Computers in Biology and Medicine 138 (2021) 104886

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Adaptive risk prediction system with incremental and transfer learning

Aki Koivu [a,*], Mikko Sairanen [b], Antti Airola [a], Tapio Pahikkala [a], Wing-cheong Leung [c], Tsz-kin Lo [d], Daljit Singh Sahota [e]

[a] *University of Turku, Department of Computing, Turun Yliopisto, 20500, Turku, Finland*
[b] *PerkinElmer, Mustionkatu 6, Turku, Finland*
[c] *Department of Obstetrics and Gynaecology, Kwong Wah Hospital, Hong Kong, China*
[d] *Department of Obstetrics and Gynaecology, Princess Margaret Hospital, Hong Kong, China*
[e] *The Chinese University of Hong Kong, Department of Obstetrics and Gynaecology, Hong Kong, China*

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Incremental learning
Transfer learning
Prenatal screening

ABSTRACT

Currently, popular methods for prenatal risk assessment of fetal aneuploidies are based on multivariate probabilistic modelling, that are built on decades of scientific research and large-scale multi-center clinical studies. These static models that are deployed to screening labs are rarely updated or adapted to local population characteristics. In this article, we propose an adaptive risk prediction system or ARPS, which considers these changing characteristics and automatically deploys updated risk models.

8 years of real-life Down syndrome screening data was used to firstly develop a distribution shift detection method that captures significant changes in the patient population and secondly a probabilistic risk modelling system that adapts to new data when these changes are detected. Various candidate systems that utilize transfer -and incremental learning that implement different levels of plasticity were tested.

Distribution shift detection using a windowed approach provides a computationally less expensive alternative to fitting models at every data block step while not sacrificing performance. This was possible when utilizing transfer learning. Deploying an ARPS to a lab requires careful consideration of the parameters regarding the distribution shift detection and model updating, as they are affected by lab throughput and the incidence of the screened rare disorder. When this is done, ARPS could be also utilized for other population screening problems.

We demonstrate with a large real-life dataset that our best performing novel Incremental-Learning-Population-to-Population-Transfer-Learning design can achieve on par prediction performance without human intervention, when compared to a deployed risk screening algorithm that has been manually updated over several years.

1. Background and significance

Population-based prenatal screening provides the means for identification of various adverse pregnancy-related outcomes early in pregnancy in order to provide prophylactic treatment, greater clinical monitoring to monitor disease progression and optimise timing of delivery. Currently, the most common prenatal screening tests are for fetal aneuploidies, specifically Down syndrome or Trisomy 21 (T21) in the 1st trimester using the combined test [1]. This test combines the results from maternal blood analysis, fetal ultrasound examination with relevant maternal obstetric history and socio-demographic information to produce a personalised estimated risk [2]. Software used to estimate this risk use published algorithms employing multivariate probabilistic models whose probability distributions are the result of decades of scientific research and large-scale multi-center clinical studies [2–4]. These probabilistic models in addition require all continuous measurements used to be standardised prior their use by expressing all measurements as a multiple of the expected median (MoM) [5] level in pregnancy, maternal physical size, current and past pregnancy history and socio-demographic factors.

In practise, when risk software are deployed to the field, the models are rarely updated [6], frequently not adapted to local population characteristics [5] and only revised if a laboratory screening performance significantly deviates from expected levels. In addition, the generalisability and therefore usefulness of probabilistic models in an unseen patient population can be limited, as they rely on the assumption

[*] Corresponding author.
E-mail addresses: aki.i.koivu@utu.fi (A. Koivu), mikko.sairanen@perkinelmer.com (M. Sairanen), ajairo@utu.fi (A. Airola), aatapa@utu.fi (T. Pahikkala), leungwc@ha.org.hk (W.-c. Leung), lotk@ha.org.hk (T.-k. Lo), daljit@cuhk.edu.hk (D.S. Sahota).

https://doi.org/10.1016/j.compbiomed.2021.104886
Received 30 June 2021; Received in revised form 3 September 2021; Accepted 20 September 2021
Available online 23 September 2021

that any unseen population will have similar characteristic to that of the original population used to create the initial model [7]. Risk models developed in one part of the world and then utilized in another can result in reduced prediction performance of the model simply due to the population differences between the two places [8]. In addition, population characteristics in a region of a screening laboratory can also change over time [9,10]. Furthermore, changing the models requires individual laboratories to have technicians who possess the requisite analytical skills to revise models and assess the impact of any on changes on screening performance. Few laboratories however possess such individuals.

One alternative would be to use an automated adaptive risk prediction system (ARPS) in parallel with the risk software to monitor, revise and assess models in the risk software. Automation would potentially allow the risk estimation software to incrementally learn (IL) and adapt models to local variance and populations over time through continuous feedback [11]. ARPS would also monitor distribution changes of the data and activate the learning process or training when it detects enough deviation, thus eliminating the need to unnecessary re-train the used model at every point in time when new data arrives, which is demanding in terms of computing resources. Systems for adapting from one domain to the next have been proposed in the past [12–15], and more than often they are developed for a specific task in mind. Moving away from the mindset of static risk prediction models and towards locally adapting systems with a quality control based on clinical significance would be the next steps for risk prediction. To our knowledge, this has not been addressed in scientific literature in terms of prenatal screening risk algorithms.

2. Objective

The objectives of this study are firstly to design a method for detecting distribution shifts in terms of feature variables used in T21 risk prediction, secondly develop a probabilistic risk modelling system that adapts to new data when these shifts are detected. To this end, statistical testing methods and incremental learning and population-to-population transfer learning with neural networks are investigated. The resulting system should achieve better automation for model updating while maintaining the clinically feasible performance. Also, it should be parameterizable for determining the sensitivity of the shift detection, and model updating should be based on clinically significant performance. The prediction performance of our proposed system is also simulated with a real-life dataset that contains data collected in over 8 years. The historical risk prediction results produced by the currently deployed lab's model [16] and our previously published static model [17] are used as a benchmark for screening performance.

3. Materials and methods

3.1. Study data

Anonymised patient screening and pregnancy outcome data of women having a singleton pregnancy were extracted from the screening database of the Obstetrics Screening Laboratory of The Department of Obstetrics and Gynaecology of The Chinese University of Hong Kong for the period between July 2011 and June 2019 inclusive. All women attended the Hong Kong Hospital Authority Universal Down Syndrome screening test at 11–13 + 6 weeks' of gestation. Fetal ultrasound biomarkers were documented in a standardised manner at the time of screening and maternal blood was analysed using standard commercial biochemistry analysers for levels of pregnancy related hormones known to be associated with occurrence of T21. The dataset contained 117 753 unaffected and 270 confirmed T21 affected pregnancies. All of the participants signed an Institutionally approved consent form specific to Aneuploidy screening. An audit and analysis of pregnancy outcome in women undergoing screening was approved by the Joint Chinese University of Hong Kong – New Territories East Cluster Clinical Research Ethics Committee (CREC Ref No. 2012.538). Table 1 summarises the anonymised data set according to T21 status. This retrospective dataset provided a record of real-world clinical use of the commonly used T21 algorithm [2] over an 8 year period. The Laboratory has previously reported that 1) it detected 90% of T21 affected pregnancies in a consistent manner since 1st trimester combined screening test was introduced in 2003; 2) that 5–6% of screened pregnancies with a risk cutoff of 1:250 are screened as high risk [16,18]. This method represents the predicate and benchmark against which the adaptive system is compared.

The anonymised dataset contained information on fetal viability, fetal nuchal translucency (NT) thickness, fetal crown rump length (CRL) and absence of major fetal abnormalities. For all patients, maternal blood samples were collected on the same day as the measurement for determination of T21-related biomarkers pregnancy-associated plasma protein A (PAPP-A) and free human chorionic gonadotropin beta (fhCGβ) concentration levels using either the KRYPTOR (ThermoFisher Scientific, Hennigsdorf, Germany) or DELFIA Xpress analyzers (PerkinElmer, Turku, Finland). Measured NT, PAPP-A and fhCGβ were converted to their MoM value using previously published expected median values in Chinese [19]. Gestational age at the time of screening was determined from CRL using a previously published Chinese dating formulae [20].

Feature selection was dictated by the clinical task, as only the variables related to T21 combined test were used in this study. Data preprocessing for this study was minimal, as we didn't want to alter the routine nature of the dataset and all data were validated against the laboratory referral form at the time of initial screening. The descriptive statistics of the feature variables in the study data are listed in Table 1.

The utilization of IL or any adaptive method requires the definition of a data block, i.e. in what increments is the data stream processed [11].

Table 1
Descriptive statistics of the feature variables. Statistical difference of cases and controls were tested with one-way ANOVA [21] for continuous variables and Chi-square test [22] for the categorical values.

	Control (N = 117753)	Trisomy 21 (N = 270)	Total (N = 118023)	p value
Maternal Age (years)				<0.001
Median (Q1, Q3)	32.45 (29.45, 35.45)	37.20 (33.84, 40.18)	32.46 (29.45, 35.46)	
Maternal Weight (kg)				0.518
Median (Q1, Q3)	54.10 (49.40, 60.20)	54.35 (49.08, 60.85)	54.10 (49.40, 60.20)	
GA sampling (wks)				0.369
Median (Q1, Q3)	88 (85, 91)	87 (85, 90)	88 (85, 91)	
Ethnicity				0.500
East Asian	116909 (99.283%)	269 (99.630%)	117178 (99.284%)	
Afro-Caribbean/ Caucasian/South Asian/Other	844 (0.717%)	1 (0.370%)	8845 (0.716%)	
Smoker				0.925
No	112379 (95.436%)	258 (95.556%)	112637 (95.436%)	
Yes	5374 (4.564%)	12 (4.444%)	5386 (4.564%)	
NT MoM				<0.001
Median (Q1, Q3)	1.02 (0.91, 1.17)	1.99 (1.48, 2.80)	1.02 (0.91, 1.17)	
fhCGβ MoM				<0.001
Median (Q1, Q3)	0.96 (0.66, 1.46)	1.76 (1.16, 2.73)	0.96 (0.66, 1.46)	
PAPP-A MoM				<0.001
Median (Q1, Q3)	1.01 (0.72, 1.41)	0.46 (0.31, 0.74)	1.01 (0.72, 1.40)	

This block size should scale to the analysis throughput of the screening lab, while considering the minimum required observations for the statistical distribution changes detection to function properly. Also, the incidence of the screened rare disorder needs to be considered when determining a proper data block size. For our research, different block sizes were estimated based on our study data, which contained 118 023 patient records or observations over 8 years, approximately 1229 observations per month. For simplifying the data block determination, the block representing a month was rounded down to 1000 observations. From there, data blocks for one day, week, quartile, half year and a year were estimated, and are represented in Table 2.

3.2. Distribution shift detection

Variables used for risk prediction can be susceptible to variance from different sources. All data have the small probability of containing data registration errors [23], which are sometimes close to impossible to recognize. Sources of variance that can be mitigated commonly relate to factors affecting sample measurement, such as laboratory-to-laboratory, instrument-to-instrument and operator-to-operator variance [24], along with seasonal effects that can affect the biochemical testing process. These are addressed to a varying degrees by the MoM procedure [5] which is used to reduce the laboratory-to-laboratory variance. If the population median is updated regularly, it can reduce other sources of variance that are affected by time. MoM is however commonly used for biochemical and biophysical measurements only [25], in order to standardize information of significant predictor of outcome.

Mother's pregnancy history and demographics are also susceptible to sources of variance, but they are usually not scaled or adapted to local differences. The proportions of different ethnicities and what is considered a common body-mass index can vary from country to country, and within country as a function of time [26]. This is evident in our study data as well; Fig. 1 depicts the monthly median of the maternal weight in screening results collected in Hong Kong over eight years, where seasonality within a year can be seen but also the steady rise of maternal weight overall.

Detection of such distribution shifts [27] and shape changes over time enables a risk system to adapt to them. A feasible detection method would monitor every feature variable, so it would have to accommodate different data types. For T21 risk prediction, feature variables consist of continuous and categorical data. Testing for differences in old and new data medians of continuous distributions can be done using a nonparametric Mood's median test [22], as the assumptions about sample variance are more relaxed when compared to the widely used one-way ANOVA [21]. As for distribution shape, a nonparametric two sample Kolmogorov-Smirnov test [28] can be used to compare the cumulative distributions of two samples. Distribution differences in categorical data on the other hand can be tested with Chi-square test of independence via a contingency table [22]. Our method consists of using these three tests for appropriate feature variables according to their data type, this is described in Table 3.

After the testing, the produced p values are adjusted with the Bonferroni correction [29]. This is done to reduce type 1 error or false positive error when conducting multiple statistical tests. After this, each of the adjusted p values are compared against a p value cutoff for significance called the global p value cutoff or GpVC, and it is a tuneable parameter of how sensitive the method is to finding differences. If any of the feature variables are found to have a distribution shift, the system continues to fitting a candidate model on the assumption that any shift is due to underlying populations change and not equipment failure. The method does not consider any prior information about the data nor the clinical relevance of the differences, it is used for finding events of numerical difference and then triggering learning and evaluation steps in the risk system. A laboratory should adapt to their own specific GpVC value, as the proper cutoff determination depends on the local patient population and the screening test used. The proposed distribution shift detection schema is depicted in Fig. 2.

Table 2
Data block sizes in the study.

Data block	Estimated time
36	Day
250	Week
1000	Month
3000	Quartile
6000	Half year
12000	Year

3.3. Incremental and transfer learning

While the data block sizes are determined by the estimated throughput of the lab, different data processing strategies can be used with IL. These relate to the way historical data is utilized; the cumulative strategy is to include all historical data during fitting the updated models, while the windowed strategy limits training data to a certain period. The sufficient amount of adaptivity demonstrated by an incrementally learning ARPS relates to the stability-plasticity dilemma [30], where the idea is that in a learning system sufficient adaptability is required for integrating new knowledge, while retaining stability to prevent forgetting of important past knowledge. Cumulative strategy can be thought to be most stable version of IL as all previous knowledge is retained, while windowed strategy forces plasticity to the model fitting process. Both strategies are experimented with our proposed ARPS.

When a shift event is detected, the system fits a candidate model from the historical data using a data processing strategy. This model is a deep fully-connected artificial neural network (DNN) with same architecture as our previously published T21 risk model, which achieved improved performance when compared to a commonly used T21 algorithm which is based on multivariate logistic regression [17]. Neural networks are also suitable for IL due to the nature of mini-batch stochastic gradient descent optimization [31], as you can continue model fitting with new data. After fitting, the performance of the candidate and current models are calculated. The chosen performance metric should be appropriate for our problem of T21 screening, which is a binary classification task with significant class imbalance because the incidence of T21 is roughly 1 in 700 [32]. Because of this, relevant literature commonly reports true positive rates (TPR) at clinically significant false positive rates (FPR) along with the more general area under the curve (AUC) from a receiver operating characteristic curve (ROC) [33]. For our model updating, the chosen metric was the partial AUC (pAUC) [34] of 0%–10% FPR, as this represents the clinically significant FPR range for this application. Plain AUC, average precision from a precision-recall curve [35] and F1 score [36] were initially investigated but not used due to lesser prediction performance when compared to pAUC of 0%–10%. The diagram of our proposed ARPS is in Fig. 3A.

The first model in ARPS that would be deployed to a screening lab without any historical data will more than likely perform poorly, either due to small training data amount because of small data block size, or that the training data contains little or no positive cases due to T21 incidence. To mitigate this, transfer learning or TL can be used with DNN models to start with an existing model fitted with data from a different problem domain [37]. The network architecture is partially frozen, and only some of the last layers are actively fitted to the study data. This way, the existing knowledge that resides in the fitted parameters of the frozen layers can be leveraged. TL has been successfully utilized in other clinical domains [38]. In our research, our previously published 2018 T21 risk model (named as NN2) [17] that is fitted to a significantly different population will serve as the basis for population-to-population TL in an IL manner that we call IL-P2P-TL, similarly to domain adaptation [39]. Class imbalance of the data during training was considered

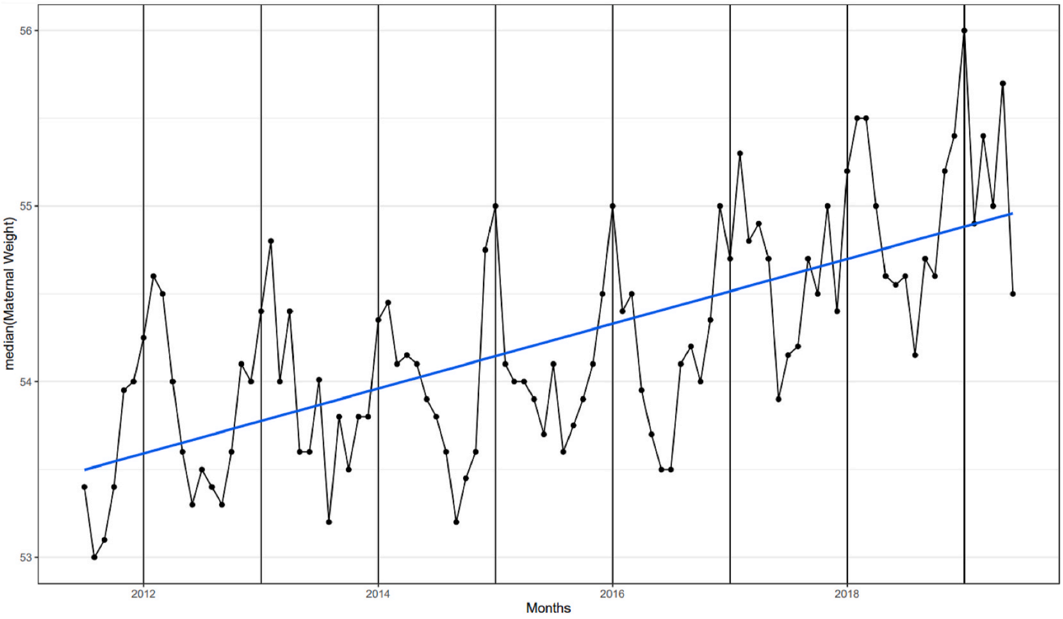

Fig. 1. Monthly median of maternal weight calculated form the study data. Each year is marked with a vertical line, and a linear regression line (blue) highlights the population change over time.

Table 3
Features variables of T21 risk prediction and statistical tests used for detecting their distribution shifts.

Feature	Test
Maternal Age	Moods median test
Maternal Weight	
GA sampling	
Ethnicity	Two sample Kolmogorov-Smirnov
Smoker	
NT MoM	Moods median test
fhCGβ MoM	
PAPP-A MoM	

by utilizing cost-sensitive learning [40]. For utilizing TL properly, same categorical levels of variables should be used. South Asian and East Asian ethnicities were recoded as Asian to achieve compatibility with our previously published model. This proposed method is depicted in Fig. 3B. A comprehensive description of our model architecture is described in the Supplementary material.

3.4. Experimental overview

Our experiments are divided into two phases: parameter and data processing strategy investigation with the proposed distribution shift detection method, and risk prediction performance evaluation of the resulting different ARPS versions. In the first phase, by conducting a grid search of a range of GpVC values and data block sizes for both

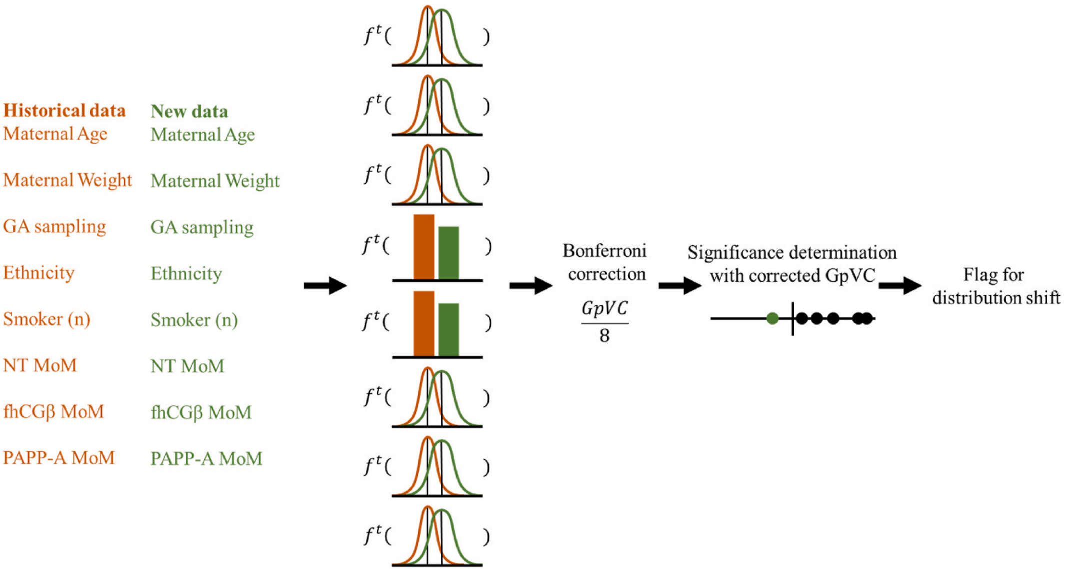

Fig. 2. Schema figure of the proposed distribution shift detection. For every feature variable, appropriate statistical test is used between historical and new data for determining the statistical significance of the difference. The chosen global p value cutoff or GpVC is then corrected of type 1 error with Bonferroni correction. After this, the values are compared against the corrected GpVC, and if any of them is at or below it then the new data is flagged for distribution shift. The GpVC can be adjusted for determining the sensitivity of the detection.

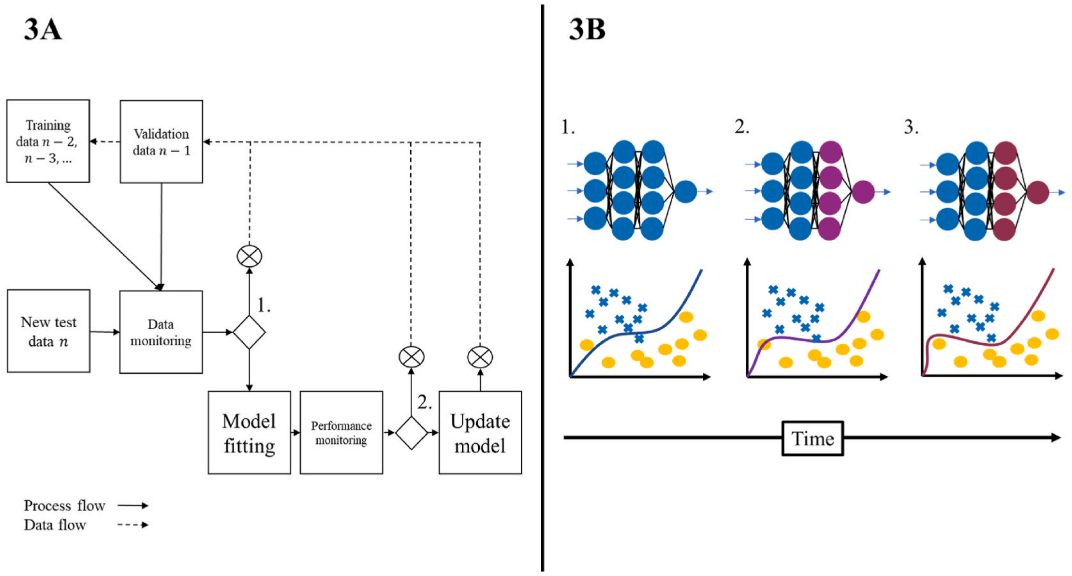

Fig. 3. ARPS diagram is depicted in 3A. New data n is introduced for data monitoring, where it is compared to existing data. If a distribution shift of any feature has been detected (1.), a candidate model is fitted. This can mean fitting a completely new model or utilize TL. $n-2$ data is used for training and $n-1$ for validation, while new data n is used for testing the performance of the current and candidate models with the pAUC metric. After this, if the candidate model produced better prediction results (2.), the current model is replaced by the candidate model. All the diagram outcomes trigger a data update procedure (dashed arrows), where new data n becomes validation data and $n-1$ is added to the training data. IL-P2P-TL is illustrated in 3B. Initially, a pretrained neural network model fitted with a different patient population is used by the lab as a starting point (1.). The first two layers of the network are freezed, i.e. their weights and biases are not updated, and this knowledge is retained. As more new data is introduced, the last layers of the network are fitted to the specific patient population (2.), until the model has adapted and surpassed the original version in terms of prediction performance (3.).

cumulative and 2-block windowed strategies while iterating through the whole study dataset, the number of detected distribution change events will be investigated. Initial testing revealed that the window sizes of three or less behaved similarly, while window sizes over three behaved similarly to the cumulative strategy, so the window size of two was chosen for experimentation. In the second phase, all combinations of distribution shift detection parameters, data processing strategies and TL utilization will be compared against the predicate method and our previously published DNN model. This will highlight whether high-stability or high-plasticity IL is beneficial, if TL is beneficial and does the best performing automatic ARPS match the performance of the predicate. In our experiments, we assume that the outcome information of the patient arrives at the same time. This is not the case in real-life screening however, where the real outcome arrives to the lab with some delay. In routine use, the distribution shift detection would function the same, while training a candidate model would wait for the outcome information to arrive. Supplementary material contains the list of used software libraries and hardware.

4. Results

4.1. Distribution shift detection

The distribution shift detection method was tested with various data block sizes representing different time points estimated from lab throughput, along with the GpVC values of 0.05–0.5 by 0.05 increments. These parameters with the resulting amounts of data distribution shift events were visualized as separate heatmaps shown in Fig. 4. From this plot we can see that the two strategies behave similarly with the block size of 36 or one day, while the windowed strategy detected smaller amount of shifts with all of the other block sizes and throughout the whole p value range. The most notable differences were with block sizes of 1000 and 3000. As the data block size was increased, the number of detected shifts decreased with the cumulative method. This was to be expected, as the number of total possible shifts decreases. However, the windowed strategy results in Fig. 4B demonstrate a decrease of detected shifts at data block of 3000, and the number of shifts is increased after this in data block of 1000. The authors speculate that seasonal effect components which can be seen from comparing the study data month to month and half-year to half-year are diluted with the data block of 3 months, and thus the window strategy finds less shifts. As for the GpVC value, when it is increased the number of detected shifts increased with both methods. This was also expected behaviour, as the cutoff value for significance becomes more lenient. All the parameter combinations presented in Fig. 4 heatmaps were experimented with in the second phase of the study, so that the relationship between the number of detected shifts and the overall method prediction performance could be investigated. The full description of the results is depicted in the Supplementary material.

4.2. System evaluation

Phase one of the experimentation demonstrated that the cumulative and windowed data processing strategies generated significantly different amount of detection events and therefore affect the systems adaptability. The strategy along with the utilization of TL produced four candidate system designs: cumulative data processing with and without TL, and windowed data processing with and without TL. These four were all used to process through the study data, and their AUC performance at each data block and over the whole data were compared against the predicate methods. Comparing against our previously published model would demonstrate if TL is beneficial and comparing against the labs screening method would demonstrate if IL is beneficial and can the ARPS reach similar performance automatically. Every combination of GpVC and data block size were experimented with, these results are listed in the Supplementary material. GpVC value was found to be more meaningful for window and window TL, as different values had no significant effect on overall AUC with cumulative strategy -based methods. This is understandable; during the simulation as the old data set cumulates into big enough size that when it is compared against one data block for statistically significant differences, they are not produced due to the sample size differences. Data block size on the other hand can

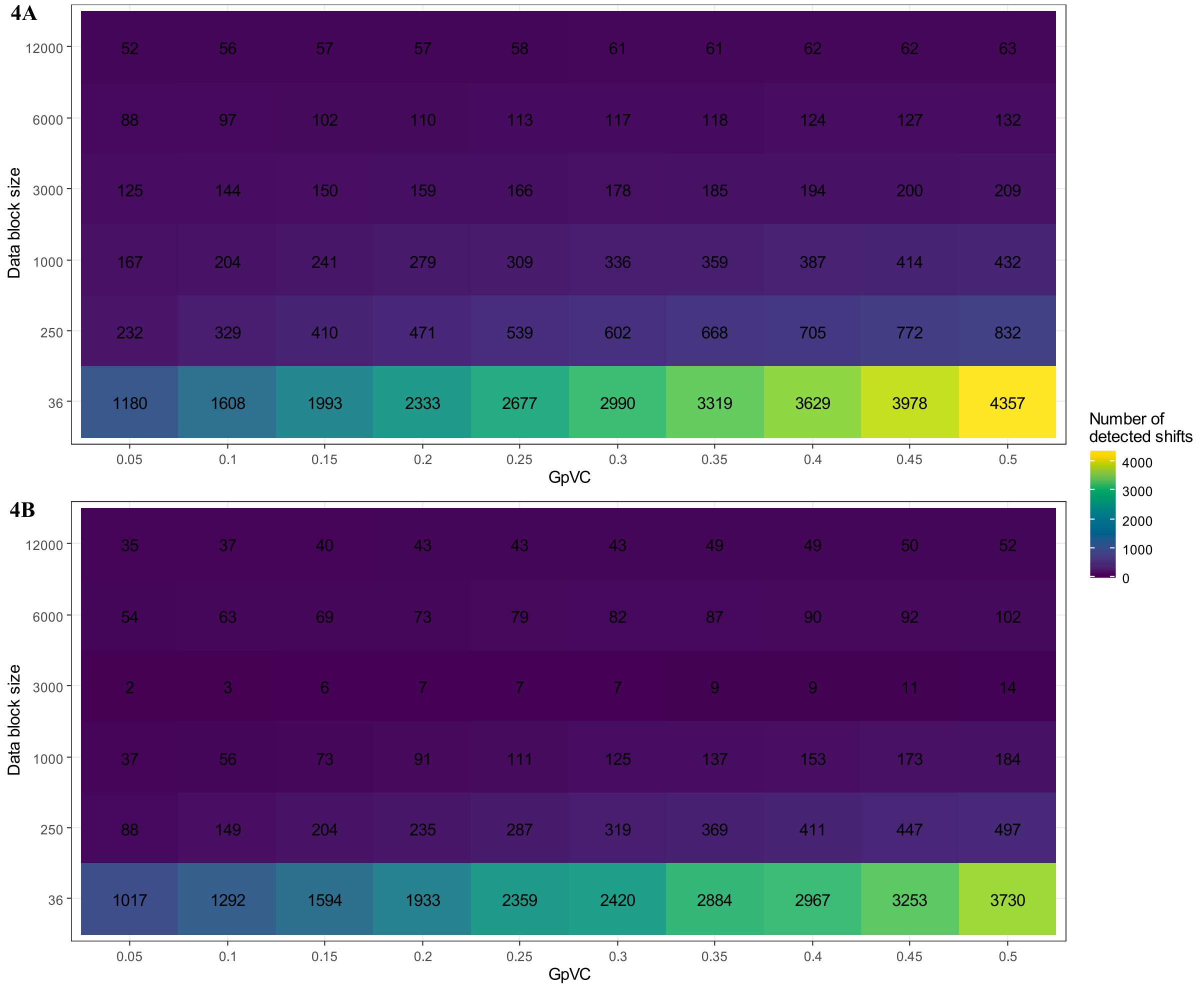

Fig. 4. Distribution shift results of the cumulative (4A) and window strategies (4B) as heatmaps. Number of detected shifts is plotted against data block size (Y-axis) and GpVC (X-axis). With both data processing strategies the number of events increases as the GpVC is increased, or as the p value cutoff gets more lenient. However, the size of the data block contributes more to the overall number of events for both strategies, as the number of maximum possible events decreases as the block size increases.

be seen as a limiting factor, as the usage of bigger data blocks result into better overall AUC with all of the predicate methods. However, laboratories that have not yet cumulated years' worth of data cannot utilize such data block sizes. The amount of data required to achieve clinically acceptable performance as early as possible is decreased when TL is used, as the window TL method achieves the performance of 0.96 AUC with GpVC of 0.05 and data block size of just 1000. These results along with all the other candidate and predicate methods are depicted in Fig. 5 as a function of time. The plot showcases the fast adaptability of TL, as it provides the candidate models with the means to give predictions of feasible performance during the early phases of adaptation (data blocks from 1 to roughly 30). Without it, the cumulative system requires a significant amount of data for achieving similar performance that of the predicates, while the windowed strategy is the most unstable.

Fig. 5 also demonstrates that with any candidate and benchmark models, at some time point some data block's AUC is less than ideal, thus producing an outlier value. We believe this is because the block contains one or more T21 positive observation which is abnormal in terms of feature variables, and thus the system produces suboptimal AUCs for that block. The number of positive observations is limited due to the incidence of T21, thus making these observations highly impactful when calculating AUC of a block. The candidate system with a top median AUC and the best IQR was the windowed system with transfer learning.

For the final comparison, performances over the whole study data with 0.05 GpVC and 1000 data block were investigated by calculating ROC curves. These results are listed in Table 4. Compared to our 2018 model that was used as the backbone for TL, windowed TL system marginally improved the prediction performance, indicating that transfer learning with windowed incremental learning is beneficial, while transfer learning with cumulative data strategy was disadvantageous. Systems without transfer learning also performed poorly compared to the 2018 model. The screening lab algorithm performs as previously reported.

5. Discussion

Deploying an adaptive risk prediction system to a lab requires careful consideration of the parameters regarding the distribution shift detection and model updating. Window strategy -based system enables the laboratory to start utilizing risk prediction early after starting their operation, and with TL the performance would reach a clinically acceptable state faster. The determination of GpVC for proper distribution shift detection is critical in this case, and less so with cumulative strategy -based methods. Also, the data block size should scale with the throughput of the lab while considering the minimum amount of data needed for probabilistic modelling, as the incidence of rare disorders dictate the number of positive observations which are available for training models. The performance metric of model updating should also

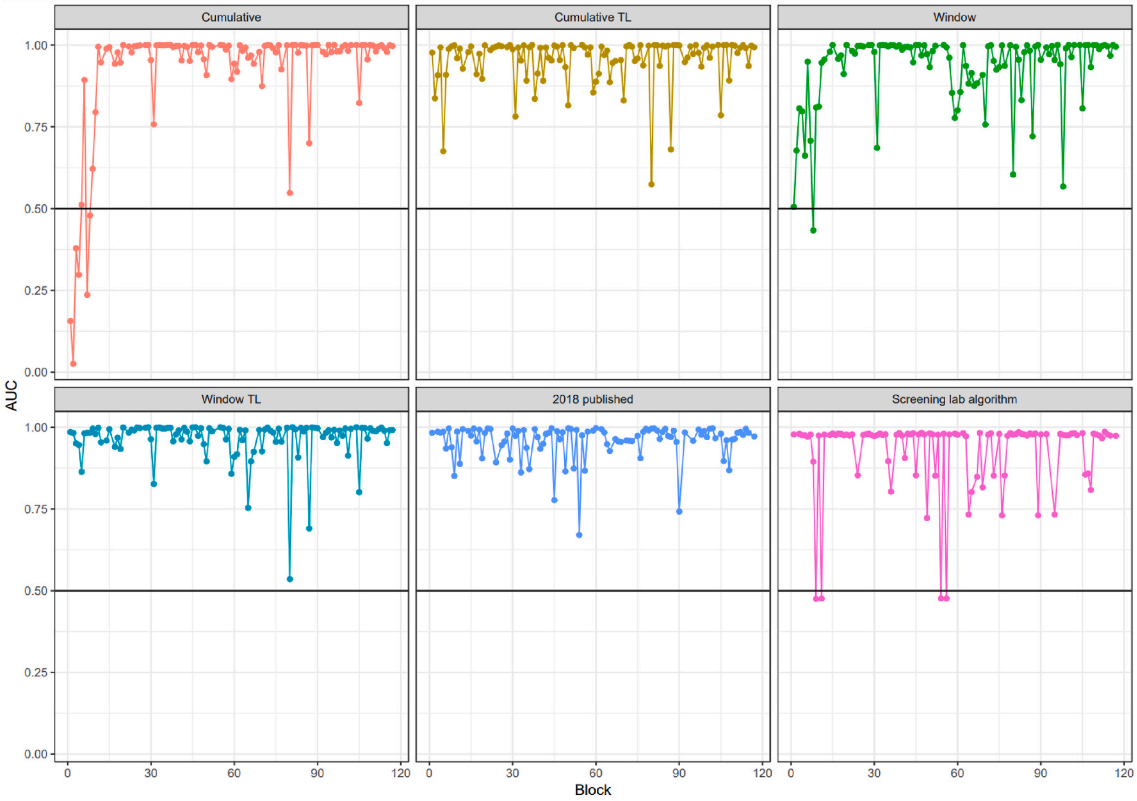

Fig. 5. AUC prediction performance as a function of time for all candidate systems and benchmark models, results per data block. The "coin flip" value of 0.5 is depicted as a horizontal line.

Table 4
Prediction performance of the predicates and candidate models calculated over the whole study data. Candidate models were parameterized with 0.05 GpVC and 1000 data block. AUC and bootstrapped 95% confidence interval (1000 bootstrap replicates) are presented, along with TPRs of 5% and 10% FPR. The best performing candidate model results are highlighted in bold.

	AUC (95% CI)	TPR of 1% FPR	TPR of 3% FPR	TPR of 5% FPR	TPR of 8% FPR	TPR of 10% FPR	TPR of 15% FPR
Predicates							
Screening lab algorithm	0.98 (0.98,0.99)	70%	85%	91%	94%	95%	98%
2018 published model	0.96 (0.95,0.97)	44%	69%	79%	84%	87%	92%
Candidates							
Cumulative system	0.92 (0.89,0.94)	**65%**	74%	77%	80%	81%	84%
Cumulative system with TL	0.90 (0.88,0.92)	52%	66%	71%	75%	77%	81%
Windowed system	0.89 (0.87,0.91)	50%	59%	61%	65%	68%	72%
Windowed system with TL	**0.96 (0.95,0.97)**	60%	**75%**	**82%**	**88%**	**90%**	**93%**

be adapted to the problem domain, where it would best represent the clinical significance. For the prediction task of T21, the pAUC of 0%–10% was experimented to yield best overall prediction performance.

Our empirical experimentation shows that when we utilize distribution shift detection, the window data processing strategy provides a computationally less expensive alternative to continuously cumulating training data that also does not perform worse in terms of prediction performance. TL has shown to enable leveraging models developed with different patient populations as a backbone for adapting to the local population over time with IL. The ARPS can adapt to data over time in a clinically significant way, however closer inspection of the learning model's fitted parameters would be the topic of future research, in addition to method generalization. We have demonstrated in the past that neural network models can perform better when compared to multivariate logistic models that are routinely used in this domain [17], however the explainability of neural network model fits and predictions are not at the same level as with logistic regression. We believe that this trade-off will be less severe in the future, as more effort is currently put into making neural network models more transparent [41].

The screening lab algorithm performed the best overall, however behind this performance there is extensive adjustment work done by one or more laboratory technician over the period of multiple years. With our best performing candidate system, we demonstrated on par performance of an adapting system that requires no human intervention. Few screening laboratories possess individuals with such analytical skills that can revise the analysis and understand the impact of those changes on screening performance. This highlights the impact of our proposed method, as screening labs with limited resources could utilize our solution and actively improve their screening performance.

The prediction of T21 was experimented with our proposed ARPS, however with proper parameters it could be utilized to other problem domains. Relating to this, we produced promising results of using our published T21 prediction model as a backbone for improving detection of other chromosomal abnormalities such as T13 and T18 [42], the results are appended in the Supplementary material. The core functionality of ARPS could also be extended further. Multiple different models fitted for different block sizes could be used to account for small and big time frame changes in the data. We have successfully applied ensemble learning in the past [43], this could also be used to increase robustness or performance. Routine screening data is also commonly highly imbalanced due to the incidence of a particular outcome, for this problem we have previously published a GAN-based solution [44] which could be integrated into ARPS. Equations used in the MoM process that standardize the biophysical or biochemical measurement by factors such as gestational age could also be adjusted within ARPS.

The main limitations of our study relate to method generalization. While our study data is extensive, it was gathered from one laboratory representing one patient population and region of sample collection. Collecting a data set of similar size from a different setting and testing the generalization of ARPS and its individual components would be the next steps of our research.

We believe that TL and IL as methods are at the maturity level that they can be utilized for improving clinical risk prediction, as demonstrated by our research. IL-P2P-TL could enable a screening lab to start with an established prediction model, but over time adapt to their local population and improve the detection of positive cases. There is also a possibility of starting with a model for one outcome, and slowly branch out new TL models that are adapted to other outcomes, which are routinely so rare that feasible models cannot be constructed otherwise. By developing our adaptive system build on familiar concepts to clinical practitioners, we believe that ARPS could be a practical way to improve risk assessment related to clinical screening in general.

6. Conclusion

In this paper we propose a novel adaptive risk prediction system for the risk assessment of T21. We demonstrate with a sufficiently large real-life dataset that our best performing design can achieve on par prediction performance without human intervention, when compared to a deployed risk screening algorithm that has been manually updated. For this system to operate in a clinical setting with a feasible stability and plasticity, we also propose a novel distribution shift detection method that can be parameterized to fit the required sensitivity of the problem domain.

Declaration of competing interest

None declared.

Acknowledgement

The study had no funding from any sources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104886.

References

[1] M.E. Weijerman, J.P. de Winter, Clinical practice. The care of children with Down syndrome, Eur. J. Pediatr. 169 (2) (2010) 1445–1452.

[2] D. Wright, K.O. Kagan, F.S. Molina, A. Gazzoni, K.H. Nicolaides, "A mixture model of nuchal translucency thickness in screening for chromosomal defects," *Ultrasound in obstetrics & gynecology*, the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 31 (4) (2008) 376–383.

[3] P. Royston, S.G. Thompson, Model-based screening by risk with application to Down's syndrome, Stat. Med. 11 (2) (1992) 257–268.

[4] K.O. Kagan, D. Wright, K. Spencer, F.S. Molina, K.H. Nicolaides, "First-trimester screening for trisomy 21 by free beta-human chorionic gonadotropin and pregnancy-associated plasma protein-A: impact of maternal and pregnancy characteristics," *Ultrasound in obstetrics & gynecology*, the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 31 (5) (2008) 493–502.

[5] N. Wald, K.H. Nicolaides, The detection of neural tube defects by screening maternal blood, Prenat. Diagn. (1976) 227–238.

[6] K.G. Moons, A.P. Kengne, D.E. Grobbee, P. Royston, Y. Vergouwe, D.G. Altman, M. Woodward, Risk prediction models: II, External validation, model updating, and impact assessment," *Heart* 98 (9) (2012) 691–698.

[7] S. Mukherjee, P. Niyogi, T. Poggio, R. Rifkin, Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization, Adv. Comput. Math. 25 (2006) 161–193.

[8] I.G. Stiell, C. Bennett, Implementation of clinical decision rules in the emergency department, Acad. Emerg. Med. 14 (11) (2007) 955–959.

[9] M. Carolan, The graying of the obstetric population: implications for the older mother, J. Obstet. Gynecol. Neonatal Nurs. : J. Obstet. Gynecol. Neonatal Nurs. 32 (1) (2003) 19–27.

[10] D.Y. LaCoursiere, L. Bloebaum, J.D. Duncan, M.W. Varner, Population-based trends and correlates of maternal overweight and obesity, Utah 1991-2001, Am. J. Obstet. Gynecol. 192 (3) (2005) 832–839.

[11] L. Bruzzone, D. Fernàndez Prieto, An incremental-learning neural network for the classification of remote-sensing images, Pattern Recogn. Lett. 20 (11–13) (1999) 1241–1248.

[12] E. Hajiramezanali, S.Z. Dadaneh, A. Karbalaygh areh, M. Zhou, X. Qian, Bayesian Multi-Domain Learning for Cancer Subtype Discovery from Next-Generation Sequencing Count Data, *arXiv preprint*, 2018, p. 1810, 09433.

[13] J.R. Finkel, C.D. Manning, Hierarchical bayesian domain adaptation, in: Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 602–610.

[14] I.B. Arief-Ang, F.D. Salim, M. Margaret Hamilton, DA-HOC: semi-supervised domain adaptation for room occupancy prediction using CO2 sensor data, in: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys '17), Association for Computing Machinery, 2017, pp. 1–10.

[15] M. Sugiyama, M. Kawanabe, Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation, MIT press, London, 2012.

[16] D.S. Sahota, W.C. Leung, W.P. Chan, W.W. To, E.T. Lau, T.Y. Leung, Prospective assessment of the Hong Kong Hospital Authority universal Down syndrome screening programme, Hong Kong medical journal = Xianggang yi xue za zhi 19 (2) (2013) 101–108.

[17] A. Koivu, T. Korpimäki, P. Kivelä, T. Pahikkala, M. Sairanen, Evaluation of machine learning algorithms for improved risk assessment for Down's syndrome, Comput. Biol. Med. 98 (2018) 1–7.

[18] T. Leung, L. Chan, L. Law, D. Sahota, T.e. a. Fung, "First trimester combined screening for trisomy 21 in Hong Kong: outcome of the first 10,000 cases," *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 (4) (2009) 300–304.

[19] D.S. Sahota, T.Y. Leung, T.Y. Fung, L.W. Chan, L.W. Law, et al., "Medians and correction factors for biochemical and ultrasound markers in Chinese women undergoing first trimester screening for trisomy 21," *Ultrasound in obstetrics & gynecology*, the official journal of the International Society of Ultrasound in Obstetrics and Gynecology 33 (4) (2009) 387–393.

[20] D.S. Sahota, T.Y. Leung, T.N. Leung, O. Chan, T.K. Lau, "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 Ultrasound in Obstetrics and Gynecology 33 (2) (2009) 157–160.

[21] D. Howell, Statistical Methods for Psychology, Duxbury, Pacific Grove, 2002.

[22] S. Siegel, N.J.J. Castellan, Nonparametric Statistics for the Behavioral Sciences, McGraw–Hill, New York, 1988.

[23] S.I. Goldberg, A. Niemierko, A. Turchin, Analysis of data errors in clinical research databases, in: AMIA Annual Symposium Proceedings, 2008. Washington, DC.

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

[25] J.C. Bishop, F.D. Dunstan, B.J. Nix, T.M. Reynolds, A. Swift, All MoMs are not equal: some statistical properties associated with reporting results in the form of multiples of the median, Am. J. Hum. Genet. 52 (2) (1993) 425–443.

[26] P.T. James, R. Leach, E. Kalamara, M. Shayeghi, The worldwide obesity epidemic, Obes. Res. 9 (S11) (2001) 228S–233S.

[27] J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recogn. 45 (1) (2012) 521–530.

[28] M.A. Stephens, EDF statistics for goodness of fit and some comparisons, J. Am. Stat. Assoc. 69 (347) (1974) 730–737.

[29] C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936.

[30] M. Mermillod, A. Bugaiska, P. Bonin, The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects, Front. Psychol. 4 (2013) 504.

[31] M. Li, T. Zhang, Y. Chen, A.J. Smola, Efficient mini-batch training for stochastic optimization, in: Proceedings Of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014. New York.

[32] C.T. Mai, J.L. Isenburg, M.A. Canfield, R.E. Meyer, A. Correa, C.J. Alverson, P. J. Lupo, T. Riehle-Colarusso, S.J. Cho, D. Aggarwal, R.S. Kirby, N.B.D.P. Network, National population-based estimates for major birth defects, 2010-2014, Birth defects research 111 (18) (2019) 1420–1435.

[33] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (8) (2006) 861–874.

[34] L.E. Dodd, M.S. Pepe, Partial AUC estimation and regression, Biometrics 59 (3) (2003) 614–623.

[35] E. Zhang, Y. Zhang, Average precision, in: Encyclopedia Of Database Systems, Springer US, Boston, MA, 2009, pp. 192–193.

[36] D.M. Powers, Evaluation: from Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation, 2020 arXiv preprint, arXiv:2010.16061.

[37] S. Bozinovski, Reminder of the first paper on transfer learning in neural networks, Informatica 44 (3) (1976) 291–302.

[38] C. Shie, C. Chuang, C. Chou, M. Wu, E.Y. Chang, Transfer representation learning for medical image analysis, in: 37th Annual International Conference Of the IEEE Engineering In Medicine And Biology Society, EMBC), Milan, Italy, 2015.

[39] I. Redko, E. Morvant, A. Habrard, M. Sebban, Y. Bennani, Advances in Domain Adaptation Theory, ISTE Press - Elsevier, 2019.

[40] C. Ling, V. Sheng, Cost-sensitive learning and the class imbalance problem, in: Encyclopedia Of Machine Learning, Springer, New York, 2010, pp. 231–235.

[41] R. Roscher, B. Bohn, M.F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216.

[42] R. Qiang, N. Cai, X. Wang, L. Wang, K. Cui, W. Wang, X. Wang, X. Li, Detection of trisomies 13, 18 and 21 using non-invasive prenatal testing, Experimental and therapeutic medicine 13 (5) (2017) 2304–2310.

[43] A. Koivu, M. Sairanen, Predicting risk of stillbirth and preterm pregnancies with machine learning, Health Inf. Sci. Syst. 8 (14) (2020) 1–12.

[44] A. Koivu, M. Sairanen, A. Airola, T. Pahikkala, Synthetic minority oversampling of vital statistics data with generative adversarial networks, J. Am. Med. Inf. Assoc. 27 (11) (2020) 1667–1674.