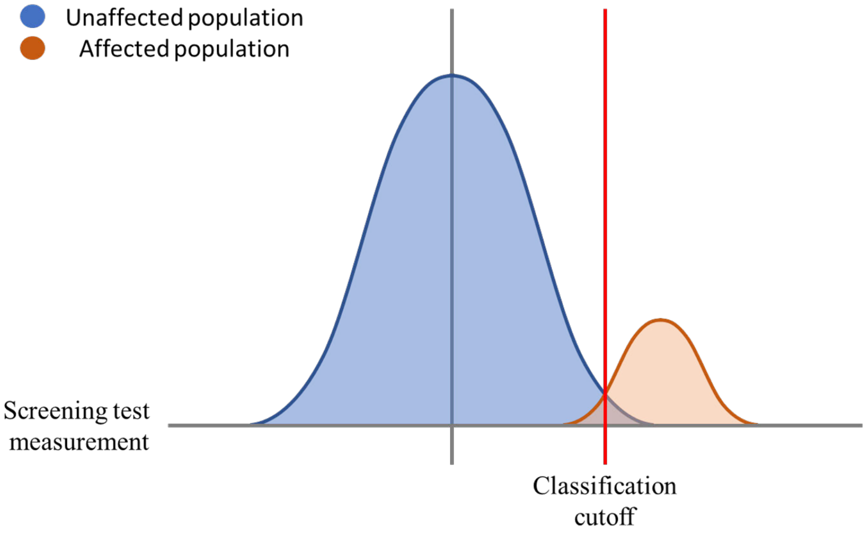

Figure 2. Typical measurement distributions of the affected and unaffected populations with a screening test. The classification threshold provides the decision value where lower values are classified as unaffected, and higher values as affected. Moving the cutoff from left to right provides better FPR at the cost of reduced TPR and moving from right to left would provide better TPR while increasing FPR. As the test doesn’t provide perfect discrimination of the two populations, the optimal classification cutoff for the measurement is determined by the performance requirements generated by the clinical prediction task.

In order to produce more accurate classification results, more information can be considered. If it can be demonstrated that the classification outcome of a screening test has a significant correlation to other factors that are known about the patient, those factors can be used to further improve the screening test’s performance. For the clinical classification algorithm used in a laboratory, one can determine multiple cutoffs for specific sub-populations, for example for each BMI group (World Health Organization, 2004). This type of manual adjustment is feasible to a point where the number of variables and their connections to the outcome stay within practical limits. For anything more involved, statistical analysis methods such as logistic regression or LR can be utilized. This enables the utilization of prediction variables with modest discrimination capability, where multivariate models of the classification problem can be constructed.

Binary LR, henceforth referred as LR, is probabilistic modelling method for binary outcomes (McCullagh & Nelder, 1987). It describes the probability of an outcome as a function of predictor variables

$$P(X) = \frac{b^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}}{1+b^{(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} \tag{2}$$

where the β_0 denotes the estimated intercept, $\beta_1, \ldots, \beta_k$ denotes the estimated slope coefficients, b denotes the base of a logarithm and P denotes a probability of the observation belonging to a category of the binary Y variable. In this context, the Y variable follows the Bernoulli distribution (McCullagh & Nelder, 1987). The LR formula calculates odds by exponentiating log-odds of base b. This formula can be further manipulated into a form

$$P(X) = \frac{1}{1+b^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}} = S_b(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) \tag{3}$$

Where S_b is the sigmoid function with base b. When the intercept and slope coefficients are chosen, the resulting function or a model can be used for binary prediction. This is essentially what fitting an LR implies, and this process can be facilitated with multiple different optimization methods (Minka, 2003).

LR is commonly used in clinical research because the frequent goal in this domain is to understand biological effects and their determinants. The binary outcome of “effect” or “no effect” in drug discovery is a good example of this. Prenatal risk assessment is not an exception to this, as most research papers utilize multivariate analysis or LR (Ylijoki, et al., 2019). The relevant binary outcome in this context is “low risk” or “high risk”, which can also be reported as odds instead of probabilities. LR as a method has been extended multiple times in the past (Wilson, 2015), and the current versions usually deploy some type of regularization when multiple predictor variables are present (Salehi, et al., 2019). Class weighting (He & Ma, 2013) can also be used to make the model aware of the imbalanced classes of the data, and thus improving fitting and performance.

In addition to LR’s applicability, it can also be used intuitively to describe the effect of a predictor to the outcome. The exponential function of a fitted regression coefficient β_k can be described as an odds ratio or OR that is associated with a one-unit increase in the predictor k (Szumilas, 2010). These ORs compare the relative odds of an outcome to the predictor variable of interest, and they are commonly interpreted as risk factors for the outcome in question. These factors can be interpreted as

$$Interpretation = \begin{cases} No\ effect & if\ OR = 1 \\ Predictor\ associated\ with\ higher\ outcome\ odds & if\ OR > 1 \\ Predictor\ associated\ with\ lower\ outcome\ odds & if\ OR < 1 \end{cases} \quad (4)$$

and their magnitude can also be compared. Comparing ORs between different studies is common in prenatal risk assessment research (Kogan, et al., 1994), and some ORs have enough data behind them for the research community to recognize them as standards, which are then established as guidelines (Coppedè, 2016).

The task of predicting risk from a highly imbalanced dataset of prenatal risk assessment results requires a method with high sensitivity. This means that the performance metric for fitting a model also needs to reflect this requirement. Receiver operating characteristic or ROC curve is the golden standard for clinical binary prediction tasks (Obuchowski, et al., 2004). It plots TPR against FPR as a curve. With a binary problem and using some probability cutoff for classification, this curve can demonstrate the trade-off of TPR and FPR when different cutoffs are utilized. The area under of this curve, AUC, can also be calculated, which can be used as a performance metric to determine a feasible prediction model for screening rare occurrences (Huang & Ling, 2005). AUC has a value range from 0 to 1, and it depicts the discrimination of two classes. AUC of 1 represents perfect discrimination within the prediction result for the two classes, while AUC of 0.5 represents an imperfect discrimination which can be compared to random guessing. AUC of 0 on the other hand represents reversed perfect discrimination, where the prediction is always the opposite of the true class. While AUC is favoured in the screening domain where TPR is frequently compared against FPR, it is insensitive to class-imbalance, meaning that proportions of the classes “affected” and “unaffected” are not considered.

Clinical screening laboratories are known for monitoring and reporting their TPR against their FPR over periods of time (De Jesús, et al., 2010). In the field, this is the most straightforward way to measure their performance. In terms of the ROC curve, the laboratories are interested in a limited set of cutoffs with fixed FPR’s for different screening tests. These are sometimes called FPR’s with clinical significance (Bigirumurame & Kasim, 2017). This set of FPR’s is unique for every screened condition because the relevance is determined by the incidence of the condition. For example, clinically significant FPR’s for T21 are narrow 1% to 5% because the incidence of the condition is most currently estimated as 1 in 700 (Mai, et al., 2019). TPR of 100% is meaningless in this context if FPR reaches unacceptably high, because over 99% of the time the true outcome is unaffected. Another aspect for the laboratory to consider when selecting FPR’s is the available resources, finding a screen positive result generates work and the screening laboratory needs to produce results in a timely manner, so that the diagnosis can be promptly made.

In every measurement conducted by a clinical screening laboratory, underlying components that produce variance can be detected (Ichihara, et al., 2008). These can include variation caused by different laboratories, patient populations, lab environments, lab technicians and instruments (Whiting, et al., 2004). To reduce the effect of unwanted variation, multiple of the median or MoM was developed (Wald, et al., 1977). This procedure divides the result of the individual patient with the median measurement of the patient population

$$MoM_{patient} = \frac{Result_{patient}}{median_{patient\ population}},\ interval\colon [0, +\infty) \tag{5}$$

and produces values that show how much the individual patient differentiates from the population. This comprehensible method was first introduced as a method to compare the results from different laboratories and has now become the golden standard for reporting prenatal screening results (Berberich, 2013). Extension of this method is to also adjust the MoM results based on other maternal factors, such as gestational age at the time of sampling and ethnicity (Sprawka, et al., 2011). This can be done by using sub-population medians or by fitting multivariate regression models.

Typical statistical analysis workflow for prenatal data would consist of using descriptive statistics for describing the study population, using univariate or multivariate analysis in order to assess the ORs of predictors and finally produce prediction performance metrics of the finalized model. Variable analysis could also be a two-step process, where the statistically significant ORs would be detected, and a final model would be constructed based on them (Nicholas, et al., 2009).

2.2 Statistical Sampling Methods

Prenatal risk data is highly imbalanced where most observations are a part of the unaffected class due to the incidence of the condition. Data sampling methods that address this limitation by attempting to increase the number of affected observations or reduce the number of unaffected in data are called oversampling and undersampling methods. In this section, the applicable sampling methods for prenatal risk assessment are described.

Routine prenatal data has a majority of unaffected observations. The most straightforward method for rebalancing the two classes is to remove unaffected observations from the training data. This is called majority undersampling (Chawla, 2010). In the case of random majority undersampling, one would remove observations belonging to the majority class without any other criteria. This has the negative effect of removing important observations that significantly contribute to the model's decision boundaries by chance. The more advanced undersampling methods such as K-means clustering (Forgy, 1965) and Tomek links (Tomek, 1976) attempt to address this by using mean- and distance-based observation elimination. Undersampling has been successfully applied to predicting adverse pregnancy outcomes by identifying factors that contribute to PTB (Dong, et al., 2020), predicting gestational diabetes or GDM (Qiu, et al., 2017) and predicting readmission following hospitalization due to hypertensive disorders (Hoffman, et al., 2021).

Aside from model-specific methods such as class weighting (He & Ma, 2013) and focal loss (Lin, et al., 2017), an alternative way to address the class-imbalance is to oversample the minority class observations. The most straightforward way of doing this is to randomly duplicate these observations (Ling & Chenghui, 1998), however duplicated observations do not enrich the training data in a way that more intricate decision boundaries can be modelled. The more involved way of generating new and unique minority class observations is to understand the feature space of the predictors and generate new points within that space. This is essentially what the method Synthetic Minority Oversampling Technique or SMOTE does (Chawla, et al., 2002). By calculating the distance between observations with a metric such as Euclidean distance, new synthetic observations can be generated between real observations in the feature space. SMOTE has been widely applied to different domains (Fernández, et al., 2018), and its methodology has been extended to multiple variations of the algorithm (Han, et al., 2005; Maciejewski & Stefanowski, 2011; He, et al., 2008) . In terms of prenatal risk assessment, SMOTE and its variants have been applied to prediction of stillbirth and miscarriage (Inyang, et al., 2020), PTB (Fergus, et al., 2016) and T21 (Ramanathan, et al., 2018).

2.3 Tree Models and Ensemble Learning

The most common alternative to using LR for constructing prenatal risk models is to use tree-based models (Wallenstein, et al., 2016; Heng, et al., 2014; Shen, et al., 2020). These models deploy a hierarchical tree structure that enables the predictor variable space to be assigned into multiple decision boundary areas (Quinlan, 1986). In this section, most prominent tree-based models that have been applied to prenatal risk assessment are described.

Tree-based learning is a widely applied statistical modelling technique (Podgorelec, et al., 2002). The overall concept of tree learning is to produce a hierarchical and sequential system of rules where conditions form if clauses that progress decision making into the next set of rules or nodes (Quinlan, 1986). These paths form branches, and at the end of them are output decisions or leaves of the model. The simplicity and nonlinear behaviour make tree-based models highly popular for classification and regression tasks. One implementation of this is classification tree, henceforth referred as decision tree or DT (Quinlan, 1986). The core principles of fitting a DT are entropy and information gain. Shannon's Entropy (Shannon, 2001) can be defined as

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i, \ interval: [0,1] \tag{6}$$

where p_i is the probability of a class value i occurrence in the data used for training the model. Entropy is a metric for disorder or uncertainty, used in fitting DT models by minimizing it while constructing the tree. Information gain, also known as Kullback-Leibler divergence (Kullback & Leibler, 1951), on the other hand can be defined as

$$IG(Y,X) = \mathrm{E(Y)} - E(Y|X), interval: [0,1] \tag{7}$$

where the conditional entropy of Y given predictor X is subtracted from the prior entropy of Y. Because information gain describes the information gained by explaining Y with X, it is used to choose the optimal predictor variables and node splits during fitting the DT model until a certain hyperparameter threshold is achieved. After the fitting process is done, the finalized information gain values per predictor variable can be inspected with the method proposed by Breiman (Breiman, 2001) for understanding variable importance of the final model. This is one of the main reasons why DT is also highly favored in medical diagnosis (Podgorelec, et al., 2002). For adverse pregnancy conditions, DT has been applied to classifying high-risk pregnancies (Lakshmi, et al., 2016), hypertension during pregnancy (Moreira, et al., 2017), stillbirth outcome of pregnancy (Malacova, et al., 2020), GDM during pregnancy (Shen, et al., 2020) and PTB (Hill, et al., 2008). The abstract visualization of a DT and its decision areas are depicted in Figure 3.

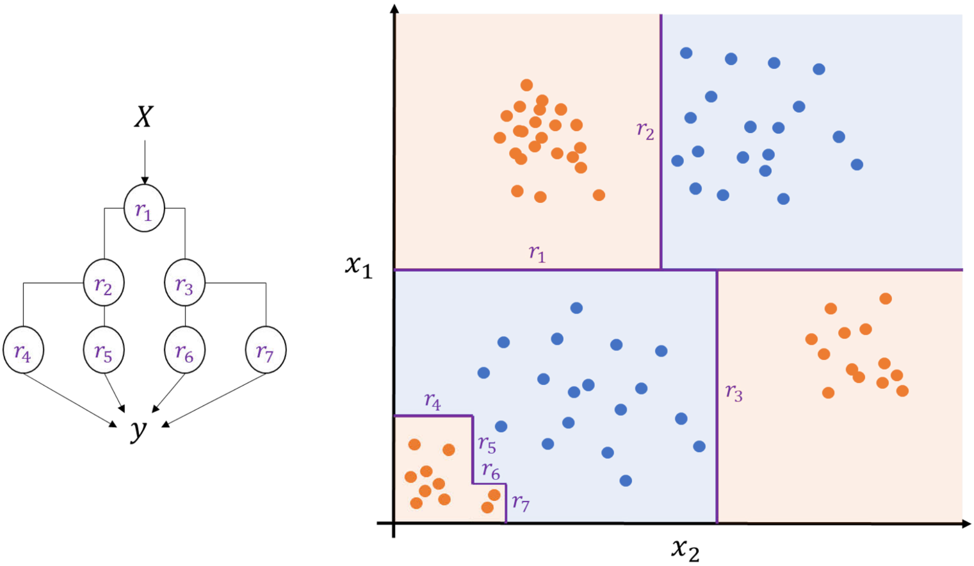

Figure 3. DT's tree structure depicted as a set of rules, which are visualized in predictor variable space of x_1 and x_2. Two classes of orange and blue are classified with the decision regions implemented by the DT's rules.

As an extension of DT, random forests or RF were first proposed by Ho in 1995 (Ho, 1995). In its initial state, the training algorithm would randomly sample the training data multiple times and fit multiple decision trees with them. This is called bootstrap aggregation (Breiman, 1996). When the model would process unseen data to predict them, average of all models would be produced in the case of regression and majority vote in the case of classification. Using multiple independent prediction models to perform superiorly compared to a single model is commonly references as ensemble learning (Piryonesi & El-Diraby, 2020). This strategy was later extended to predictors by Ho in the form of random subspace bagging or feature bagging (Ho, 1998), and it represents the current interpretation of random forests. In feature bagging, predictors are randomly sampled to individual tree models. This is proposed to increase variance of the model without introducing more bias (Ho, 1998). As for adverse pregnancy outcomes, RF has been applied to predicting PTB (Lee & Ahn, 2019), stillbirth (Malacova, et al., 2020), GDM (Shen, et al., 2020), hypertension (Ijaz, et al., 2018), congenital heart defects of the newborn (Luo, et al., 2017), pre-eclampsia (Jhee, et al., 2019) and abnormal pregnancies (Spilka, et al., 2014). Gradient boosted decision trees or GBDT are the next iteration of the tree models after random forests by extending the ensemble learning aspect (Ye, et al., 2009). Boosting is a method of combining weak models into a strong ensemble, where a weak learner is a model that has fitted poorly to the underlying problem. During each iteration of constructing a new parallel tree model, the next model is given the information of what observations were poorly classified by the previous model, in other words the prediction errors. This way, the next model is focused on correctly predicting those hard observations. The first tree model to utilize this type of boosting was called AdaBoost (Schapire, 2013), later the generalization of adaptive boosting to gradient boosting in a form of XGBoost (Chen & Guestrin, 2016) created gradient boosted decision trees, where minimizing an arbitrary differentiable loss function by adding weak learners optimized by a stochastic gradient descent was the core functionality of the algorithm. The abstract description of this process is depicted in Figure 4.

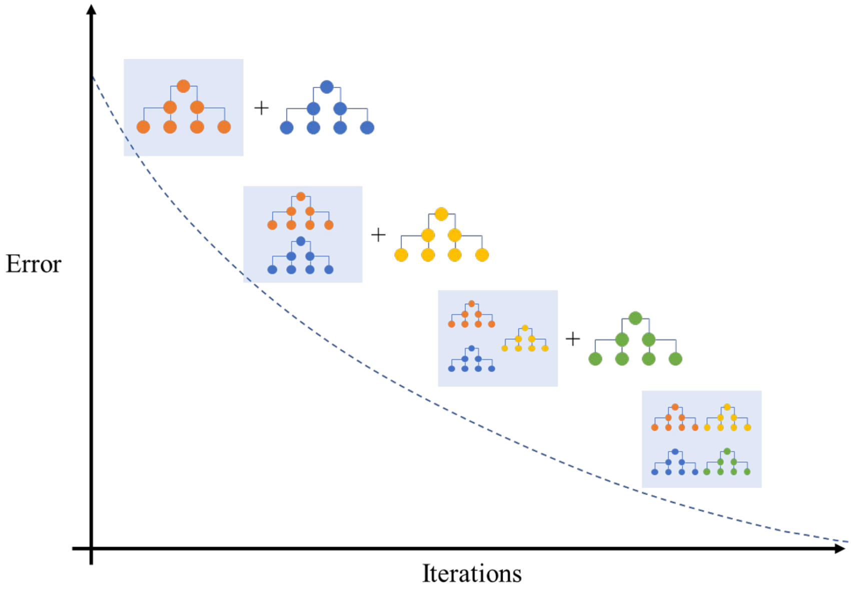

Figure 4. In gradient boosted trees, at each iteration a new parallel tree model is created that is trained on new targets, which are the errors created by the existing ensemble of trees (Inside the blue rectangle). As the ensemble grows after each iteration, the prediction errors shrink, and the model fits further.

This method has been applied to predicting adverse pregnancy outcomes in the form of GDM (Shen, et al., 2020), in vitro fertilization or IVF outcome (Qiu, et al., 2019), and PTB (Malacova, et al., 2020).

Extending ensemble learning beyond tree models can also be done with various methods (Opitz & Maclin, 1999). Ensemble averaging is one strategy that enables the use of multiple different algorithms as one (Naftaly, et al., 1997). Given a set of probabilistic prediction models trained with the same training data and binary classification task, the prediction probabilities of an unseen observation can be averaged. This strategy frequently outperforms any individual models, because various underfitting and overfitting problems of a single model are balanced out. Ensemble averaging can be extended to weighted averaging, where a set Y of predicted probabilities by different models are adjusted by a weight set A, and the weighted sum γ is calculated as

$$\gamma(x, A) = \sum_{j=1}^{p} a_j y_j(x) \tag{8}$$

where p is the number of models in the ensemble. This method produces the need to find a feasible set of weights, which requires ranking the ensemble models by some mechanism. Prior domain knowledge can be used for this, or search strategies such as the exhaustive brute-force, or the genetic algorithm (Mirjalili, 2019).

2.4 Artificial Neural Networks as Classifiers

Artificial neural networks or ANN are not commonly utilized in prenatal risk assessment modelling in the past, mostly due to their unintuitive interpretation, but their benefits can outweigh this fact. In this section, ANN's and their auxiliary methods applicable to prenatal risk assessment are described.

The human brain's main tissue component is nervous tissue that is comprised of nerve cells or neurons (Herrup & Yang, 2007). These neurons can pass electrical and chemical signals to other neurons via a structure called a synapse, and given stimulus they can produce individual or ensemble neural responses (Foster & Sherrington, 1897). While a human learns, the neuronal connections can be altered (structural neuroplasticity), or the properties of neurons are altered (functional neuroplasticity) (Schmidt-Wilcke, et al., 2010). Artificial neural network or ANN is the algorithmic simplification of this complex biological system (McCulloch & Pitts, 1943). In ANN, an artificial neuron or a node can of type hidden, input or output. Given a number of input nodes n, a hidden node k takes the output $x_1, \dots, x_n$ of those nodes as input, then multiplies them with weights $w_0, \dots, w_n$ (dot product), and their sum along with an added bias term b_k is passed to a non-linear function φ

$$y_k = \varphi\left(\sum_{i=1}^{n} x_i w_{ki} + b_k\right) \tag{9}$$

that produces the output y of the node k. φ in this context is called the activation function (Cybenko, 1989). There are several proposed activation functions (Nwankpa, et al., 2018), and most of them are designed to produce nonlinearity to the model by mimicking the action potential signalling of a neuron (Hodgkin & Huxley, 1952). The neuronal learning scheme was first derived as an unsupervised technique (Hinton, et al., 1999) from Hebbian learning theory (Hebb, 2005), and later the supervised learning method perceptron was proposed (Rosenblatt, 1958). This was then later extended to support automatic differentiation by backpropagation (Linnainmaa, 1970; Rumelhart, et al., 1986), which enabled the development of the now widely adopted ANN methods.

As opposed to hidden nodes, input nodes of the input layer do not have similar functionality, as they simply pass input values to hidden nodes. A hidden layer of a neural network can consist of one or many hidden nodes, this is usually referred as the wideness of a layer (Lee, et al., 2020). Output nodes have the same functionality as hidden ones, but their used activation and amount in the output layer are determined by the prediction task (Nwankpa, et al., 2018). For binary problems with a finite output value set of $[0,1]$, commonly a one node output layer with a sigmoid activation is used. The abstract structure of an ANN is depicted in Figure 5.

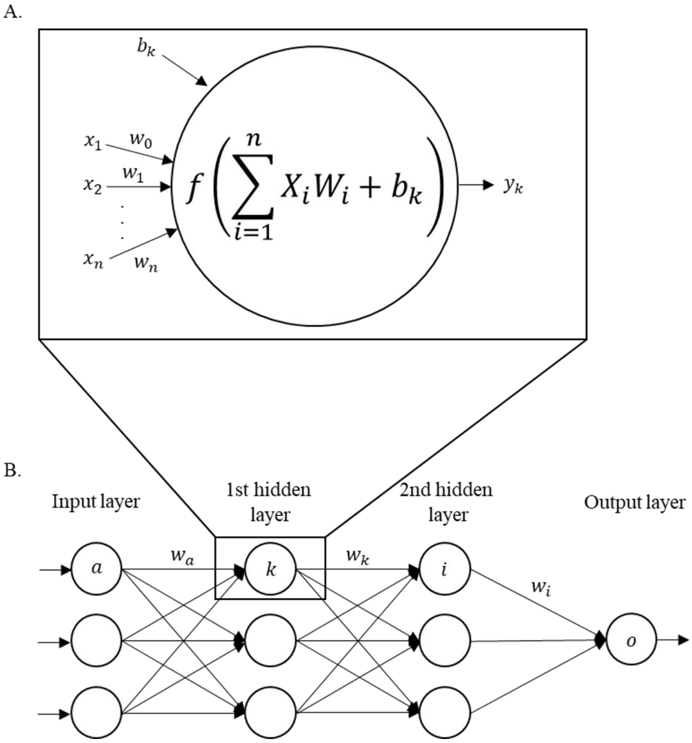

Figure 5. Subplot A depicts a hidden neuron or node k that is feeded inputs $x_1, \dots, x_n$ weighted with $w_1, \dots, w_n$, which are added together along with a bias term b_k, and their sum is passed to the activation function f. The resulting y_k is feeded to the 2nd layers nodes as input. Subplot B depicts the common network design of an ANN, which contains an input layer, 1,…,n hidden layers and an output layer.

The initialization of weights and biases for the network is usually stochastic, and during fitting the model they along with bias are updated. This is done using the method called backpropagation (Linnainmaa, 1970). In the case of a binary classification task, when the true class y of the training data observation k is known, the error of the predicted value p can be calculated with using cross entropy or log loss (Murphy, 2012) as a loss function

$$CE_k = -(y_k \ln(p_k) + (1 - y_k) \ln(1 - p_k)) \tag{10}$$

Where ln is the natural logarithm. The sum of cross entropy over all training data observations is the value that backpropagation needs to minimize. The partial derivatives of the loss function with respect to any weight or bias in the network are calculated, and then the weights and biases can be updated by the product of the learning rate constant α and a partial derivative with respect to the cost function. This algorithm is called gradient descent or GD that iterates through the whole training data before weights updates are calculated (Boyd & Vandenberghe, 2004). Currently, the most popular version of GD is stochastic gradient descent or SGD (Bottou, 1998) with mini-batches (Bertsekas, 1996), as it is computationally less expensive at deriving heuristic optimization results. Given an objective function that is differentiable or subdifferentiable, instead of calculating the gradient from the whole data set as in GD, the data set is partitioned randomly into subsets. These subsets or mini-batches are then used individually to calculate the gradient.

A fully connected ANN that consist of one input, output and hidden layer and uses a step function for output is commonly referenced as single-layer perceptron (Auer, et al., 2008). It is considered as the predecessor of modern neural networks. The full connectivity in this context means that each node of network is connected to all the nodes in the previous layer (Hastie, et al., 2009). The advantage of this is that no prior understanding of the prediction problem and its data is necessary. If one or more hidden layer to the design of an ANN is added, the depth of the network increases, and it becomes a multi-layer perceptron which can be considered as a deep neural network or DNN (Hastie, et al., 2009). Compared to less complex modelling techniques such as LR, DNN can provide more intricate decision areas that can have multiple local optima. If the data to be fitted has enough complex and high dimensional structures from which DNN fitting can benefit from, it will outperform conventional models such as LR with high probability. On the other hand, because of this complexity, the interpretability of a DNN model is worse when compared to LR, where coefficients or OR's can be inspected.

The progress of ANN's applied to tabular data has not been as rapid when compared to neural networks that are applied to more complex data types such as images and video (Voulodimos, et al., 2018). While currently the most promising methods for tabular data seem to be variations of gradient boosted decision trees, advancements such as the self-normalizing neural network or SELU network have been made (Klambauer, et al., 2017). This type of ANN utilizes scaled exponential linear unit or SELU

$$SELU(x) = \delta \begin{cases} x & if\ x > 0 \\ \varepsilon \exp(x) - \varepsilon & if\ x \leq 0 \end{cases} \tag{11}$$

as activation functions for hidden nodes. The parameters $\delta\ \approx 1.0507$ and $\varepsilon \approx 1.6733$ are solved for the equations resulting from finding a fixed point for the mapping from the mean and variance of activations from one layer to another, using the Banach fixed point theorem (Banach, 1922). They represent standard deviation of 1 and mean of 0 respectively in the normalization mapping. In addition to this, a dropout method revised by the authors of SELU needs to be used, as commonly used dropout would not result in the desired mean and variance of the activations (Klambauer, et al., 2017).

For utilizing ANNs with highly imbalanced prenatal data, in addition to using data-related methods such as under -and oversampling, cost-sensitive learning can be used (He & Ma, 2013). This can mean designing the loss function to consider the uneven importance's of the affected and unaffected observations or apply weights to the error scores based on the class of the observations, which is more straightforward. In this context, the affected class is assigned the larger weight for error, because its correct prediction result is more important. Fully connected ANNs have been proposed for prenatal risk tasks of T21 (Williams, et al., 1999), PTB (Fergus, et al., 2016; Zernikow, et al., 1998) and congenital heart disease (Li, et al., 2017).

2.5 Artificial Neural Networks as Generators

Recently, an extension of ANN's called generative adversarial networks or GAN have been developed for the task of generating synthetic data based on training data (Goodfellow, et al., 2014). In this section, the formulation of GAN and its relevance to clinical data is described.

The GAN framework was proposed by Goodfellow et al. in 2014. The method consists of training a generative neural network model G along with a discriminative neural network model D in an adversarial manner. The learning task of G is to map stochastic noise derived from a distribution to observations in the training dataset. Successfully fitting this network will then produce realistic synthetic observations by feeding it random noise. The learning task for D on the other hand is to determine if the input observation is truly from the training dataset or has it been produced by G. Successful fitting of D would produce a model that perfectly discriminates real and synthetic data. These two models are trained together with a two-objective loss function that can be formalized as

$$\min_{G} \max_{D} L\,(D, G) = \mathbb{E}_{x\,\sim\,\mathrm{P_r}}[log(D(x))] + \mathbb{E}_{\tilde{x}\,\sim\,\mathrm{P_g}}[log(1 - D(\mathrm{G}(\tilde{x})))] \tag{12}$$

where P_r is the training data distribution and P_g is the model distribution, defined by

$$\tilde{x} = G(z), z \sim p(z) \tag{13}$$

where z is the noise input sampled from the stochastic distribution p. This type of training can be compared to a zero-sum game between the networks G and D. After a successful training of both models, the discriminator network is no longer utilized, and the generator is used to produce synthetic observations. The conceptual description of the GAN framework is depicted in Figure 6.

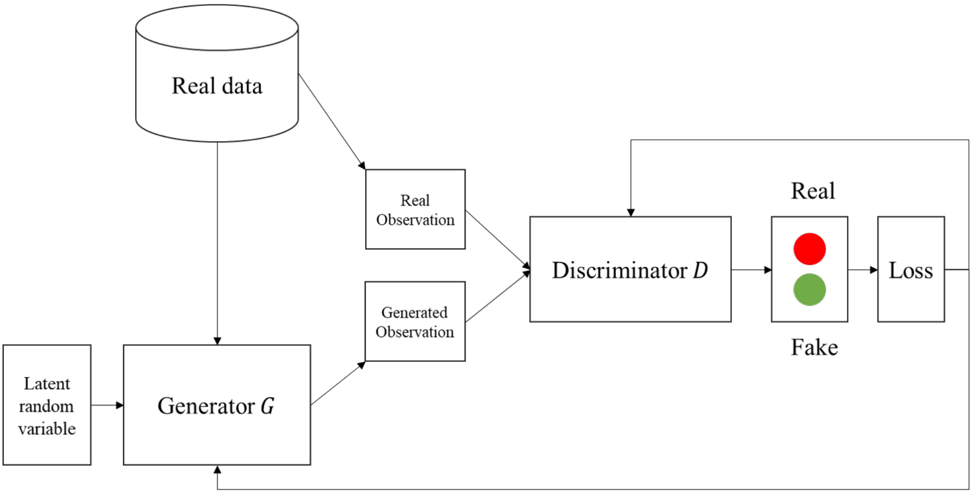

Figure 6. The generator network G is fitted to map random values to training observations of real data, so that it can generate synthetic observations. The discriminator network D is then fitted to differentiate these observations from the real ones. These two networks are trained in unison, better fake observations are generated by G while they are more accurately discriminated from real ones by D. This results in a G that can create synthetic observations highly similar to real data, while D is discarded.

Training GAN's proposed by Goodfellow et al. without alterations have been proven to be fragile and unstable, as problems such as mode collapse and diminishing gradient of the generator arise (Arjovsky & Bottou, 2017). In 2017, extension of GAN called Wasserstein GAN or WGAN was proposed to address these problems (Arjovsky, et al., 2017). The major difference was to replace the discriminator network D with a critic network C that scores observations as being real or synthetic by learning a K-Lipschitz function to compute Wasserstein distance between the probability distributions of real and fake samples (Fournier & Guillin, 2015). Decreasing this function during training implicates that the resemblance of the G network output to true training data increases, the loss function is then defined as

$$W(P_r, P_g) = \frac{1}{K} \sup_{\|f\|_L \leq K} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_\mathrm{r}} [f(x)] - \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_\mathrm{g}} [f(\tilde{x})] \tag{14}$$

where sup is supremum and K is the Lipschitz constant for function f, and is made to satisfy $||f||_L \leq K$, or K-Lipschitz continuity (Hager, 1979). WGAN provided much needed stability to model training, however preserving the K-Lipschitz continuity opposed challenges, as weight clipping was utilized to limit weight updates to a small value range.

WGAN was further improved upon by Gulrajani et al. as they proposed WGAN with gradient penalty or WGAN-GP (Gulrajani, et al., 2017). Gradient penalty replaced weight clipping for the method of maintaining K-Lipschitz continuity during weight updates. A differentiable function f can be considered 1-lipschitz if and only if it has gradients everywhere with a norm of at most 1. In order to achieve this, the loss function was redesigned to increase the generated loss if the gradient norm moved away from 1. This new loss was defined as

$$W(P_r, P_g) = \frac{1}{K} \underset{\|f\|_L \leq K}{sup} \underset{x \sim \mathbb{P}_r}{\mathbb{E}}[f(x)] - \underset{\tilde{x} \sim \mathbb{P}_g}{\mathbb{E}}[f(\tilde{x})] + \lambda \underset{\hat{x} \sim \mathbb{P}_{\hat{x}}}{\mathbb{E}}[(\|\Delta_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{15}$$

where $\mathbb{P}_{\hat{x}}$ is sampled from generator distribution $\mathbb{P}_g$ and data distribution $\mathbb{P}_r$ with t uniformity between the range of 0 and 1, so that

$$\hat{x} = t\tilde{x} + (1 - t)x \text{ when } 0 \leq t \leq 1 \tag{16}$$

And λ is the penalty coefficient. This modification stabilized model training even further and removed the need to set a hyperparameter for weight clipping (Gulrajani, et al., 2017).

As of now, several variations of GAN methods have been proposed (Jabbar, et al., 2020). While these have been more widely applied for image and video-based tasks, they are applicable to clinical tabular data that is used in prenatal risk assessment. The major problem to solve in this research space is to make the generator network aware of the different data types of variables, as discrete variables should have different generation rules when compared to continuous ones. GAN designs that take this to account have been proposed (Xu & Veeramachaneni, 2018; Xu, et al., 2019).

Clinical data applications of GAN networks have also been proposed. Data-driven research based on electronic health records could be improved by generating less noisy and more complete data sets (Rashidian, et al., 2020). Availability of more confidential clinical data could also be improved by GAN networks (Allen & Salmon, 2020). GAN-generated data can also be used to improve model performance for clinical prediction tasks (Lazaridis, et al., 2021). Development of feasible and useful GAN methods in the domain of clinical risk assessment represents the most contemporary research, as there is currently limited scientific literature available.

2.6 Incremental and Transfer Learning

The applicability of a model from one patient population to the next is crucial for its clinical significance. Prediction models that function feasibly within one dataset have little value if it cannot generalize to other datasets. Often the lack of generalizability of a model is caused by the insufficiency of the training data (Therrien & Doyle, 2018). The real world is constantly changing and a method that tries to model it should also constantly adapt to it. In this section, the covariate shift problem that also affects prenatal risk data is described, along with modelling techniques that try to solve it.

Changes in the general patient population over time are evident (Gebremariam, et al., 2018). The natural temporal change of a population alters the relationship of the dependent or outcome and independent or predictor variables (Sugiyama, et al., 2007). Training data will always be a snapshot over a certain time frame, and a probabilistic model fitted to it will also learn only the observed predictor-outcome relationship of that data. The prediction performance of that model will deteriorate over time as the temporal change affects the predictor variable's value distribution. This phenomenon is called covariate shift (Sugiyama, et al., 2007) due to temporal change, and it is evident in population screening tasks such as prenatal risk assessment.

Probabilistic modelling that addresses the fact that data becomes available as a function of time has had many names in the past, online learning, continual learning and incremental learning for example. The two main topics of research in this domain have been algorithms that work with unavailable data due to physical memory constraints of a computer, and algorithms that work with unavailable data due to temporal availability (Bottou & LeCun, 2004). Incremental learning algorithms have been proposed to address covariate shift due to temporal change, as they adapt to new available data without completely forgetting existing knowledge. Often, the relevance of old training data can be parameterized. As for ANN's, the commonly used mini-batch SGD technically uses incremental batch learning for gaining computational advantage compared to other GD methods (Bottou, 1998), so in essence it supports incremental learning. The amount of training epochs and possible decay terms need to be parameterized correctly, however. An abstract description of incremental training GD in batches is depicted in Figure 7.

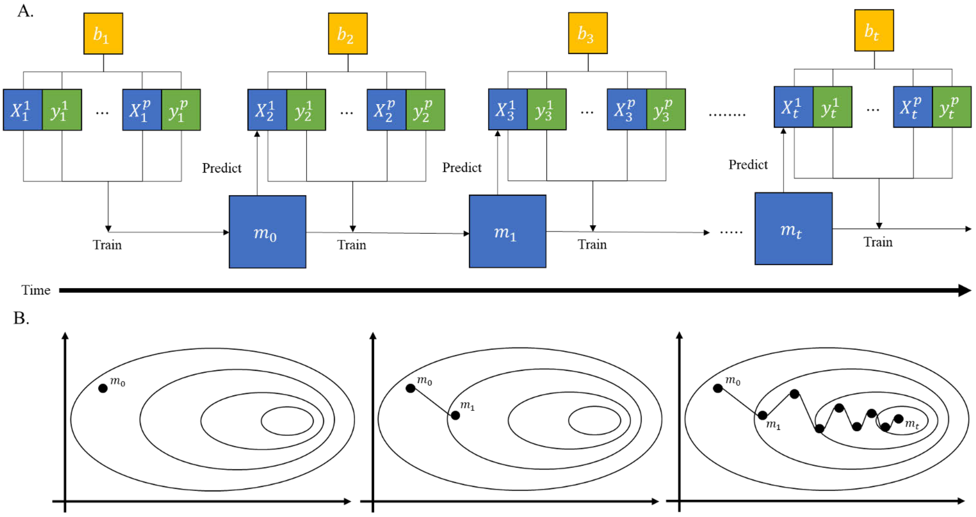
A.
b_1
b_2
b_3
b_t
X_1^1 y_1^1 … X_1^p y_1^p
X_2^1 y_2^1 … X_2^p y_2^p
X_3^1 y_3^1 … X_3^p y_3^p
....
X_t^1 y_t^1 … X_t^p y_t^p
Predict
Predict
Predict
Train
m_0
Train
m_1
Train
.....
m_t
Train
Time
B.
m_0
m_0
m_1
m_0
m_1
m_t

Figure 7. The incremental batch learning process with GD over multiple iterations in time. Subplot A demonstrates that during the first iteration, no model is yet present, so b_1 data batch is used for training model m_0. After this, m_0 is used for predicting the next batch, which is followed by incorporating batch b_2 into the training process, which results into model m_1. This model uses the next batch for prediction, and after that for training. As more time passes, more data is available, which enables better model fitting produced with GD. Subplot B showcases these training iterations as points in a nonlinear optimization space that is traversed by GD.

Covariate shift due to real-world limitations can also occur. An example of this would be a prediction model trained with enough training data but within one geographical location. This limitation causes the model to perform poorly when used in another location that has a significantly different patient population. Therefore, multi-center and multi-institutional studies are important as they produce richer datasets where population characteristics and artefacts of a single study site can be described. These are however resource-heavy endeavours, so often pilot or case studies are done with a selected population and geographical location. Prediction models of these studies can perform well within their study data and even with unseen data from the same population, but often the models struggle or fail completely when they are applied to another patient population.

Modelling techniques that can adapt from population to population, or more generally from domain to domain have been proposed (Torrey & Shavlik, 2010). This is called transfer learning, where existing knowledge can feasibly be applied to new problems. This improves the computational efficiency of modelling; instead of collecting an exhaustive dataset of the new problem, partial dataset can be enough as a model that has been fitted to a similar problem now only has to adapt to minor discrepancies between the two problems. Transfer learning can address covariate shift due to real-world limitations. Prediction models developed in another geographical location with a distinct patient population could be adapted to a new patient population with partial training data. Transfer learning with ANN's commonly amount to fixing or freezing the hidden layer weights and biases, removing the output layer of the network, adding one or more hidden layer along with a new output layer, and fitting the weights and biases of the new layers with the dataset of the new problem (Torrey & Shavlik, 2010). This enables leveraging existing information about the previous problem in the form of fitted weights and biases of the hidden layer nodes, while adapting to the new problem by fitting new layers. Also, not freezing and removing layers and simply continuing model fitting is called fine-tuning. This is commonly done when a pretrained model is used for the same task but fine-tuned to a new data set. An abstract description of both TF and fine-tuning is depicted in Figure 8.

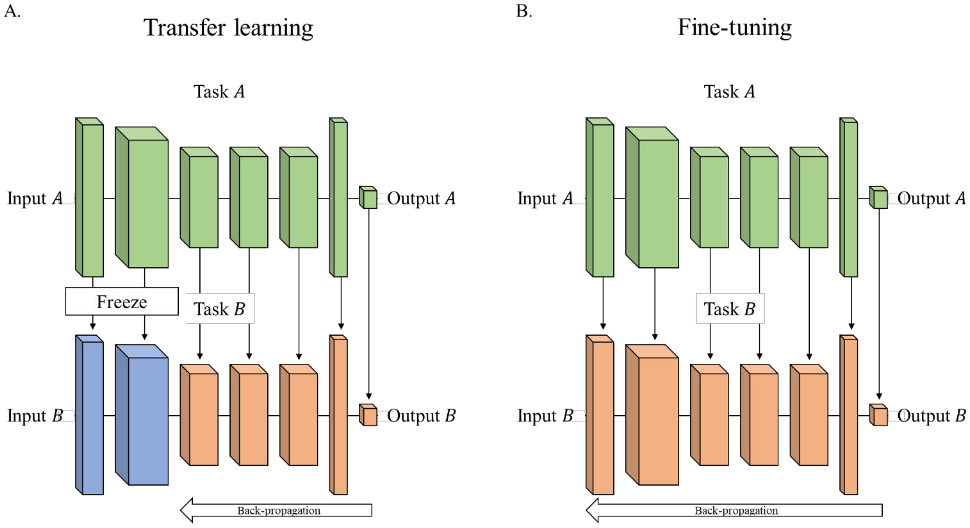

Figure 8. The process of transfer learning and fine-tuning with ANNs. In subplot A, given a completed ANN model for task A, the input layer and some of the layers are fixed or frozen. This way the back-propagation process does not update the weights and biases of these layers, thus retaining their fitted information during training for task B. In subplot B, the whole pretrained model is used for back-propagating weights while training for task B.

When incremental and transfer learning are utilized to adapt to new unseen information, often the amount of adaptation needs to be addressed. The well-known stability-plasticity dilemma relates to this, as it addresses the key constraint of learning by biological and artificial neural systems (Carpenter & Grossberg, 1987). It states that a learning system needs plasticity in order to integrate new information, but also stability retain previous learned knowledge. There seems to be a "golden mean" of the two where old data is not being constantly forgotten and new data is being adapted to. The two extremes are the undesirable learning outcomes of a system; too much plasticity results to catastrophic forgetting of previous essential information and too much stability causes the entrenchment effect where older information is more important compared to newer (Ratcliff, 1990).

When the learning parameters related to IL and TL are adjusted for the prediction problem and the temporal aspect of data availability, automation of the adaption could be achieved. A system designed to automate modelling that reacts to new training data in a meaningful way, by evaluating the necessity of a model update could be deployed to a clinical setting such as a prenatal screening lab safely. This would be done using IL and an evaluation for updating, which would be based on clinically significant performance increase. TL would enable the usage of prior prediction models trained with different populations. In an optimal situation, a new screening lab could immediately start the usage of an existing NN prediction model instead of waiting for the completion of their own sufficient data set, and via this automated system the model could be incrementally updated as screening data would be generated by the lab. This increases operational availability of the prediction model and requires less manual oversight by the lab staff.

3 Materials and Methods

This chapter contains a summary of the materials and methods used in this study. Complete descriptions can be found in the original Publications **I-IV**.

3.1 Datasets

3.1.1 Publication I

Dataset for the first publication (**I**) was a combination of data from three clinical studies, which originally was collected to assess the risk prediction performance of routine T21 screening. This data was retroactively analysed for modelling purposes. Rights to analyse and publish the results were given to two individual datasets from Canada and one from the UK by clinical collaborators. All the studies were approved by local ethics committees, also the participants of the studies gave their informed consent for the whole data life cycle of collection, analysis and publication of results. The data was also irreversibly anonymized.

The combined dataset was used to evaluate the benefits of using machine learning modelling techniques in prenatal risk calculation for T21. Compared to the typical population incidence of T21, the number of positive cases is highly over-represented in the dataset. This enriched case population was caused by the data collection method of the original studies and enabled feasible usage of machine learning with a limited dataset. For training the candidate models, two of the three datasets were used while the third was used for model evaluation. Also, during the training process, k-fold cross-validation (Hastie, et al., 2009) was used to test the generalizability of the chosen hyperparameters.

Variables for the analysis were selected based on the predictor set of the benchmark method, which in our case was LR with the parameters that were established during the SURUSS study (Wald, et al., 2003). This set of predictors contained maternal history, demographics and T21-relevant biomarkers NT thickness (Souka, et al., 2005), PAPP-A (Breathnach & Malone, 2007) and fHCGβ (Ong, et al., 2000). AFP (Tomasi, 1977) and PlGF (Shibuya, 2008) were also partially present in the data, although not utilized due to incomplete data. Summaries of the three study datasets are listed in Table 1 of Publication **I**.

Before any modelling could take place, the study data was preprocessed so that it would applicable for modelling. Predictor variables were inspected to having enough values not missing, AFP and PlGF measurements were excluded because of this reason. Missing values of NT, PAPP-A and fhCGβ were sparse, so they were imputed with zero values. Missing ethnicities were recoded as “Other/Unknown”, while missing smoking status was replace with “No”. Weight, gestational age and maternal age values that were missing were imputed by their mean values. For the biomarker measurements, MoM values and raw concentration values were available, however MoM values were selected based on association tests to the outcome, which in this case were linear correlation and chi-squared tests. The tests compared both versions of the measurement to the outcome of T21 by calculating Cramer’s V and R^2.

3.1.2 Publication II

Dataset for the second publication (**II**) consisted of two routinely collected infant birth datasets that contained the outcomes of the pregnancies. The first dataset contained reported pregnancies during the years 2013 to 2016 in the United States, and it was released by the Centers for Disease Control and Prevention or CDC. We accessed this data via their National Vital Statistics System (Centers for Disease Control and Prevention, 2019). The data originates from the yearly reported birth and death certificates. CDC anonymizes data before publicly releasing it, and our study complied with their data user agreement. The second dataset contained yearly reported pregnancies from 2014 to 2016 in New York City. It was requested from the New York City Department of Health and Mental Hygiene or NYC DOHMH. Similarly to the CDC dataset, this data is also based on the reported birth and death certificates. IRB approval was obtained for the analysis of the dataset, and it was anonymized by NYC DOHMH.

The two datasets were used for constructing predictive models for PTB, early and late stillbirth. The incidences for PTB and infant death were 9.6% and 0.58% in the CDC dataset, while for NYC data they were 8.7% and 0.15%. These incidences correspond to the reported national averages of PTB and infant death in the United States (Purisch & Gyamfi-Bannerman, 2017). The larger CDC dataset was used for training the models, while NYC data was used for model evaluation.

Biomarker measuring is not routinely reported by the hospitals that provide the certificates, so the study data was limited to infections, maternal history and demographics. Initial variable selection was made based on existing literature (Trudell, et al., 2017; Yerlikaya, et al., 2016; Kayode, et al., 2016; Purisch &

Gyamfi-Bannerman, 2017), along with pragmatic reasoning. After this, predictor sets for the models were selected by utilizing correlation and univariate analysis. Chosen variables for the study are listed in Table 1 of Publication **II**.

Inclusion criteria for the observations consisted of six rules. Mothers would have to be 18 of age or older, cases of maternal death were excluded, multiple birth pregnancies were excluded, fetal death outcomes which were reported to being caused by external causes were excluded, postnatal death outcomes were excluded and pregnancies with reported alive babies with less than 21 weeks of gestational age were excluded. The final observations counts are listed in Table 2 of Publication **II**.

The CDC study data was divided into four different sets for four different purposes; feature selection data, training data, validation data and testing data. They were used for conducting feature selection, training the models, validating the models while training and evaluating the finished models respectively. This was an applicable approach since the amount of data was substantial. The NYC dataset was added to the testing dataset, this way also the generalizability to other data could be evaluated. Class-imbalance had to considered when splitting the data, so class-stratified split was used to split the CDC data into partitions of 10%, 70%, 10% and 10% for feature selection, training, validation and testing respectively. This established individual sets for all the different phases of the analysis. The final partitions are listed in the Table 2 of Publication **II**.

3.1.3 Publication III

Dataset for the third publication (**III**) was the aforementioned NYC dataset. It was used to develop a novel minority oversampling method, with the focus of applying it to predicting early stillbirth within the dataset. Incidence for early stillbirth in the data was 0.04%, or 1 in 2500. This significant class-imbalance reflects the routine clinical situation where the method was aimed for and was therefore applicable for experimentation.

Predictor set selection for early stillbirth prediction was made based on existing literature (Trudell, et al., 2017; Yerlikaya, et al., 2016; Kayode, et al., 2016) . Chosen variables for the study are listed in Table 1 of Publication **III**. Same inclusion criteria were used as what is depicted in chapter 3.1.2. The data was randomly splitted in a class-stratified manner into two equal-sized sets, one for method hyperparameter optimization and fitting the finalized model, and one for model evaluation. Preprocessing was done to each variable according to their data type. One-hot encoding was used with nominal features and continuous variables were standardized by unit-variance and zero-mean normalization. Parameters for the preprocessing were calculated using the hyperparameter dataset, and after this they were applied to both datasets.

3.1.4 Publication IV

Dataset for the fourth publication (**IV**) was collected by the Hong Kong Hospital Authority Universal Down Syndrome screening program. First trimester T21 screening data was acquired from a screening database, which was provided by the Obstetrics Screening Laboratory of The Department of Obstetrics and Gynaecology of The Chinese University of Hong Kong. The time frame of the acquired data was from July 2011 to June 2019. All participants signed an institutionally approved consent form, which was specific to screening aneuploidies. The audit and analysis related to the pregnancy outcomes of the women undergoing aneuploidy screening was authorized by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (CREC Ref No. 2012.538). The observations were irreversibly anonymized before the analysis took place.

The dataset was used for a retrospective analysis of the use of an adaptive risk prediction system, as it provided a real-world example of screening T21 from a population. The incidence of T21 was calculated being 0.22% or roughly 1 in 440, which is more than the anticipated T21 incidence of 1 in 700 (Mai, et al., 2019). Characteristics for the study population are listed in Table 1 of Publication **IV**.

As the origin of the data was a T21 screening program, the relevant biochemical and biophysical measurements were included alongside with demographics and maternal history. NT, PAPP-A and fHCGβ concentrations and MoM derivatives were present in the data. Fetal crown rump length or CRL was also present, which was used for determining gestational age using a previously published Chinese dating formulae (Sahota, et al., 2009). For a proper comparison against the laboratory’s existing method, the same set of feature variables was used. Preprocessing used in this retrospective study for the dataset was chosen to be minimal, because it was already validated and standardized against the referral form of the laboratory, at the time when the initial data collection was conducted. Characteristics for the variables are listed in Table 1 of Publication **IV**.

Due to the nature of the experimentation, the dataset was partitioned multiple times into different length data blocks. These blocks were iterated one by one by the adaptive risk prediction system, which would simulate real-world usage of the system over some time period. Data block sizes representing one day, one week, one month, one quartile, half year and one year were estimated from the study data, and they are presented in Table 2 of Publication **IV**.

3.2 Risk Modelling

3.2.1 Publication I

Seven different machine learning methods were used to assess the risk of T21 (Figure 2D, publication **I**). These included K-nearest Neighbour, Decision Tree, Random Forest, Naïve Bayes, L2-regularized LR, Support Vector Machine and Feed-forward ANN. The finalized design of all the methods were found after rigid heuristic experimentation with the study data. Along with most feasible hyperparameters, several versions of the modelling methods were experimented with. Naïve Bayes was investigated with and without Laplace smoothing (Sorkine, et al., 2004). Support vector machines with different kernel functions and penalty factors were investigated. Lastly, for the feed-forward ANN, different activation functions and network structures were investigated. The complete description of the used methods is listed in publication **I**.

All of the models were used in conjunction with the k-fold cross-validation procedure. Because of this, tuning the parameters within the training data was possible, as all of the tested algorithms were experimented upon to produce better finalized models for the test data evaluation. The appropriate k value for our study data was determined to be 10, and the median value of ROC AUC calculated from all folds was used as the result for parameter tuning. Algorithms that produced a feasible level of performance in the cross-validation process were then used to produce risk predictions from the test data set, this represented independent evaluation of classification performance per model. The resulting AUC scores were used to compare against the predicate method of the study, the results of LifeCycle™ risk assessment software. In addition to comparing AUC's, TPR's calculated at FPR's of 1, 3, 5 and 8% were also compared against the predicate. The reason for this was to cover the relevant FPR range for prenatal screening of T21, which is derived from the incidence of the predicted outcome. The summary of results is listed in chapter 4.1, while the comprehensive description can be found in Publication **I**.

3.2.2 Publication II

Correlation and univariate analysis were conducted for the partitioned feature analysis dataset. In the former, all predictor variables were tested in the combinations of two for linear dependency. This was due to the fact highly correlated predictors have the same effect on the response variable (Hall, 2000). Correlations of less than -0.5 and more than 0.5 flagged for the removal of the other variable. By doing this, redundancy was removed from the data set, which should result in more stable and accurate modelling. The univariate analysis consisted of using LR to construct a binary prediction models for all study outcomes, so that individual ORs with their confidence intervals of 2.5% and 97.5% and p values could be determined. The univariate analysis was not however used for feature selection in the case of ML, because it was noted that beneficial feature dependencies found and utilized by ML models were not necessarily detected with LR-based analysis.

After the determination of feature variables, binary risk modelling of affected and unaffected in the case of PTB, late stillbirth and early stillbirth with different algorithms was conducted. These included LR, GBDT algorithm called LGBM (Ke, et al., 2017) and a two different deep fully-connected ANN, details are listed in Publication **II**.

LGBM was chosen to represent the tree models because according to its author, it provided a concrete increase in execution speed without losing significant amount of accuracy (Ke, et al., 2017), which was relevant because of the data size of the study. Different TPR values at clinically significant FPR's were experimented as an alternative, but they did not provide any significant improvements when compared to AUC.

Two ANN models were considered. The first was a Leaky ReLU-based (Maas, et al., 2013) deep two-layer feed-forward ANN, referenced as the deep ANN henceforth, that we had previously shown in Publication **I** to perform in a feasible manner in the task of predicting risk of T21. The second one was deep four-layer feed-forward self-normalizing neural network, referenced as the SELU network henceforth. It was designed to use the scaled exponential linear units or SELU activation function in its hidden nodes, which the author of the SELU network has shown to achieve superior performance when compared to other prominent feed-forward ANN versions (Klambauer, et al., 2017). The author of SELU network also suggests that architectures which are deeper produce better results, so instead of using two layers like in our first ANN model proposed in Publication **I**, four hidden layers were used with the SELU network. For every hidden layer, the number of nodes was set to be the same as the number of predictor variables, and all of the nodes contained the SELU activation function.

Class-imbalance was present in the study data, so the training data set was used to derive class weights w with

$$w = s/(c * f(y)) \tag{17}$$

where s is number of samples, c is the number of predicted classes and $f(y)$ is the frequency of said classes in data labels y. These class weights were utilized with all of the modelling algorithms. Cross validation was not deemed necessary with the substantially large independent test data set.

Ensemble learning of the different types of models was also considered. As there were two ANN models which were experimented with, one of them will be chosen for ensemble learning based on their individual performance. The average and weighted average strategies or AE and WA were experimented in order to test if different modelling methods with differing structures, priors and assumptions would complement each other and ultimately produce superior risk predictions. In AE, prediction probabilities of the multiple models were averaged together, creating a new ensemble prediction. In WA, the set of probabilities created by different models or y is used with a set of predetermined weights or α to calculate the weighted sum $\tilde{y}$ with the formula

$$\tilde{y}(x; a) = \sum_{j=1}^{p} a_j y_j(x). \quad (18)$$

Since we didn't possess any prior information on the optimal or even sub-optimal set of weights to be used, all possible weight combinations were calculated using an exhaustive grid search, where the objective function was to maximize AUC of the ensemble prediction result. From there, the most optimal set of weights would be selected for WA. Results summary is listed in chapter 4.2, and the comprehensive description can be found in Publication **II**.

3.3 Data Augmentation

3.3.1 Publication III

For oversampling the minority class of a mixed-type data set, a variation of the SMOTE algorithm was used called SMOTE-Nominal Continuous or SMOTE-NC (Chawla, et al., 2002). Our study dataset contained continuous feature variables and nominal feature variables. The design of SMOTE-NC considers this by containing a separate logic for nominal features, which is designed to specifically penalize differences in nominal features. Theoretically, this should produce more precise synthetic features if they are nominal. SMOTE-NC was selected to be the benchmark in our study, as it represents the current standard.

In order to generate synthetic observations of the affected class that take into account the variable-specific constrains which are present in tabular mixed-type data, the design of WGAN-GP was altered for this purpose. The most common application of GANs is image data, where one variable is in the form of a pixel. Compared to tabular data, a set of pixels is more straightforward to work with, since they are essentially continuous variables with no specific constraints. In our domain, mixed-type data is more challenging, as it can contain both nominal and ordinal variables that are affected by explicit rules, for example the value of a variable should be a non-negative integer value. Preprocessing can also introduce rules to the data, for example one-hot encoding produces multidimensional representations that can have conditional properties. One example of this would be a stochastic vector, where the representation should always add up to 1. Given enough training data, the generator model of a GAN will learn all of the imposed rules. However, if there is not a sufficient amount of data, such as in a minority oversampling situation, the rules affecting the data cannot be learned in a feasible manner. To overcome this problem, we proposed an alteration of WGAN-GP where the output layer is activation-specific, or actGAN.

Build upon the WGAN-GP architecture which combines gradient penalty with Wasserstein loss, actGAN was designed to learn and generate mixed-type data in a minority learning setting. In actGAN, different variable types dictate the activation functions of the output layer neurons in the generator model. For generating continuous and discrete variables, the method uses the RELU function (Agarap, 2018)

$$ReLU(x) = \begin{cases} x \text{ if } x > 0, \\ 0 \text{ otherwise,} \end{cases} \tag{19}$$

Which structure guarantees that only non-negative values are produced. Binary variables on the other hand were created with the logistic function

$$f(x) = \frac{1}{1+e^{-x}} \tag{20}$$

as its output can be interpret as a binary value in a feasible manner. One-hot encoding representations were deemed appropriate for nominal features, they were generated with the softmax function (Agarap, 2018)

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{21}$$

where for $i = 1, \dots, K$ and $\mathrm{z} = (\mathrm{z}_1, \dots, \mathrm{z}_\mathrm{K}) \in \mathbb{R}^\mathrm{K}$. Softmax is useful in this context because its structure ensures that the output vector's σ(z) elements sum is 1. Customizing the output layer's activation functions made it possible to inject domain knowledge which is known in prior of the generated variables to the model structure, which enabled actGAN to achieve a better model fit when a limited amount of training data was available. Scaled exponential linear unit (SELU) functions that utilize LeCun normal weight initialization (Klambauer, et al., 2017) were used in the hidden layer nodes. The design of the critic network was kept simple on purpose, because the task of discrimination is simpler when compared to generation. Random noise input would be used with the finalized generator in order to generate synthetic observations. The generator and critic model designs are depicted in Figure 1 of Publication **III**.

Data generation methods were compared by inspecting the added prediction performance they provide. This was done by choosing two different classifier models that would use the generated training data for the prediction task. LR which is commonly used in ML benchmarking and a NN -variant called the SELU network were selected, the latter representing the state of the art of fully-connected feed-forward ANN's. During testing, hyperparameters were also tuned appropriately. Cost-sensitive learning was also utilized in the form of class weights due to the magnitude of class-imbalance found in the study data. Chapter 4.3 contains the summary of results, and the full description can be found in Publication **III**.

3.4 Adaptive Risk Prediction System

3.4.1 Publication IV

When designing a method for the detection of distribution shifts, different sources of variance needed to be addressed. Mitigable sources of variance are usually related to factors related the sample measurement of biomarkers. In a clinical setting, to common ones are laboratory-to-laboratory, instrument-to-instrument and operator-to-operator variance (Munson & Rodbard, 1978). In addition to this, the biochemical -and physical testing can be affected by temporal variance, such as seasonal effects. To a varying degree, these sources are addressed by the MoM procedure (Wald & Nicolaides, 1976), which is commonly used by clinical entities for the reduction of laboratory-to-laboratory variance, and thus making their results more comparable. However, it is not specifically designed to address temporal variance sources. If the population median used by the MoM procedure is updated in a regular manner, this can actively reduce seasonality. The MoM procedure, while applicable to any continuous variable, is commonly only used for biochemical -and physical measurements (Bishop, et al., 1993), as the standardization of the most predictive risk prediction features is deemed most important by most clinical entities.

For a risk system to adapt to changes evident in the data over some period of time, they need to be detected first. A feasible method of detection should consider every feature variable, regardless of their data types. In the task of predicting risk for T21, used features are commonly continuous and categorical variables. For continuous features, testing for differences in two distributions can be done by inspecting the medians with a nonparametric Mood's median test (Siegel & Castellan, 1988). This test is more applicable for our problem when compared to one-way ANOVA, because the assumptions related to sample variance are more relaxed (Howell, 2002). Distribution shape is another aspect of the data that can be monitored with a nonparametric two sample Kolmogorov-Smirnov test (Stephens, 1974) by comparing the cumulative distributions. For categorical data, Chi-square test of independence used to a contingency table can be utilized (Siegel & Castellan, 1988). Our distribution shift detection method utilizes all of the three aforementioned tests with the feature variables of suitable data type. The method is described in Table 3 of Publication **IV**.

The distribution shift detection method which utilizes the three tests produces a set of p values as a result. In order to reduce the type 1 error produced by calculating multiple statistical tests, the values are adjusted with the Bonferroni correction (Bonferroni, 1936). The resulting set of adjusted p values is then compared against a cutoff value, which determines their significance in terms of detecting a shift event. This global p value cutoff or GpVC is a tunable parameter in our method, and it determines how sensitive the determination of significance is. If one or more feature is found to have a significant shift event, the prediction system fits a candidate model with the currently iterated data set. At this point, it is assumed that any of the detected events are caused by underlying populations changes, and not from other sources such as equipment failure. The shift detection method is naïve in a sense it does not use any prior information about the data, nor does it address the clinical relevance of the found differences. The latter is considered later by the model updating mechanism in the prediction system, which is triggered by the distribution shift detection method. The proposed schema of the method for detection distribution shifts is found in Figure 2 of Publication **IV**.

When a shift event is found by the detection method, the prediction system constructs a candidate model from the available historical data. This is done according to one of the data processing strategies. The cumulative strategy utilizes all of the available historical data during fitting, as opposed to the windowed strategy which limits the used training data to a certain sample of the historical data, determined by the data block size. These two strategies implement two different points in the stability-plasticity scale (Carpenter & Grossberg, 1987). The most stable way of conducting IL would be to use the cumulative data strategy, since all historical data retained. The windowed strategy on the other hand forces plasticity in a form of moving training data window. Experimentation with both strategies was done with our proposed prediction system.

Deep fully-connected ANN (DNN) was used as the basis of our learning system, and it had similar architecture as our best performing T21 risk model, proposed in Publication **I**. This was due to the demonstrated improvement of the performance in the T21 risk prediction task, when the comparison was done against a commercially used T21 algorithm implemented in the LifeCycle™ software. Within our adaptive risk prediction system or ARPS, the new candidate model would be compared performance-wise to the latest existing prediction model. The performance metric related to this model update rule needs to be appropriate for fitting a model with rare disorder screening data. In this type of data, a significant class-imbalance is commonly found. This is caused by the incidence of T21 in a patient population, which is estimated being roughly 1 in 700 (Mai, et al., 2019). Due to this, prior literature of this topic commonly reports TPR's at clinically significant FPR's, along with the more general ROC AUC (Obuchowski, et al., 2004). In order to represent the clinically significant FPR range, the partial AUC or pAUC (Dodd & Pepe, 2003) of 0% to 10% FPR was selected for our ARPS performance metric. Average precision from a precision-recall curve (Zhang & Zhang, 2009), F1 score (Powers, 2020) and plain AUC were also investigated initially, but they were ruled out due to inferior performance when compared to the selected pAUC of 0% to 10% FPR. The proposed ARPS is showcased in in Figure 3A of Publication **IV**.

It should be noted that the first model that ARPS would generate and deploy to production would probably perform in a suboptimal way. This is caused by the small training data amount (small throughput laboratory or small data block size), or because that the training data simply contains a small amount of T21 cases due to its rare occurrence. To avoid this “cold start”, TL can be used with the ANN models. Existing models that are fitted to a different population or to a different clinical problem all together can be used as the backbone for fitting. We have proposed a T21 risk model published in the past, which was included in Publication **I**. In our experimentation regarding TL, this model was used as the backbone model. The training was done in an IL manner, where the backbone would be used for inference in the first time point, and then retrained in the next time points with the local data. It should be noted that this combination of TL and IL is similar to the research topic of domain adaptation (Redko, et al., 2019), where the utilization also imposes restrictions for the used data. For TL to function properly, the data was formatted in a similar way. Categorical variables needed to contain same levels, this meant that ethnicities of South and East Asian were recoded as Asian to attain compatible standardization across the two data sets. The proposed method for utilizing TL is showcased in Figure 3B of Publication **IV,** also a more complete description of our used model architecture is listed in the Supplementary material of Publication **IV**.

The experiments in Publication **IV** are reported in two phases: firstly, the parameters relating to distribution shift detection and the two data processing strategies were investigated. Secondly, the evaluation of differing ARPS systems risk prediction performance was done. During the first phase, the relationship of data block sizes and GpVC values and detected distribution change events was investigated. This was done for both data processing strategies. In the second phase, these strategies and the utilization of TL were investigated, and finally evaluated against the predicate methods. The testing done in the second phase showcased which IL is more advantageous for our prediction task, a system with high stability or plasticity. Also, the usefulness of TL was assessed, and lastly the performance of the automatic ARPS was compared to see if it can match the performance of the predicates. It should be noted, that in our study the assumption was made that the outcome of the patient is available in the modelling environment at the same time as the predictor information. However, this is not the case when screening is conducted in real-life. The real patient outcome will most probably arrive to the screening lab with some delay, as the entities conducting screening and providing the clinical outcome are usually different. If our proposed ARPS would be implemented to a laboratory, the detection of the distribution shifts would function similarly, while there would be some delay in fitting the candidate models, as this step would be pending for the arrival of the patient outcomes. The list of hardware and software libraries used to construct ARPS for this study is listed in the Supplementary material of Publication **IV.** Chapter 4.4 summarizes the results of Publication **IV**, which contains the full description of the results.

4 Results and Discussion

The summary of the results for Publications **I-IV** are presented and elaborated on in this chapter. Full descriptions can be found in the corresponding publications, and their supplementary materials.

4.1 Publication I

The cross-validation procedure with the study training data was done, the results are highlighted in Figure 9. For Naïve Bayes, it was found that the best performance was achieved when Laplace smoothing was used. It was speculated that since the smoothing ensures that posterior probabilities never reach zero which improves the model's generalization with unseen data values, and that the test set contained the ethnicity value of "Other / Unknown" which was not present in the training data, the smoothing enabled better generalization and thus better performance.

For SVM, ANOVA kernel function (Wahba, 1990) and the penalty factor C (Cortes & Vapnik, 1995) were chosen for the finalized model as they presented superior performance with the test data set. Based on the commonly used radial basis kernel (Cortes & Vapnik, 1995), the ANOVA kernel has been shown to perform well with multi-dimensional data used for non-linear estimation (Stitson, et al., 1997), a result that was replicated in our experiments with our data. Figure 2A of Publication **I** highlights the findings on the hyperparameters related to the ANOVA kernel.

It should be noted that for the feed-forward ANN, Leaky ReLU activation function (Maas, et al., 2013) was selected because it demonstrated better performance with the test data set when compared to the others. The SGD was used as the weight and bias optimization algorithm during fitting and learning rate decrease or decay was utilized. It can be seen in Figure 2C of Publication **I** that the low decay of learning rate contributed towards a favourable ROC AUC, and that if we increase the number of epochs iterated during fitting the model, it does not increase the performance in a meaningful way. In addition to this, prediction performance decreased with the utilization of node dropout if and only if the amount of dropout was set to over 50%. Also, increasing the decay of learning rate speeds up this phenomenon. This is highlighted in Figure 2B of Publication **I**. Three ANN model architectures were experimented with; NN1 which had a modest number of hidden nodes for both layers, NN2 with substantial amount of hidden nodes while node dropout was used and the training was 100 epochs long, and lastly NN3 which was similar to NN2 but with only 20 epochs during training. The used parameters for the three finalized ANN models are listed in Table 2.

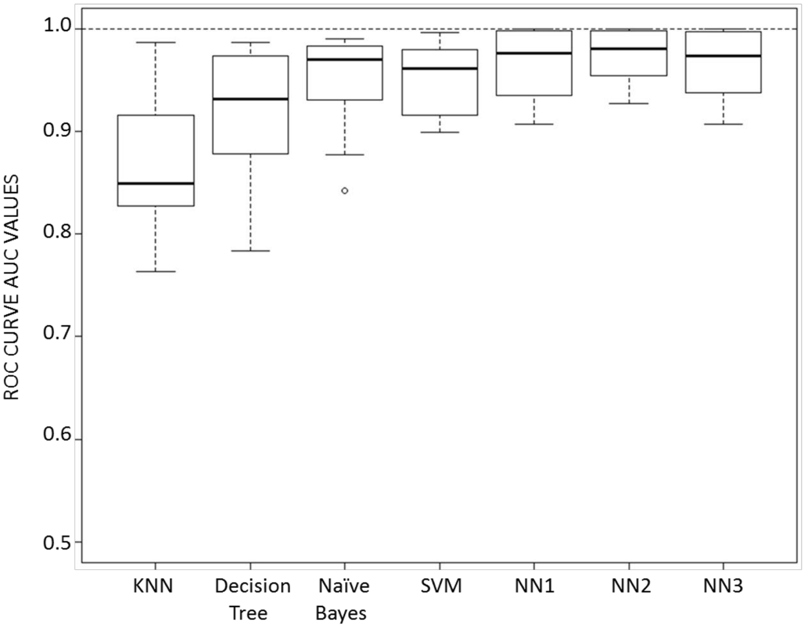

Figure 9. Results produced by the 10-fold cross-validation as boxplots. Median AUC values from training phase for KNN, Decision Tree, Naïve Bayes, SVM, NN1, NN2 and NN3 algorithms were 0.85, 0.93, 0.97, 0.96, 0.97, 0.98 and 0.97, respectively. From Publication **I**.

Results from the cross-validation were that the best performance was achieved with the ANN models; the resulting median AUCs were 0.97 or better (Figure 9). SVM and Naïve Bayes also produced good results with median AUCs of 0.96 and 0.97, but the latter had one serious outlier of 0.84 during the cross-validation. LR also performed adequately by producing a median AUC of 0.97, but the cross-validation produced an outlier of 0.89 for it. It was concluded from these results that the SVM, LR and NN models were used for the comparison of LifeCycle™ classification results in the test data evaluation phase.

These results confirmed the previous findings where SVM and ANN models performed more advantageously in the domain of prenatal risk assessment when compared to KNN, Decision Tree and Naïve Bayes models (Uzun, et al., 2013). This could be because the subtle non-linear relationships between the demographic information and the tested biomarkers are captured in more detail. To avoid local minima and overfitting during the model optimization of classification error, careful design of the optimization parameters and the usage of techniques such as node dropout for ANN's were deemed important if not crucial while analysing data of this size.

Table 2. Used model hyperparameters and the resulting performance metrics from different algorithms with the test sample set, presented as AUC from ROC and TPR at one, three, five and eight FPR values. From Publication **I**.

Algorithm	Attributes		TPR at			
		AUC (CI 95%)	1% FPR	3% FPR	5% FPR	8% FPR
LC4.0	Population parameters from the SURUSS study	0.96 (0.94 - 0.98)	66%	78%	85%	87%
SVM	ANOVA Kernel, sigma=0.31, C=0.057	0.95 (0.93 - 0.98)	61%	75%	81%	86%
NN1	30-20, a=1, dec=0.1, ep=20	0.95 (0.92 - 0.98)	68%	78%	86%	89%
NN2	80-70, a=1, dec=0.1, drop=0.5, ep=100	0.96 (0.94 - 0.99)	72%	82%	85%	92%
NN3	80-70, a=1, dec=0.1, drop=0.5, ep=20	0.96 (0.93 - 0.98)	78%	80%	84%	88%

The results of the test data evaluation are listed in Table 2. The ANOVA SVM, LR and NN1 models performed slightly worse when they were compared to LifeCycle™, however NN1 had similar TPR's at low FPR's. When compared, NN2 and NN3 models demonstrated comparable AUC values, and they also achieved significant improvements in TPR's at multiple different fixed FPR cut-offs. Comparing different FPR results, NN2 model performed moderately better overall, however TPR at 1% FPR was worse. Because NN2 and NN3 differ only by how long the fitting was parameterized with epochs, the results would indicate that the number of epochs in ANN training contributed to the TPR at low FPR's. This is then compensated in higher FPR's, which ultimately results to similar results in terms of ROC AUC. This effect is also showcased in the ROC curve shape in Figure 3 of Publication **I**.

The current "golden standard" model of T21 risk assessment utilizes decades worth of domain knowledge and multi-site studies, and they are commonly LR-based (Verweij, et al., 2013). While ML-based models for this domain have been proposed in the past (Westreich, et al., 2010; Uzun, et al., 2013; Neocleous, et al., 2016) , they don't seem to be favoured by clinical entities. The speculation is that the modelling methods have not advanced into more involved and contemporary algorithms for two reasons; firstly the interpretability of such methods is usually weaker when compared to commonly used LR, and the primary users of such methods are laboratory workers that favour straightforward solutions in general. The second reason is that these new methods can be thought of having miniscule improvement in prediction performance. In our study, this is related to the primary limiting factor of relatively small number of observations. The benefit of utilizing deep learning algorithms is fully manifested only when sufficiently large training data set is used. However, the results we produced indicate that even with a limited data set, one can fit an ML model that is either on par or superior to the current "golden standard".

Experimentation also revealed that the number of hidden layers mattered more when compared to the number of nodes within one layer. There is prior literature that supports this claim (Chiu, et al., 1996). The biggest improvement was from advancing from one hidden layer to two. The ANN structure should be re-investigated when new predictor variables are added, for example in the case of the discovery of a new predictive biomarker. ANN experimentation should also be present during this discovery work, as biomarkers that might not initially appear useful or predictive when analysed with traditional methods can be beneficial with ANN fitting.

While prediction of T21 with the information from the combined test is fairly optimized to a point of saturation, improving FPR while maintaining the "golden standard" TPR has relevance. Small improvements in population screening can produce significant cost savings for the clinical entity involved. An example of this would be that in a hospital with 30,000 annual prenatal screens, reducing FPR from 3% to 1% with our method would mean 600 fewer unnecessary invasive procedures, which amounts to roughly 450000€ in saved costs when using numbers from (Chitty, et al., 2016). Future work relating to T21 ML models would include developing intuitive explainability methods for clinicians to use, and adding new biomarkers as more predictor variables would support the usage of deep learning algorithms even further. This way predictive performance in a prenatal screening program could be improved while the overall cost would be kept minimal.

4.2 Publication II

From the proposed feature variables, mother's BMI was selected overweight with the results from correlation analysis. The utilization of assisted reproductive technology or ART and infertility drugs were correlated to infertility treatment, this is straightforward as they are alternative forms of this type of treatment. Other significant correlations, i.e. less than − 0.5 or more than 0.5, were not found.

The univariate analysis revealed that for the prediction of early stillbirth with the selected variables, 18 out of the total 26 had a statistically significant OR's. Table 4 of Publication **II** demonstrates that these variables with notable OR's were risk factors, along with ART, infertility treatment and marital status. For late stillbirth, the analysis showed that 14 variables out of the total had a statistically significant OR. Notable significant feature variables were also risk factors, along with ART, infertility treatment and marital status. For the prediction of PTB, the infection status of Hepatitis B was the only one not found to be statistically significant. Same variables with notable OR were found when compared to the two other outcomes, but it seems that infections have a bigger association in predicting PTB. This was to be expected, as infectious diseases are associated with about 25%-30% of the preterm pregnancy cases (Goldenberg, et al., 2008). Our results have a similar conclusion, as every infectious disease other than Hepatitis B became statistically significant with PTB and no other outcomes. It should also be noted that the level of education had a positive effect of lowering the calculated risk, this was evident for all of the three adverse outcomes. The finalized variable sets that were used in the risk prediction phase are highlighted in Table 4 of Publication **II**.

The risk prediction performance of the used models was most promising with early stillbirth; LGBM and SELU network models both achieved 0.75 and 0.76 AUC, which was better when compared to LR and deep ANN, however SELU network had slightly better TPR at 10% FPR. With the external NYC dataset, the performance of the different models was similar. Late stillbirth prediction produced the worst results of the three outcomes. From them, LGBM produced the best result of 0.60 and 0.61 for CDC and NYC data. Lastly, PTB classification results were from in between: LGBM and SELU network both produced 0.64 and 0.67 AUC's. The complete listing of results can be found in Tables 5 and 6 of Publication **II**.

Out of the two ANN models, the SELU network performed better when compared to the deep ANN in each of the experiments. As the purpose of using ensemble learning was to have a diverse set of different algorithms that could potentially complement each other, SELU network was chosen to represent ANN models over the deep ANN. AE achieved similar performance when compared to the best models for PTB and early stillbirth classification, in both datasets. The best AUC for late stillbirth was 0.63, and the 10% TPR was the same as the best individual model. Rest of the related results are listed in Tables 5 and 6 of Publication **II**.

Similarly to AE, WA ensemble reached similar performance when compared to the best performing models, but it achieved it for all outcomes with CDC data. Differing results were achieved in late stillbirth prediction, as the WA ensemble produced a noteworthy TPR increase of 26% at 10% FPR, when the second best TPR for this FPR was 22%, which was created by the LGBM model. For this ensemble, the used weights used were 0.0 for LR, 0.2 for SELU network and 0.8 for LGBM. The results of both ensemble methods would indicate that ensemble learning provided significant increase in predicting late stillbirth, however it was the only outcome that was improved upon. This can also be seen from the weight grid search experiments in Figure 3 of Publication **II**, as it demonstrates to being the only outcome that showed some effect when different weights were iterated, while models for PTB and early stillbirth were mostly unresponsive. Tables 5 and 6 of Publication **II** contain the full listing of results.

The SELU network used did not show any improvements in prediction performance after adding more hidden layers beyond four. This result contradicted the conclusions provided by the authors of the SELU method (Klambauer, et al., 2017). Also, it elaborated on the ANN results of Publication **I** by demonstrating that an ANN's deepness can be increased to a point of saturation. Therefore, one should iteratively design the network architecture by assessing the prediction task at hand.

Our study findings further establish the role of ML models as methods that can achieve improved risk prediction performance over more conventional LR. The main limitations in the study were that the study population was limited to US citizens, and the inability to assess data quality and integrity due to the nature of data released by NYC and CDC. Errors in data entry and other random artefacts can be present in the study data, however the size of the data sets should reduce these types of errors to insignificant levels.

Ensemble of three different ML methods achieved on par results or results with minor improvements. The biggest effect of improving TPR at 10% FPR for late stillbirth with WA ensemble could not be replicated for other outcomes, however the improvement was similar within the CDC and NYC data. This finding suggests that the improvement was task-specific, so future work would include investigating the possibility of using WA ensemble of ML models for other prenatal outcomes to gain additional prediction performance from existing predictor variables.

4.3 Publication III

Six different combinations were used with the evaluation dataset, the prediction results are showcased in Table 3. Comparing the benchmark models side by side, LR was able to achieve better performance over the SELU network. The utilization of SMOTE-NC was reduced to no utilization at all during the optimization of SMOTE-NC and LR, this explains why it produced identical performance to LR. SMOTE-NC combined with SELU network was able to improve performance over the SELU network regarding some of the metrics, while decreasing the others. actGAN combined with LR produced similar conflicting results. This can be seen as a drop in TPR at higher FPRs, shown in Figure 3 of Publication **III**. actGAN combined SELU network was able to improve all of the four metrics when compared to plain SELU network, and it produced the best performance of the whole experiment. Table

2 of the Supplementary material of Publication **III** showcases variable importance results of the classifiers.

Table 3. Model results calculated with the evaluation data set. TPR's at clinically significant FPRs are presented, along with ROC AUC. Models without training data oversampling are presented and models that utilized either SMOTE-NC or actGAN oversampling. From Publication **III**.

Name	AUC (CI 95%)	TPR at 1% FPR	TPR at 3% FPR	TPR at 5% FPR
LR	0.688 (0.620 - 0.756)	9%	16%	20%
SELU Network	0.659 (0.590 - 0.728)	7%	16%	20%
SMOTE-NC & LR	0.688 (0.620 - 0.756)	9%	16%	20%
SMOTE-NC & SELU	0.663 (0.594 - 0.733)	6%	17%	23%
actGAN & LR	0.637 (0.562 - 0.712)	9%	16%	24%
actGAN & SELU	0.704 (0.635 - 0.772)	9%	23%	27%

Applying more intricate modelling-based minority oversampling techniques to vital statistics data or prenatal screening data can be effortful simply because of the restricted amount of affected data available, and mixed-type data to be modelled. This can be demonstrated with SMOTE-NC, which results imply that it could produce improved TPR's at specific FPR's while decreasing it at others. Our proposed actGAN produced more robust results when paired with the SELU network classifier, performance with LR was mixed. actGAN generator network was optimized during the Bayesian hyperparameter optimization to be more complex when paired with SELU network as opposed to LR, this result implies that the generator network architecture needs to reflect the complexity of the used classifier method in order for any beneficial performance to manifest. Compared to SMOTE-NC, the activation-specific output layer of actGAN mostly outperformed the nominal mechanism of SMOTE-NC with both classifier methods.

The limitations of the study are related to the data set used. Similarly as in Publication **II**, due to the release method the quality and integrity of the data cannot be assessed by external entities. Also, the study population was limited to US citizens which were recorded by the NYC health system at the time. This means that the generalizability of actGAN is not exhaustively investigated. This would be the topic for future work.

The applicability of actGAN to any probabilistic modelling task where rare conditions are predicted is the method's biggest upside. By utilizing domain knowledge to improve the modelling of a minority class, any prenatal screening risk assessment model can potentially be improved upon without adding more data, along with clinical models from other domains. This is especially beneficial with rare disorders, as gathering data from rare occurrences is problematic.

4.4 Publication IV

In phase one of the study, the method for detecting distribution shifts was experimented with a range of data block sizes. All the different values represented different durations of time which were estimated from laboratory sample throughput. GpVC was also iterated over a value range, to see the effect of the sensitivity of significance determination to the overall shifts detected. These two parameters were plotted as heatmap coordinates, shown in Figure 10. The plot shows the resulting amounts of distribution shift events that each parameter pair produces.

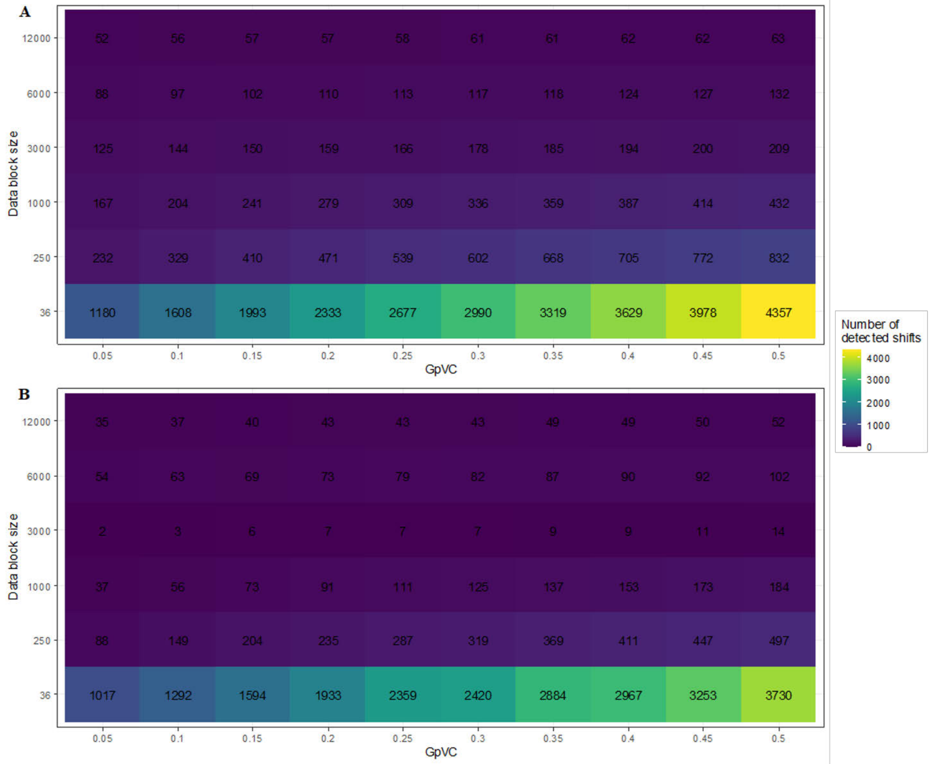

Figure 10. The resulting distribution shifts of the windowed (10A) and cumulative strategies (10B). Number of shift triggers is visualized for data block size and GpVC value ranges. From Publication **IV**.

From Figure 10, it can be seen that with both cumulative and windowed data strategies, the amount of shifts increases when the GpVC or cutoff for significance is increased, this was to be expected as the required level for significance gets more lenient. The results show that from the two parameters, data block size is more important to the number of detected events. This was true with both data strategies. This is also to be expected, as the maximum number of possible events increases. Biggest difference in the two strategies was within the block size of 1000 and 3000 observations, where number of detected events with the 2-block windowed strategy was significantly smaller with any GpVC. The complete listing of the results can be found in the Supplementary material of Publication **IV**.

Results from phase one indicated that the number of detection events from two data processing strategies can differ, given different parameters. What strategy to use and if TL was used or not formulated into four candidate system designs: cumulative strategy with and without TL, and windowed strategy with and without TL. The four systems were all tested to process the whole study data, and all data block size and GpVC value combination were investigated. Complete results of this are listed in the Supplementary material of Publication **IV**. Data block of 1000 and GpVC of 0.05 were investigated in more detail, as it highlighted the differences between the windowed and the cumulative strategies. The four predicate system's AUC performance at each data block were investigated and then evaluated against the predicate methods. Comparison to our published deep ANN model would showcase if TL would be beneficial, and comparison against the laboratory's routine screening method would demonstrate if IL would be stable enough to be used. The latter would also showcase if the automated system could reach similar clinically acceptable performance automatically. These results are plotted in Figure 5 of Publication **IV**. It demonstrates that TL gives the candidate systems the ability to give feasible predictions during the initial phases of operation. Without TL, the results show that the candidate system using cumulative data strategy requires a substantial amount of training data for producing similar performance when compared to the predicates. The plot also shows the volatility of the windowed strategy without TL.

If we calculate ROC curves over the whole study data, we can compare the performance of the predicates and candidates across the whole time span. These results are listed in Table 4.

Table 4. Predicate and candidate system performance metrics calculated over the whole study data. ROC AUC, its bootstrapped 95% confidence interval and TPRs of 5% and 10% FPR are shown. The best performing candidate system is highlighted in bold. From Publication **IV**.

Predicates	**AUC (CI 95%)**	**TPR at 5% FPR**	**TPR at 10% FPR**
Screening lab algorithm	0.98 (0.98 - 0.99)	91%	95%
2018 published model	0.96 (0.95 - 0.97)	79%	87%
Candidates			
Cumulative system	0.92 (0.89 - 0.94)	77%	81%
Cumulative system with TL	0.90 (0.88 - 0.92)	71%	77%
Windowed system	0.89 (0.87 - 0.91)	61%	68%
Windowed system with TL	**0.96 (0.95 - 0.97)**	**82%**	**90%**

Compared to our previously published deep ANN model in Publication **I** that was the TL backbone, the windowed TL system managed to slightly improve the prediction performance with the study data. This would indicate that TL combined with windowed IL is beneficial, while TL with cumulative data strategy was unfavourable. Also when we compare systems without TL, it can be seen that they performed poorly compared to the deep ANN of Publication **I**. Our results also indicate that the screening lab algorithm performs as previously reported (Leung, et al., 2009).

The deployment of an automated adaptive system such as what we experimented with in this study requires thoughtful examination of the used parameters relating to how the distribution shift detection and model updating behave. As the shift detection was parameterized for our application, the sensitivity of the detection needs to be adjusted for the specific task and environment. In addition to this, data block determination should be done according to a screening lab’s throughput, while taking into account the minimum required amount of data for deploying any type of risk modelling to a clinical setting. Also, the performance metric which drives the model updating procedure should be decided accordingly with the prediction task at hand. For the prediction task of T21, the pAUC of 0% to 10% FPR was tested to produce the best overall performance with our study data.

Our empirical testing shows that if we use our proposed method for the detection of distribution shifts where model fitting is based on differences found from data over time, the window strategy provides an alternative to fitting models while cumulating training data which is less computationally expensive. We demonstrated that TL could leverage models fitted with other patient populations as a starting point for adapting to a local population over time. Our proposed ARPS can achieve this adaptation over time in a clinically significant way. Also, it is not limited to T21 prediction, as any clinical probabilistic screening model used in a lab could benefit from it. It is also possible to extend ARPS to include actGAN synthetization for added performance, or the utilization of Shapley additive explanations or SHAP values (Lundberg & Lee, 2017) could be added to improve model transparency, and possibly explain in great detail the differences between models generated within ARPS. Causal representation learning has been recently proposed to address domain adaptation (Schölkopf, et al., 2021), this framework could also be integrated to ARPS in the future.

Generalizability of the proposed method is the main limitation in this study. The data set used in the study is exceptionally large, however it contains a specific patient population. Collecting a data set from another population of equal proportions and evaluating ARPS with this external data set would be the emphasis of future work. Also, implementations to other prenatal outcomes such as stillbirth and preterm birth could be investigated.

The assessment of the study is that IL and TL are at the maturity level where they can be used for in clinical risk assessment in a robust way. ARPS could potentially enable a screening laboratory to start their operation with an existing prediction model that has been fitted to a different population, or to a similar but different clinical risk prediction task, and over time fit or adapt to their local patient population, and improve the risk assessment accuracy of the affected cases. Rare disorders that have not been feasible to build prediction models for in the past could now be within reach, as similar and more common disorder models could be used as a backbone with TL. This was also experimented in our study for Trisomy 13 and 18 (Lakovschek, et al., 2011), where a system using our published T21 model was used as the backbone with TL. The results showed improved prediction performance over the model fitting without TL, and they can be found in the Supplementary material of Publication **IV**. By designing our adaptive system to be built on concepts that are familiar to clinical practitioners, we believe ARPS has the potential to introduce more involved risk assessment to clinical screening situations in general, while utilizing a complex ANN-based system.

5 Conclusions

Risk prediction for the adverse outcomes of pregnancy is a clinical domain where probabilistic modelling is utilized. Different entities relating to risk prediction of adverse outcomes of pregnancy have their own use cases and goals for the modelling methods they utilize. More classical statistical techniques used are selected for their prediction performance and explainability, as they are more easily monitored and regulated. More intricate modelling techniques such as artificial neural networks are commonly not considered, as their explainability is not as good, and the added benefit hasn't been investigated thoroughly in prior literature.

In this thesis, methods that could improve the existing risk prediction of adverse outcomes of pregnancy were investigated. This included the evaluation of the applicability of ML methods to be used by the centralized screening labs and clinicians in a hospital environment, and the proposal of novel methods which address some of the key modelling obstacles during development by researchers and manufacturers. In addition to this, increasing automation related to risk model usage in a screening lab was also investigated. Clinical significance was the driving factor in all of the studies. Pregnancy-related outcomes such as Trisomy 21 or Down syndrome, Trisomy 13 & 18, stillbirth and preterm birth were investigated. The primary goal of this thesis was to consider all entities in the clinical analysis workflow and provide novel methods and applications which could be implemented and used routinely in the real world.

ML has the capability to improve risk prediction modelling in multiple ways; maintaining the clinically sufficient true positive rate while reducing false positive rate in a screening situation, improving the current modelling results by generating synthetic affected observations to learn from, and providing more autonomy to model updating processes.

Based on the original publications **I-IV,** the main conclusions of the thesis are:

I&II: ML methods can improve the existing performance of risk prediction for adverse outcomes of pregnancy. This was demonstrated also with T21 or Down syndrome, where the current standard of prediction performance is considerably high. The amount and complexity of the data gathered which relate to pregnancy will probably increase over time, and in order to match this growth the related analysis methods need to be scalable and efficient. This can also mean utilizing ensemble learning of different modelling techniques.

I&II: In terms of the architecture of a fully-connected feedforward artificial neural network, 2 to 4 hidden layers is appropriate for the domain of prenatal risk prediction. This finding reflects on the complexity of the data commonly used in this context.

III: Screening data is heavily imbalanced in terms of affected and unaffected classes; this is caused by the rare incidence of the affected outcomes. Probabilistic modelling from this type of data can be challenging, as one of the predicted classes is not well represented in the data. To combat this modelling problem, our proposed synthetic minority oversampling method actGAN could be used to mitigate it.

IV: Prenatal screening in a laboratory environment requires extensive knowledge of the local patient population and statistical expertise. This includes improvements and updates of the used risk algorithms. Autonomous adaptation of a risk model which is computationally efficient within a laboratory environment is feasible with our proposed ARPS architecture, which could improve existing screening strategies in general.

The digitalization of healthcare will affect the processes and technologies related to detecting adverse outcomes of pregnancy in the future. More data is collected and processed, which means that more relevant information can be incorporated into the prediction step (Gil, et al., 2015). Methods that scale sufficiently with this phenomenon are needed, and ANN-based methods seem to provide the necessary capabilities for this. They can also enable better utilization of routinely generated data of rare disorders. These benefits can be the necessary advancement needed for achieving clinically significant performance of a prediction model for an outcome, so that it can be implemented for routine clinical use, when no alternative is available.

To conclude, the clinical analysis workflow related to risk prediction of adverse outcomes of pregnancy could potentially benefit from various ML methods presented in original publications **I-IV**.

Acknowledgements

The doctoral research described in this thesis was conducted at the Department of Computing, University of Turku, during the years 2018-2022. The project was done in collaboration with PerkinElmer Wallac Oy. This thesis is the result of the interdisciplinary contribution of prenatal screening experts, health professionals, R&D chemists, statisticians, machine learning experts and one PhD student. I consider myself extremely lucky to have met all the people that were ultimately involved in the publications, as I couldn't have achieved this level of excellence on my own.

Being associated with the Department of Computing has been a privilege. I would like to thank Professor Ville Leppänen for starting my academic career by encouraging me to write my first publication, and for helping me starting my PhD journey. I would also like to thank Lauri Koivunen and Shohreh Hosseinzadeh for the support while writing said publication, as I came to the realization that a PhD was something attainable for me. Lastly, I would like to thank the department for giving me a grant for finalizing my thesis.

I would like to thank both of my supervisors Assistant Professor Antti Airola and Associate Professor Tapio Pahikkala, and my research director Timo Knuutila for their insight. I'm deeply thankful for Tapio and Antti for being a part of my remote research group and for helping me to reach this level of academic professionalism.

I am deeply grateful for my pre-examiners Assistant Professor Pekka Marttinen and Assistant Professor Smisha Agarwal for their hard work. Thank you Pekka for the thorough list of constructive comments and thank you Smisha for the acclaim and for accepting the position in such a short notice.

I would also like to thank all the co-authors which made the publications possible in the first place: Dr. Teemu Korpimäki, Petri Kivelä, Dr. Tapio Pahikkala, Dr. Mikko Sairanen, Dr. Antti Airola, Dr. Wing-cheong Leung, Dr. Tsz-kin Lo and Dr. Daljit Singh Sahota. I would like to thank Dr. Sahota especially for the collaboration regarding the Adaptive Risk Prediction System paper.

The collaboration with PerkinElmer Wallac Oy was crucial for the completion of this thesis, and for that I will always be grateful to the company and the people in it. Thank you Tuukka Pölönen for the input from a statistician's perspective. I would

also like to thank Petri Kivelä and Janne Seppälä for making this possible. Lastly, I would like to thank my New Technologies team co-workers Dr. Tero Lehtonen, Dr. Joona-Pekko Kakko, Dr. Henna Päkkilä, Dr. Ville Veikkolainen, Mikko Aaltoranta, Juuso Huhtala, Dr. Teemu Korpimäki and Dr. Mikko Sairanen for their support, input and comradery. The time spent with you was precious to me. I would also like to thank Mikko Sairanen in particular for helping me from start to finish and mentoring me to overcome any obstacle.

I wish to extend my warmest thanks to my family that supported me during the thesis process. Mom and dad, you always encouraged me to focus on school, as you wanted your children to have a better starting point in life than what you had. For this I am eternally grateful, because I was able to reach achievements beyond my wildest dreams.

Lastly, I would like to thank my dear wife Marika. You are the reason I embarked on this journey in the first place. You and I weathered this storm together, and for that I am forever grateful. Thank you for believing in me in times when I didn't.

09.02.2022, Turku
Aki Koivu

List of References

Abbag, F. I., 2006. Congenital heart diseases and other major anomalies in patients with Down syndrome. *Saudi medical journal,* 27(2), p. 219.

Abdel Razik, M. et al., 2016. Prophylactic treatment for preeclampsia in high-risk teenage primigravidae with nitric oxide donors: a pilot study. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians,* 29(16), pp. 2617-2620.

Abramovici, A., Cantu, J. & Jenkins, S. M., 2012. Tocolytic therapy for acute preterm labor. *Obstetrics and gynecology clinics of North America,* 39(1), p. 77–87.

Agarap, A. F., 2018. Deep learning using rectified linear units (relu). *arXiv preprint,* p. arXiv:1803.08375.

Alexander, G. R. & Kotelchuck, M., 2001. Assessing the role and effectiveness of prenatal care: history, challenges, and directions for future research. *Public health reports,* 116(4), p. 306.

Alfirevic, Z., Navaratnam, K. & Mujezinovic, F., 2017. Amniocentesis and chorionic villus sampling for prenatal diagnosis. *The Cochrane database of systematic reviews,* 9(9).

Alfirevic, Z., Stampalija, T. & Medley, N., 2017. Cervical stitch (cerclage) for preventing preterm birth in singleton pregnancy. *The Cochrane database of systematic reviews,* 6(6).

Allen, M. & Salmon, A., 2020. Synthesising artificial patient-level data for Open Science-an evaluation of five methods. *medRxiv preprint.*

Aminu, M., Unkels, R., Mdegela, M. & Utz, B., 2014. Causes of and factors associated with stillbirth in low- and middle-income countries: a systematic literature review. *BJOG : an International Journal of Obstetrics and Gynaecology,* Issue 121, pp. 141-153.

Arabin, B. & Alfirevic, Z., 2013. Cervical pessaries for prevention of spontaneous preterm birth: past, present and future. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology,* 42(4), p. 390–399.

Arjovsky, M. & Bottou, L., 2017. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv preprint,* p. arXiv:1701.04862.

Arjovsky, M., Chintala, S. & Bottou, L., 2017. Wasserstein GAN. *arXiv preprint,* p. arXiv:1701.07875.

Auer, P., Burgsteiner., H. & Maass, W., 2008. A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks,* 21(5), p. 786–795.

Banach, S., 1922. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae,* Volume 3, p. 133–181.

Berberich, S. L., 2013. *Using Multiples of the Median (MoM) for Normalization of TREC Results Meets the Need for Standardized SCID Reporting,* Atlanta, GA: 2013 Joint Meeting of the Newborn Screening and Genetic Testing Symposium and the International Society for Neonatal Screening.

Bertsekas, D. P., 1996. Incremental least squares methods and the extended Kalman filter. *SIAM Journal on Optimization,* 6(3), pp. 807-822.

Bigirumurame, T. & Kasim, A. S., 2017. Can testing clinical significance reduce false positive rates in randomized controlled trials? A snap review. *BMC research notes,* 10(1), p. 775.

Bishop, J. C. et al., 1993. All MoMs are not equal: some statistical properties associated with reporting results in the form of multiples of the median. *American journal of human genetics,* 52(2), pp. 425-443.

Bonferroni, C. E., 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze,* Volume 8, pp. 3-62.

Bottou, L., 1998. Online Algorithms and Stochastic Approximations. *Online Learning,* Volume 5, pp. 6-41.

Bottou, L. & LeCun, Y., 2004. Large scale online learning. *Advances in neural information processing systems,* Volume 16, pp. 217-224.

Boyd, S. & Vandenberghe, L., 2004. Unconstrained Minimization. *Convex Optimization,* p. 466–474.

Breathnach, F. M. & Malone, F. D., 2007. Screening for aneuploidy in first and second trimesters: is there an optimal paradigm?. *Current opinion in obstetrics & gynecology,* 19(2), pp. 176-182.

Breiman, L., 1996. Bagging predictors. *Machine learning,* 24(2), pp. 123-140.

Breiman, L., 2001. Random forests. *Machine learning,* 45(1), pp. 5-32.

Carpenter, G. A. & Grossberg, S., 1987. ART 2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics,* 26(23), p. 4919–4930.

Centers for Disease Control and Prevention, 2019. *National Vital Statistics System.* [Online] Available at: https://www.cdc.gov/nchs/nvss/ [Accessed 05 04 2019].

Chaiworapongsa, T. et al., 2013. Maternal plasma concentrations of angiogenic/antiangiogenic factors in the third trimester of pregnancy to identify the patient at risk for stillbirth at or near term and severe late preeclampsia. *American journal of obstetrics and gynecology,* 208(4), pp. 1-15.

Chang, H. H. et al., 2013. Preventing preterm births: analysis of trends and potential reductions with interventions in 39 countries with very high human development index. *The Lancet,* 381(9862), p. 223–234.

Chawla, N. V., 2010. Data Mining for Imbalanced Datasets: An Overview. In: O. Maimon & L. Rokach, eds. *Data Mining and Knowledge Discovery Handbook.* New York: Springer, p. 875–886.

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* Volume 16, pp. 321-357.

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

Chitty, L. S. et al., 2016. Uptake, outcomes, and costs of implementing non-invasive prenatal testing for Down's syndrome into NHS maternity care: prospective cohort study in eight diverse maternity units. *BMJ,* Volume 354, p. i3426.

Chiu, C., Li, X. & Mhaskar, H., 1996. Limitations of the approximation cababilities of neural networks with one hidden layer. *Advances in Computational Mathematic,* 5(1), pp. 233-243.

Considine, E. C., Khashan, A. S. & Kenny, L. C., 2019. Screening for Preterm Birth: Potential for a Metabolomics Biomarker Panel. *Metabolites,* 9(5), p. 90.

Coppedè, F., 2016. Risk factors for Down syndrome. *Archives of toxicology,* 90(12), pp. 2917-2929.

Cortes, C. & Vapnik, V. N., 1995. Support-Vector Networks. *Machine Learning,* 20(3), p. 273–297.

Coste, J. & Pouchot, J., 2003. A grey zone for quantitative diagnostic and screening tests. *International journal of epidemiology,* 32(2), p. 304–313.

Cuckle, H. & Benn, P., 2010. Multianalyte Maternal Serum Screening for Chromosomal Defects. In: *Genetic Disorders and the Fetus: Diagnosis, Prevention and Treatment..* Baltimore, MD: Johns Hopkins University Press, pp. 771-818.

Cuckle, H. & Maymon, R., 2016. Development of prenatal screening—A historical overview. *Seminars in perinatology,* 40(1), pp. 12-22.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems,* 2(4), p. 303–314.

Davis, C., Cuckle, H. & Yaron, Y., 2014. Screening for Down syndrome--incidental diagnosis of other aneuploidies. *Prenatal diagnosis,* 34(11), pp. 1044-1048.

De Jesús, V. R., Mei, J. V., Bell, C. J. & Hannon, W. H., 2010. Improving and assuring newborn screening laboratory quality worldwide: 30-year experience at the Centers for Disease Control and Prevention. *Seminars in perinatology,* 34(2), p. 125–133.

Dodd, J. M., McLeod, A., Windrim, R. C. & Kingdom, J., 2013. Antithrombotic therapy for improving maternal or infant health outcomes in women considered at risk of placental dysfunction. *The Cochrane database of systematic reviews,* Volume 7.

Dodd, L. E. & Pepe, M. S., 2003. Partial AUC estimation and regression. *Biometrics,* 59(3), pp. 614-623.

Dong, S. et al., 2020. Using Undersampling with Ensemble Learning to Identify Factors Contributing to Preterm Birth. *arXiv preprint,* p. arXiv:2009.11242.

Duley, L., Henderson-Smart, D. J., Meher, S. & King, J. F., 2007. Antiplatelet agents for preventing pre-eclampsia and its complications. *The Cochrane database of systematic reviews,* Volume 2.

Fergus, P., Idowu, I., Hussain, A. & Dobbins, C., 2016. Advanced artificial neural network classification for detecting preterm births using EHG records. *Neurocomputing,* Volume 188, pp. 42-49.

Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research,* Volume 61, pp. 863-905.

Fiscella, K., 1995. Does prenatal care improve birth outcomes? A critical review. *Obstetrics & Gynecology,* 85(3), p. 468–479.

Food and Drug Administration, 2011. 21 CFR 809 - IN VITRO DIAGNOSTIC PRODUCTS FOR HUMAN USE. *Code of Federal Regulations.*

Forgy, E. W., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics,* Volume 21, pp. 768-769.

Foster, M. & Sherrington, C., 1897. *Textbook of Physiology, volume 3.* 7th ed. London: Macmillan.

Fournier, N. & Guillin, A., 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields,* 162(3), pp. 707-738.

Gebremariam, M. K. et al., 2018. Change in BMI Distribution over a 24-Year Period and Associated Socioeconomic Gradients: A Quantile Regression Analysis. *Obesity,* 26(4), pp. 769-775.

Ghezzi, F. et al., 2002. Elevated amniotic fluid C-reactive protein at the time of genetic amniocentesis is a marker for preterm delivery. *American journal of obstetrics and gynecology,* 186(2), p. 268–273.

Gil, M. M. et al., 2015. Analysis of cell-free DNA in maternal blood in screening for fetal aneuploidies: updated meta-analysis. *Ultrasound in obstetrics & gynecology,* 45(3), pp. 249-266.

Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R., 2008. Epidemiology and causes of preterm birth. *Lancet,* 371(9606), pp. 75-84.

Goodfellow, I. et al., 2014. Generative Adversarial Networks. *Proceedings of the International Conference on Neural Information Processing,* pp. 2672-2680.

Greenhalgh, T., 1997. How to read a paper. Papers that report diagnostic or screening tests. *Bmj,* 315(7107), p. 540–543.

Gulrajani, I. et al., 2017. Improved Training of Wasserstein GANs. *arXiv preprint,* p. arXiv:1704.00028.

Hager, W. W., 1979. Lipschitz continuity for constrained processes. *SIAM Journal on Control and Optimization,* 17(3), pp. 321-338.

Hall, M. A., 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. *Proceedings of the Seventeenth International Conference on Machine Learning,* pp. 359-366.

Han, H., Wang, W. Y. & Mao, B. H., 2005. *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning.* Berlin, Springer.

Harper, M. et al., 2010. Omega-3 fatty acid supplementation to prevent recurrent preterm birth: a randomized controlled trial. *Obstetrics and gynecology,* 115(2), p. 234–242.

Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hebb, D. O., 2005. *The organization of behavior: A neuropsychological theory.* Psychology Press.

He, H., Bai, Y., Garcia, E. A. & Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE international joint conference on neural networks,* pp. 1322-1328.

He, H. & Ma, Y., 2013. *Imbalanced learning: foundations, algorithms, and applications.* 1st ed. New Jersey: Wiley-IEEE Press.

Heng, Y. J. et al., 2014. Whole blood gene expression profile associated with spontaneous preterm birth in women with threatened preterm labor. *PLoS One,* 9(5).

Herrup, K. & Yang, Y., 2007. Cell cycle regulation in the postmitotic neuron: oxymoron or new biology. *Nature Reviews, Neuroscience,* 8(5), p. 368–78.

Hill, J. L. et al., 2008. Prediction of preterm birth in symptomatic women using decision tree modeling for biomarkers. *American journal of obstetrics and gynecology,* 198(4).

Hinton, G. E., Sejnowski, T. J. & (Eds.), 1999. *Unsupervised learning: foundations of neural computation.* MIT press.

Hodgkin, A. L. & Huxley, A. F., 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology,* 117(4), p. 500–544.

Hoffman, M. K., Ma, N. & Roberts, A., 2021. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. *American Journal of Obstetrics & Gynecology MFM,* 3(1), p. 100250.

Honest, H. et al., 2009. Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health technology assessment,* 13(43), p. 1–627.

Ho, T. K., 1995. Random decision forests. *IEEE Proceedings of 3rd international conference on document analysis and recognition,* Volume 1, pp. 278-282.

Ho, T. K., 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 20(8), p. 832–844.

Howard, A., 1987. *Elementary Linear Algebra.* 5th ed. New York: Wiley.

Howell, D., 2002. *Statistical Methods for Psychology.* Pacific Grove: Duxbury.

Huang, J. & Ling, C. X., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering,* 17(3), pp. 299-310.

Iams, J. D. et al., 2001. The preterm prediction study: can low-risk women destined for spontaneous preterm birth be identified?. *American journal of obstetrics and gynecology,* 184(4), p. 652–655.

Ichihara, K. et al., 2008. Sources of variation of commonly measured serum analytes in 6 Asian cities and consideration of common reference intervals. *Clinical chemistry,* 54(2), p. 356–365.

Ijaz, M. F., Alfian, G., Syafrudin, M. & Rhee, J., 2018. Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest. *Applied Sciences,* 8(8), p. 1325.

Inyang, U. G. et al., 2020. Comparative Analytics of Classifiers on Resampled Datasets for Pregnancy Outcome Prediction. *International Journal of Advanced Computer Science and Applications,* 11(6).

Jabbar, A., Li, X. & Omar, B., 2020. A Survey on Generative Adversarial Networks: Variants, Applications, and Training. *arXiv preprint,* p. arXiv:2006.05132.

James, D. K., Steer, P. J., Weiner, C. P. & Gonik, B., 2010. *High risk pregnancy e-book: Management options-expert consult.* 1st ed. Elsevier Health Sciences.

Jhee, J. H. et al., 2019. Prediction model development of late-onset preeclampsia using machine learning-based methods. *PloS one,* 14(8), p. e0221202.

Kayode, G. et al., 2016. Predicting stillbirth in a low resource setting. *BMC Pregnancy and Childbirth,* Volume 16, pp. 274-284.

Ke, G. et al., 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30,* p. 3149–3157.

Klambauer, G., Unterthiner, T., Mayr, A. & Hochreiter, S., 2017. Self-normalizing neural networks. *arXiv preprint,* p. arXiv:1706.02515.

Kogan, M. D., Alexander, G. R., Kotelchuck, M. & Nagey, D. A., 1994. Relation of the content of prenatal care to the risk of low birth weight: maternal reports of health behavior advice and initial prenatal care procedures. *Jama,* 271(17), pp. 1340-1345.

Kullback, S. & Leibler, R. A., 1951. On information and sufficiency. *The annals of mathematical statistics,* 22(1), pp. 79-86.

Lakovschek, I. C., Streubel, B. & Ulm, B., 2011. Natural outcome of trisomy 13, trisomy 18, and triploidy after prenatal diagnosis. *American Journal of Medical Genetics Part A,* 155(11), pp. 2626-2633.

Lakshmi, B. N., Indumathi, T. S. & Ravi, N., 2016. A Study on C. 5 decision tree classification algorithm for risk predictions during pregnancy. *Procedia Technology,* Volume 24, pp. 1542-1549.

Lazaridis, G., Lorenzi, M., Ourselin, S. & Garway-Heath, D., 2021. Improving statistical power of glaucoma clinical trials using an ensemble of cyclical generative adversarial networks. *Medical Image Analysis,* Volume 68, p. 101906.

Lee, J. et al., 2020. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment,* Volume 12, p. 124002.

Lee, K. S. & Ahn, K. H., 2019. Artificial neural network analysis of spontaneous preterm labor and birth and its major determinants. *Journal of Korean medical science,* 34(16).

Leung, T. et al., 2009. First trimester combined screening for trisomy 21 in Hong Kong: outcome of the first 10,000 cases. *The journal of maternal-fetal & neonatal medicine,* 22(4), pp. 300-304.

Li, H. et al., 2017. An artificial neural network prediction model of congenital heart disease based on risk factors: a hospital-based case-control study. *Medicine,* 96(9).

Ling, C. X. & Chenghui, L., 1998. Data mining for direct marketing: Problems and solutions. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining,* Volume 98, pp. 73-79.

Linnainmaa, S., 1970. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Department of Computer Science, University of Helsinki.*

Lin, T. Y. et al., 2017. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision,* pp. 2980-2988.

Lundberg, S. & Lee, S. I., 2017. A unified approach to interpreting model predictions. *arXiv preprint,* p. arXiv:1705.07874.

Luo, Y. et al., 2017. Predicting congenital heart defects: A comparison of three data mining methods. *PloS one,* 12(5), p. e0177811.

Maas, A. L., Hannun, A. Y. & Ng, A. Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Model. *ICML Workshop on Deep Learning for Audio, Speech and Language Processing,* 30(1), p. 3.

Maciejewski, T. & Stefanowski, J., 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. *2011 IEEE symposium on computational intelligence and data mining (CIDM),* pp. 104-111.

Mai, C. T. et al., 2019. National population-based estimates for major birth defects, 2010–2014. *Birth Defects Research,* Volume 111, p. 1420–1435.

Malacova, E. et al., 2020. Stillbirth risk prediction using machine learning for a large cohort of births from Western Australia. *Scientific reports,* 10(1), p. 1980–2015.

Malacova, E. et al., 2020. Stillbirth risk prediction using machine learning for a large cohort of births from Western Australia, 1980-2015. *Scientific reports,* 10(1), pp. 1-8.

Martin, J. A. et al., 2009. Births: Final Data for 2006. *National Vital Statistics Reports,* Volume 57, pp. 1-25.

McClure, E. M. et al., 2011. Epidemiology of stillbirth in low-middle income countries: a Global Network Study. *Acta obstetricia et gynecologica Scandinavica,* 90(12), p. 1379–1385.

McClure, E. M., Saleem, S., Pasha, O. & Goldenberg, R. L., 2009. Stillbirth in developing countries: a review of causes, risk factors and prevention strategies. *The journal of maternal-fetal & neonatal medicine : the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians,* 22(3), p. 183–190.

McCullagh, P. & Nelder, J. A., 1987. Generalized linear models. *Biometrical Journal,* 29(2), pp. 323-3847.

McCulloch, W. & Pitts, W., 1943. A Logical Calculus of Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics,* 5(4), p. 115–133.

Medley, N., Vogel, J. P., Care, A. & Alfirevic, Z., 2018. Interventions during pregnancy to prevent preterm birth: an overview of Cochrane systematic reviews. *The Cochrane database of systematic reviews,* Volume 11.

Meertens, L. et al., 2018. Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation. *Acta obstetricia et gynecologica Scandinavica,* 97(8), p. 907–920.

Menéndez, M., 2005. Down syndrome, Alzheimer's disease and seizures. *Brain and Development,* 27(4), pp. 246-252.

Minka, T. P., 2003. *A comparison of numerical optimizers for logistic regression.* [Online] Available at: https://tminka.github.io/papers/logreg/ [Accessed 16 12 2020].

Mirjalili, S., 2019. Genetic algorithm. In: *Evolutionary algorithms and neural networks.* Cham: Springer, pp. 43-55.

Moreira, M. W. et al., 2017. Predicting hypertensive disorders in high-risk pregnancy using the random forest approach. *IEEE International Conference on Communications,* pp. 1-5.

Munson, P. J. & Rodbard, D., 1978. An elementary components of variance analysis for multi-centre quality control. *Radioimmunoassay and related procedures in medicine,* 10(22).

Murphy, K. P., 2012. *Machine learning: a probabilistic perspective.* Cambridge: MIT press.

Naftaly, U., Intrator, N. & Horn, D., 1997. Optimal ensemble averaging of neural networks. *Network: Computation in Neural Systems,* 8(3), p. 283–296.

Neocleous, A. C., Nicolaides, K. H. & Schizas, C. N., 2016. First trimester Noninvasive Prenatal Diagnosis: A Computational Intelligence Approach. *IEEE Journal of Biomedical and Health Informatics,* 20(5), pp. 1427-38.

Nicholas, S. S. et al., 2009. Predicting adverse neonatal outcomes in fetuses with abdominal wall defects using prenatal risk factors. *American journal of obstetrics and gynecology,* 201(4), pp. 383-e1.

Norwitz, E. R. & Caughey, A. B., 2011. Progesterone supplementation and the prevention of preterm birth. *Reviews in obstetrics & gynecology,* 4(2), p. 60–72.

Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S., 2018. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint,* p. arXiv:1811.03378.

Obuchowski, N. A., Lieber, M. L. & Wians Jr, F. H., 2004. ROC curves in clinical chemistry: uses, misuses, and possible solutions. *Clinical chemistry,* 50(7), pp. 1118-1125.

Olivia Kim, U. et al., 2019. Smartphone-based prenatal education for parents with preterm birth risk factors. *Patient education and counseling,* 102(4), pp. 701-708.

Ong, C. Y. et al., 2000. First trimester maternal serum free beta human chorionic gonadotrophin and pregnancy associated plasma protein A as predictors of pregnancy complications.. *BJOG : an international journal of obstetrics and gynaecology,* 107(10), p. 1265–1270.

Opitz, D. & Maclin, R., 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research,* Volume 11, p. 169–198.

Orthmann Bless, D. & Hofmann, V., 2020. Abortion in women with Down syndrome. *Journal of Intellectual Disability Research,* 64(9), pp. 690-699.

Padula, F. et al., 2014. Retrospective study evaluating the performance of a first-trimester combined screening for trisomy 21 in an Italian unselected population. *Journal of prenatal medicine,* 8(3-4), pp. 50-56.

Parritz, R. H. & Troy, M. F., 2018. *Disorders of childhood: Development and psychopathology.* 3rd ed. Boston, MA: Cengage Learning.

Patterson, D., 2009. Molecular genetic analysis of Down syndrome. *Human Genetics,* 126(1), pp. 195-214.

Piryonesi, S. M. & El-Diraby, T. E., 2020. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements,* 146(2), p. 04020022.

Podgorelec, V., Kokol, P., Stiglic, B. & Rozman, I., 2002. Decision trees: an overview and their use in medicine. *Journal of medical systems,* 26(5), p. 445–463.

Powers, D. M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint,* p. arXiv:2010.16061.

Purisch, S. E. & Gyamfi-Bannerman, C., 2017. Epidemiology of preterm birth. *Seminars in Perinatology,* 41(7), pp. 387-391.

Qiu, H. et al., 2017. Electronic health record driven prediction for gestational diabetes mellitus in early pregnancy. *Scientific reports,* 7(1), pp. 1-13.

Qiu, J. et al., 2019. Personalized prediction of live birth prior to the first in vitro fertilization treatment: a machine learning method. *Journal of translational medicine,* 17(1), pp. 1-8.

Quinlan, J. R., 1986. Induction of decision trees. *Machine learning,* 1(1), pp. 81-106.

Radnai, M. et al., 2009. Benefits of periodontal therapy when preterm birth threatens. *Journal of dental research,* 88(3), p. 280–284.

Ramanathan, S., Sangeetha, M., Talwai, S. & Natarajan, S., 2018. Probabilistic Determination Of Down's Syndrome Using Machine Learning Technique. *IEEE International Conference on Advances in Computing, Communications and Informatics,* pp. 126-132.

Rashidian, S. et al., 2020. SMOOTH-GAN: Towards Sharp and Smooth Synthetic EHR Data Generation. *International Conference on Artificial Intelligence in Medicine,* pp. 37-48.

Ratcliff, R., 1990. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review.,* 97(2), p. 285–308.

Redko, I. et al., 2019. *Advances in Domain Adaptation Theory.* ISTE Press - Elsevier.

Reisner, H., 2013. Developmental and genetic diseases. In: *Essentials of Rubin's Pathology.* Lippincott Williams & Wilkins, pp. 129-131.

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review,* 65(6), p. 386.

Rose, N. C., Kaimal, A. J., Dugoff, L. & Norton, M. E., 2020. Screening for Fetal Chromosomal Abnormalities. *ACOG Practice Bulletin,* 136(4), pp. 48-69.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J., 1986. Learning representations by back-propagating errors. *Nature,* 323(6088), pp. 533-536.

Sahota, D. S. et al., 2009. Fetal crown-rump length and estimation of gestational age in an ethnic Chinese population. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultras,* 33(2), pp. 157-160.

Salehi, F., Abbasi, E. & Hassibi, B., 2019. The impact of regularization on high-dimensional logistic regression. *arXiv preprint,* p. arXiv:1906.03761.

Say, L., Gülmezoglu, A. M. & Hofmeyr, G. J., 2003. Maternal oxygen administration for suspected impaired fetal growth. *The Cochrane database of systematic reviews,* Volume 1.

Schapire, R. E., 2013. Explaining adaboost. *Empirical inference,* pp. 37-52.

Schmidt-Wilcke, T. et al., 2010. Distinct patterns of functional and structural neuroplasticity associated with learning Morse code. *Neuroimage,* 51(3), pp. 1234-1241.

Schölkopf, B. et al., 2021. Toward causal representation learning. *Proceedings of the IEEE,* 109(5), pp. 612-634.

Shannon, C. E., 2001. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review,* 5(1), pp. 3-55.

Shen, J. et al., 2020. An Innovative Artificial Intelligence–Based App for the Diagnosis of Gestational Diabetes Mellitus (GDM-AI): Development Study. *Journal of Medical Internet Researc,* 22(9), p. e21573.

Shen, J. et al., 2020. An Innovative Artificial Intelligence–Based App for the Diagnosis of Gestational Diabetes Mellitus (GDM-AI): Development Study. *Journal of Medical Internet Research,* 22(9).

Sherman, S., Allen, E., Bean, L. & Freeman, S., 2007. Epidemiology of Down syndrome. *Developmental Disabilities Research Reviews,* 13(3), pp. 221-227.

Shibuya, M., 2008. Vascular endothelial growth factor-dependent and -independent regulation of angiogenesis. *BMB Reports,* 41(4), p. 278–86.

Siegel, S. & Castellan, N. J. J., 1988. *Nonparametric statistics for the behavioral sciences.* New York: McGraw–Hill.

Smith, G., 2017. Screening and prevention of stillbirth. *Best practice & research. Clinical obstetrics & gynaecology,* Volume 38, pp. 71-82.

Smith, G. C., 2001. Life-table analysis of the risk of perinatal death at term and post term in singleton pregnancies. *American journal of obstetrics and gynecology,* 184(3), p. 489–496.

Smith, G. C. et al., 2004. First-trimester placentation and the risk of antepartum stillbirth. *JAMA,* 292(18), p. 2249–2254.

Smith, G. C. et al., 2007. Maternal and biochemical predictors of antepartum stillbirth among nulliparous women in relation to gestational age of fetal death. *BJOG : an international journal of obstetrics and gynaecology,* 114(6), p. 705–714.

Smith, G. et al., 2002. Early Pregnancy Levels of Pregnancy-Associated Plasma Protein A and the Risk of Intrauterine Growth Restriction, Premature Birth, Preeclampsia, and Stillbirth. *The Journal of Clinical Endocrinology & Metabolism,* 4(1), p. 1762–1767.

Soneji, S. & Beltrán-Sánchez, H., 2019. Association of Maternal Cigarette Smoking and Smoking Cessation With Preterm Birth. *JAMA network open,* 2(4).

Son, M. & Miller, E. S., 2017. Predicting preterm birth: Cervical length and fetal fibronectin. *Seminars in perinatology,* 41(8), p. 445–451.

Sorkine, O. et al., 2004. Laplacian surface editing. *In Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing,* pp. 175-184.

Sosa, C., Althabe, F., Belizán, J. & Bergel, E., 2004. Bed rest in singleton pregnancies for preventing preterm birth. *The Cochrane database of systematic reviews,* Volume 1, p. CD003581.

Souka, A. P. et al., 2005. Increased nuchal translucency with normal karyotype. *American journal of obstetrics and gynecology,* 192(4), p. 1005–1021.

Souza, R. T. et al., 2019. Trace biomarkers associated with spontaneous preterm birth from the maternal serum metabolome of asymptomatic nulliparous women - parallel case-control studies from the SCOPE cohort. *Scientific reports,* 9(1), p. 13701.

Spencer, R. N., Carr, D. J. & David, A. L., 2014. Treatment of poor placentation and the prevention of associated adverse outcomes--what does the future hold?. *Prenatal diagnosis,* 34(7), p. 677–684.

Spilka, J. et al., 2014. Discriminating normal from "abnormal" pregnancy cases using an automated fhr evaluation method. *Hellenic Conference on Artificial Intelligence,* pp. 521-531.

Sprawka, N. et al., 2011. Adjustment of maternal serum alpha-fetoprotein levels in women with pregestational diabetes. *Prenatal diagnosis,* 31(3), pp. 282-285.

Stanton, C. et al., 2006. Stillbirth rates: delivering estimates in 190 countries. *The Lancet,* 367(9521), pp. 1487-1494.

Steegers, E. A., Von Dadelszen, P., Duvekot, J. J. & Pijnenborg, R., 2010. Pre-eclampsia. *The Lancet,* 376(9741), pp. 631-644.

Stephens, M. A., 1974. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association,* 69(347), pp. 730-737.

Stitson, M. O. et al., 1997. *Support Vector Regression with ANOVA Decomposition Kernels,* Surrey: MIT Press.

Sugiyama, M., Krauledat, M. & Müller, K. R., 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research,* 8(5).

Szumilas, M., 2010. Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent,* 19(3), p. 227–229.

The American College of Obstetricians and Gynecologists, 2009. Acog practice bulletin no. 102: Management of stillbirth. *Obstetrics and gynecology,* 113(3), pp. 748-761.

The American College of Obstetricians and Gynecologists, 2020. *Patient education: How your fetus grows during pregnancy.* [Online] Available at: https://www.acog.org/womens-health/faqs/how-your-fetus-grows-during-pregnancy [Accessed 23 May 2021].

Therrien, R. & Doyle, S., 2018. Role of training data variability on classifier performance and generalizability. *In Medical Imaging 2018: Digital Pathology,* Volume 10581, p. 1058109.

Tomasi, T. B., 1977. tructure and function of alpha-fetoprotein. *Annual Review of Medicine,* Volume 28, p. 453–65.

Tomek, I., 1976. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics,* Volume 6, p. 769–772.

Torrey, L. & Shavlik, J., 2010. Transfer learning. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques,* pp. 242-264.

Trudell, A. S. et al., 2017. A stillbirth calculator: Development and internal validation of a clinical prediction model to quantify stillbirth risk. *PloS one,* 12(3).

Turnbull, E. et al., 2011. Causes of stillbirth, neonatal death and early childhood death in rural Zambia by verbal autopsy assessments. *Tropical medicine & international health,* 16(7), p. 894–901.

Uzun, Ö., Kaya, H., Gürgen, F. & Varol, F., 2013. Prenatal Risk Assessment of Trisomy 21 by Probabilistic Classifiers. *Signal Processing and Communications Applications Conference,* pp. 1-4.

Versi, E., 1992. "Gold standard" is an appropriate term. *BMJ,* 305(6846), p. 187.

Verweij, E. J. et al., 2013. European Non-Invasive Trisomy Evaluation (EU-NITE) study: a multicenter prospective cohort study for non-invasive fetal trisomy 21 testing. *Prenatal Diagnosis,* 33(10), pp. 996-1001.

Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience.*

Wahba, G., 1990. Spline Models for Observational Data. *CBMS-NSF Regional Conference Series in Applied Mathematics,* p. 177.

Wald, N. J. et al., 1977. Maternal serum-alpha-fetoprotein measurement in antenatal screening for anencephaly and spina bifida in early pregnancy. Report of U.K. collaborative study on alpha-fetoprotein in relation to neural-tube defects. *The Lancet,* 1(8026), pp. 1323-1332.

Wald, N. J. et al., 2003. First and second trimester antenatal screening for Down's syndrome: the results of the Serum, Urine and Ultrasound Screening Study (SURUSS). *Health Technology Assessment,* 7(11).

Wald, N. & Nicolaides, K. H., 1976. The detection of neural tube defects by screening maternal blood. *Prenatal Diagnosis,* pp. 227-238.

Wallenstein, M. B. et al., 2016. Inflammatory biomarkers and spontaneous preterm birth among obese women. *The Journal of Maternal-Fetal & Neonatal Medicine,* 29(20), pp. 3317-3322.

Waller, D. K. et al., 1996. The association between maternal serum alpha-fetoprotein and preterm birth, small for gestational age infants, preeclampsia, and placental complications. *Obstetrics and gynecology,* 88(5), p. 816–822.

Weijerman, M. E. & de Winter, J. P., 2010. Clinical practice. The care of children with Down syndrome. *European journal of pediatrics,* 169(12), p. 1445–1452.

Werner, E. F. et al., 2011. Universal cervical-length screening to prevent preterm birth: a cost-effectiveness analysis. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology,* 38(1), p. 32–37.

Westreich, D., Lessier, J. & Funk, M. J., 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology,* 63(8), pp. 826-833.

Whiting, P. et al., 2004. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. *Annals of internal medicine,* 140(3), p. 189–202.

Williams, C. J., Lee, S. S., Fisher, R. A. & Dickerman, L. H., 1999. A comparison of statistical methods for prenatal screening for Down syndrome. *Applied Stochastic Models in Business and Industry,* 15(2), pp. 89-101.

Wilson, J. a. L. K., 2015. Short History of the Logistic Regression Model. In: *Modeling Binary Correlated Responses using SAS, SPSS and R.* Cham: Springer, pp. 17-23.

World Health Organization, 2004. Appropriate body-mass index for Asian populations and its implications for policy and intervention strategies.. *The Lancet,* 363(9403), p. 157–163.

World Health Organization, 2014. *10th revision of International Statistical Classification of Diseases and Related Health Problems,* Geneva: WHO.

Xu, L., Skoularidou, M., Cuesta-Infante, A. & Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems,* pp. 7335-7345.

Xu, L. & Veeramachaneni, K., 2018. Synthesizing tabular data using generative adversarial networks. *arXiv preprint,* p. arXiv:1811.11264.

Yaron, Y. et al., 1999. Second-trimester maternal serum marker screening: maternal serum α-fetoprotein, β-human chorionic gonadotropin, estriol, and their various combinations as predictors of pregnancy outcome. *American journal of obstetrics and gynecology,* 181(4), pp. 968-974.

Ye, J., Chow, J. H., Chen, J. & Zheng, Z., 2009. Stochastic gradient boosted distributed decision trees. *Proceedings of the 18th ACM conference on Information and knowledge management,* pp. 2061-2064.

Yerlikaya, G. et al., 2016. Prediction of stillbirth from maternal demographic and pregnancy characteristics. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology,* 48(5), p. 607–612.

Ylijoki, M. K., Ekholm, E., Ekblad, M. & Lehtonen, L., 2019. Prenatal Risk Factors for Adverse Developmental Outcome in Preterm Infants-Systematic Review. *Frontiers in psychology,* Volume 10, p. 595.

Yoon, P. W. et al., 2001. The National Birth Defects Prevention Study. *Public health reports,* 116(Suppl 1), pp. 32-40.

Zernikow, B. et al., 1998. Artificial neural network for risk assessment in preterm neonates. *Archives of Disease in Childhood-Fetal and Neonatal Edition,* 79(2), pp. F129-F134.

Zhang, E. & Zhang, Y., 2009. Average Precision. In: *Encyclopedia of Database Systems.* Boston: Springer US, pp. 192-193.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2022

ISBN 978-951-29-8771-9 (PRINT)
ISBN 978-951-29-8772-6 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/Online)