Heliyon 9 (2023) e16244

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Predicting facility-based delivery in Zanzibar: The vulnerability of machine learning algorithms to adversarial attacks

Yi-Ting Tsai [a,*], Isabel R. Fulcher [b,c], Tracey Li [d], Felix Sukums [e], Bethany Hedt-Gauthier [a,b]

[a] Department of Biostatistics, Harvard Chan School of Public Health, Boston, USA
[b] Department of Global Health and Social Medicine, Harvard Medical School, Boston, USA
[c] Harvard Data Science Initiative, Harvard University, Cambridge, USA
[d] D-tree International, Zanzibar, Tanzania
[e] Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania

ARTICLE INFO

Keywords:
Machine learning
Adversarial attack
Digital health
Maternal health
Facility delivery
Community health worker intervention

ABSTRACT

Background: Community health worker (CHW)-led maternal health programs have contributed to increased facility-based deliveries and decreased maternal mortality in sub-Saharan Africa. The recent adoption of mobile devices in these programs provides an opportunity for real-time implementation of machine learning predictive models to identify women most at risk for home-based delivery. However, it is possible that falsified data could be entered into the model to get a specific prediction result – known as an "adversarial attack". The goal of this paper is to evaluate the algorithm's vulnerability to adversarial attacks.

Methods: The dataset used in this research is from the *Uzazi Salama* ("Safer Deliveries") program, which operated between 2016 and 2019 in Zanzibar. We used LASSO regularized logistic regression to develop the prediction model. We used "One-At-a-Time (OAT)" adversarial attacks across four different types of input variables: binary – access to electricity at home, categorical – previous delivery location, ordinal – educational level, and continuous – gestational age. We evaluated the percent of predicted classifications that change due to these adversarial attacks.

Results: Manipulating input variables affected prediction results. The variable with the greatest vulnerability was previous delivery location, with 55.65% of predicted classifications changing when applying adversarial attacks from previously delivered at a facility to previously delivered at home, and 37.63% of predicted classifications changing when applying adversarial attacks from previously delivered at home to previously delivered at a facility.

Conclusion: This paper investigates the vulnerability of an algorithm to predict facility-based delivery when facing adversarial attacks. By understanding the effect of adversarial attacks, programs can implement data monitoring strategies to assess for and deter these manipulations. Ensuring fidelity in algorithm deployment secures that CHWs target those women who are actually at high risk of delivering at home.

* Corresponding author. 5 Columbia St, Apt 612, Cambridge, MA, 02139, USA.
E-mail address: yitingtsai@hsph.harvard.edu (Y.-T. Tsai).

https://doi.org/10.1016/j.heliyon.2023.e16244
Received 4 October 2022; Received in revised form 1 May 2023; Accepted 10 May 2023
Available online 13 May 2023

1. Introduction

The decrease in maternal mortality in sub-Saharan Africa (SSA) over the past two decades [1–3] is partially attributed to the increasing number of women delivering at healthcare facilities where women can receive immediate emergency care if needed [4–16]. Community health worker (CHW) interventions have been a crucial factor for connecting women to health facilities for facility-based delivery, particularly in low- and middle-income countries [17–21].

The recent adoption of digital tools, such as mobile devices [22–27], in CHW-led maternal health programs, provides an opportunity for real-time implementation of predictive models [28–35]. These predictions can then assist the implementation of targeted interventions for those most at risk for home-based delivery. Other groups have led work to predict facility-based delivery in SSA [29], but to our knowledge, none of these algorithms are currently integrated into program implementation. As programs transition to integrating algorithms into targeted maternal health care, it is important to understand the vulnerability of these algorithms to incorrect or manipulated data.

Our previous research applied machine learning techniques to predict health facility delivery among pregnant women enrolled in *Uzazi Salama* ("Safer Deliveries"), a CHW-led, digitally-supported maternal health program in Zanzibar, Tanzania [36,37]. The overarching goal is to use information recorded at the program enrollment visit to provide a real-time prediction of delivery location. Women identified as high risk of delivering at home can then immediately receive additional support, such as the CHWs providing additional birth planning counseling or women receiving money to cover transport costs.

Given that this algorithm would be implemented in real-time, it is possible that falsified data could be entered into the model [38] in order to get a specific prediction result. For example, a CHW may try to influence a prediction of home-based delivery to have a heavier workload for which they are paid more or to influence a prediction of facility-based delivery to avoid providing additional services. On the other hand, a woman may try to influence a prediction of home-based delivery with the goal of receiving more support. In this paper, we evaluate this algorithm's vulnerability to falsified responses to produce a specific prediction, often referred to as "adversarial attacks" in the machine learning literature [39–43]. The findings from this analysis will enable us to quantify the susceptibility of our developed algorithm to adversarial attacks, which can be used to inform methods to monitor for or mitigate the impact of these attacks when the algorithm is deployed.

2. Material and methods

2.1. Description of Uzazi Salama and the resulting dataset

The *Uzazi Salama* program was established in 2016 in Zanzibar to reduce maternal and neonatal mortality by increasing facility delivery rates and to support women during their antenatal care period. In January 2020, the program was expanded to provide additional postpartum and child health services, and was renamed *Jamii ni Afya* ("Communities are Health"). Women enrolled in this program were assigned CHWs, who provided women three home visits during pregnancy to provide birth planning support, and three visits during the postpartum period to monitor the woman's recovery and child's health.

In this study, we included women enrolled in *Uzazi Salama* between January 1, 2017 and June 30, 2019. After excluding women who delivered on the way to the facility and women with no birth location documented, we have a sample size of 38,787, of which 77% of women delivered in a health facility. A wide range of information is available for the women at the time of registration, including both individual- and community-level variables. Individual-level variables are those depending on a specific woman, such as demographic characteristics, self-reported clinical characteristics, current pregnancy characteristics, and program-associated characteristics. Community-level variables are calculated based on program-level information including all women who had been enrolled and delivered up to the time of registration of the specific women under consideration, such as the rate of facility delivery in a woman's shehia (local area).

2.2. Machine learning model development

To develop the algorithm, we followed the training and data pre-processing procedure outlined in previous work [37]. We randomly sampled 80% of the data, stratified by home or facility delivery, to form the *training set*, while the remaining 20% of data served as the *test set*. Missing values were imputed by K-Nearest Neighbor (KNN) imputation [44–46] with $K = 5$, which used observations from the five most "similar" women to impute a plausible value for missing data. In order to calculate the "similarity" between women, categorical and ordinal variables need to be transformed into numerical representations. For categorical variables, we performed one-hot encoding, where we created a new column for each category within a variable, and assigned 1 if a woman belongs to this category, and 0 otherwise. For ordinal variables, we assigned each category with an integer, starting from 1, according to the order of the categories.

Three different techniques were employed to address the outcome imbalance problem in earlier work: undersampling, oversampling, and a synthetic minority oversampling technique (SMOTE) [37]. Since they all had similar prediction accuracy and the sample size is sufficiently large, we used an undersampling approach during the model training phase by randomly selecting a subset of those who delivered in a health facility of an equal size to the number who delivered at home in the training dataset.

Earlier work explored four prediction models [37], including logistic regression, logistic regression with LASSO regularization, random forest, and neural network, all resulting in similar performance properties, with overall accuracy ranging between 72.4% and 74.6%. In this paper, the LASSO regularized logistic regression [47] is used as our prediction model because of its high performance

and feasibility for "real-time" predictions on smartphones. The overall accuracy of the final prediction model is 73.7%, with sensitivity at 71.0%, specificity at 74.5%, and area under the receiver operating characteristic (ROC) curve at 0.799.

For the final model, the ten variables that best discriminated between home deliveries and facility deliveries were as follows: previous delivery location, rate of facility delivery among women with the same CHW, rate of facility delivery in shehia, name of the recommended delivery facility, residential district, parity (categorized as 0 previous births, 1–2 previous births, 3–4 previous births, 5–7 previous births, and 8 or more previous births), education level, floor material, access to electricity at home, and payment fee charged at first antenatal care (ANC) visit.

The values for the variables parity and previous delivery location are related, since women with parity zero would not have a previous delivery location. To avoid overlap in the two variable values, we combined these two variables into an interaction term. In this new variable, parity is represented in the tenth digit: 0* indicates no previous births, 1* indicates 1–2 previous births, 2* indicates 3–4 previous births, 3* indicates 5–7 previous births, and 4* indicates 8 or more previous births. Previous delivery location is on the units digit: *1 indicates previously delivered at home, *2 indicates previously delivered on the way to a facility, *3 indicates previously delivered at a facility, and *4 indicates no previous delivery information. This categorizes women with parity equals zero and women without a previous delivery location into the same group "04". In modeling, this variable is treated as a non-ordinal categorical variable.

2.3. Adversarial attacks

2.3.1. Definition

The action of trying to deceive a prediction model by using adversarial samples is called "adversarial attack", where adversarial samples are manipulated inputs to a machine learning (ML) model that result in erroneous outputs [39]. The attacks can be categorized into various categories based on the goal of the adversary and the stage at which the adversarial samples are used [39].

In this paper, we are exploring a "targeted, black-box evasion attack", where an individual tries to change the output of the ML classifier to a specific class (hence, targeted), has no knowledge about the ML model (hence, black-box), and only has influence on the model execution and not the model development (hence, evasion). For our specific problem, the prediction model will already be developed and deployed on the phone, and the CHW or woman may alter their replies to influence the final classification. Even though both individual-level and community-level variables are included in the prediction model, only individual-level variables are candidates for adversarial attacks, since the community-level variables are fixed and could not be easily manipulated by CHWs.

Table 1
Data manipulation process and variables remained in the LASSO logistics regression model by input variable type.

Binary variables	
Description: Yes/No response	
Manipulation: Change a Yes into a No or a No into a Yes	
Variables:	
Prior c-section	RCH card obtained at visit
Access to electricity at home†	HIV status is unknown
Drinking water source	Partner gave permission for facility delivery
Roof material	Danger sign mother: abdominal pain
Savings acquisition type: personal savings	Danger sign mother: vaginal bleeding
Savings acquisition type: Husband/relatives	Danger sign mother: severe headache
Savings acquisition type: VSLA loan	Danger sign mother: difficult breathing
Savings acquisition type: Other	
Categorical variables	
Description: Multiple categories without order	
Manipulation: Change each category into all the other possible categories	
Variables:	
District*†	The time of CHW experience (in months)*
Delivery facility name recommendation†	The interaction term between parity† and previous delivery location†
Delivery facility type recommended	
Ordinal variables	
Description: Multiple ordered categories	
Manipulation: Change all the data into the highest/lowest possible level	
Variables:	
Education level†	Total ANC visits prior to enrollment
Continuous variables	
Description: Any real number	
Manipulation: Add/Subtract a fixed number and stop if the upper/lower limit is reached	
Variables:	
Facility delivery rate based on shehia*†	The product of delivery rate based on shehia and CHW*
Facility delivery rate based on CHW*†	
Gestational age at enrollment	

*indicates a community-level variable (not manipulable); †indicates one of the ten most influential variables.

2.3.2. Notation

We use x to denote the value of the manipulated variable and an arrow (→) to denote the direction that the variable is manipulated. We use $\hat{y} = 1$ to indicate that the woman was originally (before an attack) predicted to have facility delivery, and $\hat{y} = 0$ to indicate that the woman was originally predicted to have home delivery. We use y to denote the value of the prediction after an attack.

2.3.3. Constructing an adversarial attack

To quantify the sensitivity of a model to an adversarial attack, we will manipulate one input variable and hold all other variables fixed, known as "One-At-a-Time (OAT)" adversarial attack [47–49]. We consider different data manipulation processes for four types of input variables: binary, categorical, ordinal, and continuous (Table 1).

For demonstration purposes, we manipulate one binary variable (access to electricity at home), one categorical variable (previous delivery location), one ordinal variable (education level), and one continuous variable (gestational age at enrollment) respectively for our adversarial attacks. The criterion for choosing these variables includes: inclusion in the final LASSO model (Table 1), being a top ten most predictive variable, and being an individual-level variable, as community-level variables are fixed on phones and could not be misreported. Although gestational age is not in the top ten most predictive variables, we still included it because no other continuous variable was among the ten most predictive variables and it was the most predictive of any of the continuous variables. Note that the gestational age data used in this study is the reported gestational age, not the actual gestational age verified by ultrasound, which may be inconsistent in some cases [50].

Manipulating access to electricity. Let $x = 1$ indicate a household has access to electricity, and $x = 0$ indicate a household has no access to electricity. In the evaluation, we changed all women whose households have access to electricity to not having access to electricity ($x = 1 \rightarrow 0$), and vice versa ($x = 0 \rightarrow 1$).

Manipulating previous delivery location. Let $x = 1$ indicate a woman previously delivered at home, $x = 2$ indicate a woman previously delivered on the way to a facility, and $x = 3$ indicate a woman previously delivered at a facility. In the evaluation, we manipulate the previous delivery location in each subset into the other two categories. For example, in the subset with $x = 1$, the manipulation includes $x = 1 \rightarrow 2$ and $x = 1 \rightarrow 3$.

Manipulating education level. Let $x = 1$ indicate having some primary education, $x = 2$ indicate completed primary education, $x = 3$ indicate having some secondary education, $x = 4$ indicate completed secondary education, and $x = 5$ indicate having higher education. In the evaluation, we change all the non-extreme education levels into the most primary one ($x = 2,3,4,5 \rightarrow 1$) or the most advanced one ($x = 1,2,3,4 \rightarrow 5$).

Manipulating gestational age at enrollment. Let x be the gestational age at enrollment in weeks, with values ranging from 1 week to 40 weeks. We set $x = 10$ as the lower limit and $x = 30$ as the upper limit of our data manipulation threshold. In the evaluation, we then conduct three different adversarial attacks scenarios, where the manipulation magnitude increases from (a) to (c):

(a) Add/Subtract 5 to each x and stop if the upper/lower limit is reached
(b) Add/Subtract 10 to each x and stop if the upper/lower limit is reached
(c) Change each x directly into the upper/lower limit

2.3.4. Evaluating an adversarial attack

We show the impact of adversarial attacks under various conditions based on the manipulated variable x and their values of the original prediction $\hat{y}$. To quantify the susceptibility of algorithms, we report the "attacker success rate" as the proportion of prediction results changed due to the attack within a subset of women. The higher the attacker success rate, the more vulnerable the model is to this particular manipulation.

3. Results

Manipulating access to electricity at home. Among women with access to electricity ($x = 1$) and original prediction to have a facility delivery ($\hat{y} = 1$), the attacker success rate is 4.4% when changing to no access to electricity ($x = 1 \rightarrow 0$). When conditioning on women with no access to electricity ($x = 0$) and original prediction to have a facility delivery ($\hat{y} = 1$), the attacker success rate is 0% when changing to having access to electricity ($x = 0 \rightarrow 1$). Among women with access to electricity ($x = 1$) and original prediction to have a home-based delivery ($\hat{y} = 0$), the attacker success rate is 0% when changing to no access to electricity ($x = 1 \rightarrow 0$). When conditioning

Table 2
Adversarial attacks for access to electricity at home.

Condition	Number of observations	Attack	New prediction	Attacker success rate
$\hat{y} = 1, x = 1$	10,883	$x = 1 \rightarrow 0$	$y = 1$: 10,400 $y = 0$: 483	483/10,883 = 4.44%
$\hat{y} = 1, x = 0$	13,890	$x = 0 \rightarrow 1$	$y = 1$: 13,890 $y = 0$: 0	0%
$\hat{y} = 0, x = 1$	2213	$x = 1 \rightarrow 0$	$y = 1$: 0 $y = 0$: 2213	0%
$\hat{y} = 0, x = 0$	11,801	$x = 0 \rightarrow 1$	$y = 1$: 1300 $y = 0$: 10,501	1300/11,801 = 11.02%

on women with no access to electricity (x = 0) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rate is 11.0% when changing to having access to electricity (x = 0 → 1) (Table 2).

Manipulating previous delivery location. When we condition on women with previous home delivery (x = 1) and original prediction to have a facility delivery ($\widehat{y} = 1$), the attacker success rate is 0% when we manipulate previous delivery location into facility delivery (x = 1 → 3). When we condition on women with previous facility delivery (x = 3) and original prediction to have a facility delivery ($\widehat{y} = 1$), the attacker success rates are 55.7% when we manipulate previous delivery location into home delivery (x = 3 → 1). When we condition on women with previous home delivery (x = 1) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rate is 37.6% when we manipulate previous delivery location into facility delivery (x = 1 → 3). When we condition on women with previous facility delivery (x = 3) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rates are 0% when we manipulate previous delivery location into home delivery (x = 3 → 1) (Table 3).

Manipulating education level. Among women with more than some primary education (x ≠ 1) and original prediction to have a facility delivery ($\widehat{y} = 1$), the attacker success rate is 3.9% when changing education level to some primary education (x = 2,3,4,5 → 1). When conditioning on women who have not had higher education (x ≠ 5) and original prediction to have a facility delivery ($\widehat{y} = 1$), the attacker success rate is 0% when changing education level to higher education (x = 1,2,3,4 → 5). Among women with more than some primary education (x ≠ 1) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rate is 0% when changing education level to some primary education (x = 2,3,4,5 → 1). When conditioning on women who have not had higher education (x ≠ 5) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rate is 11.6% when changing education level to higher education (x = 1,2,3,4 → 5) (Table 4).

Manipulating gestational age at enrollment. Among women with larger gestational age at enrollment (x > 10) and original prediction to have a facility delivery ($\widehat{y} = 1$), the attacker success rate is 0% when applying a downward manipulation, no matter the magnitude of the manipulation. Among women with smaller gestational age at enrollment (x < 30) and original prediction to have facility delivery ($\widehat{y} = 1$), the attacker success rate is 2.0%, 3.5%, 4.3% when applying an upward manipulation of 5 weeks, 10 weeks, and replacing to 30 weeks, respectively. Among women with larger gestational age at enrollment (x > 10) and original prediction to have a home delivery ($\widehat{y} = 0$), the attacker success rate is 4.3%, 7.7%, 11.0% when applying a downward manipulation of 5 weeks, 10 weeks, and replacing to 10 weeks, respectively. Among women with smaller gestational age at enrollment (x < 30) and original prediction to have home delivery ($\widehat{y} = 0$), the attacker success rate is 0% when applying an upward manipulation, no matter the magnitude of the manipulation (Table 5).

4. Discussion

Our paper demonstrates the effects of adversarial attacks on four different variables when predicting facility-based delivery for women enrolled in a CHW-led maternal health program in Zanzibar. For the four exemplar variables, the variable manipulations that led to change in predicted values were intuitive; for example, access to electricity at home is a proxy for wealth, and wealth is associated with facility delivery. Therefore, changing from access to no access to electricity did change some women's prediction from facility-based delivery to home-based delivery. The fact that these effects are intuitive further emphasizes our concern that CHWs or

Table 3
Adversarial attacks for previous delivery location.

Condition	Number of observations	Attack	New prediction	Attacker success rate
$\widehat{y} = 1$, x = 1	577	x = 1 → 2	$y = 1$: 577 $y = 0$: 0	0%
		x = 1 → 3	$y = 1$: 577 $y = 0$: 0	0%
$\widehat{y} = 1$, x = 2	171	x = 2 → 1	$y = 1$: 53 $y = 0$: 118	118/171 = 69.01%
		x = 2 → 3	$y = 1$: 171 $y = 0$: 0	0%
$\widehat{y} = 1$, x = 3	16,593	x = 3 → 1	$y = 1$: 7359 $y = 0$: 9234	9234/16,593 = 55.65%
		x = 3 → 2	$y = 1$: 15,651 $y = 0$: 942	942/16,593 = 5.68%
$\widehat{y} = 0$, x = 1	7369	x = 1 → 2	$y = 1$: 2318 $y = 0$: 5051	2318/7369 = 31.46%
		x = 1 → 3	$y = 1$: 2773 $y = 0$: 4596	2773/7369 = 37.63%
$\widehat{y} = 0$, x = 2	167	x = 2 → 1	$y = 1$: 0 $y = 0$: 167	0%
		x = 2 → 3	$y = 1$: 18 $y = 0$: 149	18/167 = 10.78%
$\widehat{y} = 0$, x = 3	4994	x = 3 → 1	$y = 1$: 0 $y = 0$: 4994	0%
		x = 3 → 2	$y = 1$: 0 $y = 0$: 4994	0%

Table 4
Adversarial attacks for education level.

Condition	Number of observations	Attack	New prediction	Attacker success rate
$\hat{y} = 1$, x $\neq$ 1 (x = 2,3,4,5)	19,668	x = 2,3,4,5 → 1	$y = 1$: 18,895 $y = 0$: 773	773/19,668 = 3.93%
$\hat{y} = 1$, x $\neq$ 5 (x = 1,2,3,4)	24,427	x = 1,2,3,4 → 5	$y = 1$: 24,427 $y = 0$: 0	0%
$\hat{y} = 0$, x $\neq$ 1 (x = 2,3,4,5)	7197	x = 2,3,4,5 → 1	$y = 1$: 0 $y = 0$: 7197	0%
$\hat{y} = 0$, x $\neq$ 5 (x = 1,2,3,4)	13,942	x = 1,2,3,4 → 5	$y = 1$: 1616 $y = 0$: 12,326	1616/13,942 = 11.60%

Table 5
Adversarial attacks for gestational age at enrollment.

Condition	Number of observations	Attack	New prediction	Attacker success rate
$\hat{y} = 1$, x > 10	23,878	x = x → max(x-5, 10)	$y = 1$: 23,878 $y = 0$: 0	0%
		x = x → max(x-10, 10)	$y = 1$: 23,878 $y = 0$: 0	0%
		x = x → 10	$y = 1$: 23,878 $y = 0$: 0	0%
$\hat{y} = 1$, x < 30	21,564	x = x → min(x+5, 30)	$y = 1$: 21,125 $y = 0$: 439	439/21,564 = 2.04%
		x = x → min(x+10, 30)	$y = 1$: 20,809 $y = 0$: 755	755/21,564 = 3.50%
		x = x → 30	$y = 1$: 20,648 $y = 0$: 916	916/21,564 = 4.25%
$\hat{y} = 0$, x > 10	13,504	x = x → max(x-5, 10)	$y = 1$: 576 $y = 0$: 12,928	576/13,504 = 4.27%
		x = x → max(x-10, 10)	$y = 1$: 1049 $y = 0$: 12,455	1049/13,504 = 7.68%
		x = x → 10	$y = 1$: 1487 $y = 0$: 12,017	1487/13,504 = 11.01%
$\hat{y} = 0$, x < 30	12,285	x = x → min(x+5, 30)	$y = 1$: 0 $y = 0$: 12,285	0%
		x = x → min(x+10, 30)	$y = 1$: 0 $y = 0$: 12,285	0%
		x = x → 30	$y = 1$: 0 $y = 0$: 12,285	0%

women could identify variables to misreport in order to achieve a desired prediction.

Manipulating previous delivery location had the highest attacker success rate. This is not surprising given that previous delivery location is the most predictive variable in the algorithm. If prediction algorithms are deployed in practice, these variables that are highly vulnerable to adversarial attacks must be monitored to ensure the distributions of input data match expected distributions. Possible ways to secure input data quality include detecting and removing out-of-distribution samples using Gaussian Discriminant Analysis before carrying out the prediction or implementing manual assessments to ensure accurate reporting from the CHWs.

4.1. Limitations

Several limitations should be considered when interpreting these results. First, we imputed missing values with KNN imputation. Although KNN imputation is intuitive for continuous variables, we needed additional manipulations for categorical and ordinal variables. Different approaches for KNN with categorical and ordinal variables may yield different predictive models and different vulnerabilities to adversarial attacks. Second, we focused our evaluation on one-at-a-time adversarial attacks, which gives us the attacker success rate when changing only a single variable while holding all other variables constant. This simplified setup limits our ability to model the more complicated real-world scenario, where individuals may manipulate multiple responses simultaneously [51]. In future work, we will consider a more generalizable setting by taking the dynamic interactions between variables into account. Third, this paper specifically looked at adversarial attacks on a binary classification algorithm in the context of LASSO regularized logistic regression model because this is the current algorithm implemented by our collaborators in Zanzibar. In future research, we will examine how different algorithms could potentially reduce the vulnerability to adversarial attacks and expand our research to understand the impact of adversarial attacks on a multiple-level classification algorithm. Finally, this paper focuses narrowly on how the prediction results would change under adversarial attacks once this prediction model is built; future research should consider the potential of model poisoning, whereby data is manipulated during the model building stages, and also strategies to monitor for and prevent adversarial attacks during prediction.

4.2. Conclusion

This paper investigates the vulnerability of an algorithm to predict facility-based delivery when facing adversarial attacks. We found that manipulating input variables indeed has a large impact on the prediction results. By understanding the effect of adversarial attacks and the most vulnerable variables, programs can implement data monitoring strategies to deter these manipulations. Ensuring fidelity in algorithm deployment secures that CHWs target those women who are actually at high risk of delivering at home.

Author contribution statement

Yi-Ting Tsai: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Isabel R. Fulcher, Bethany Hedt-Gauthier: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Tracey Li, Felix Sukums: Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data will be made available on request.

Additional information

No additional information is available for this paper.

Funding statement

Yi-Ting Tsai was supported by the Harvard Data Science Initiative Public Service Data Science Graduate Fellowship.

Isabel R. Fulcher was supported by the Harvard Data Science Initiative Postdoctoral Fellows Program.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We would like to thank the Zanzibar Ministry of Health for their collaboration in the *Uzazi Salama* and *Jamii ni Afya* programs.

Summary table

What was already known on the topic.

- The increasing adoption of mobile devices in community health worker-led maternal health programs provides an opportunity for real-time implementation of machine learning predictive models to identify those at risk for poor outcomes thus to devise additional interventions.
- Research groups have developed prediction models to identify women at risk for home-based delivery in sub-Saharan Africa.

What this study added to our knowledge.

- This paper is the first to assess the vulnerability of algorithms to predict home-based versus facility-based delivery to adversarial attacks.
- Manipulation of all variables resulted in changes in the model's predicted results; the variable previous delivery location had the highest attacker success rate among all the examined variables.
- When these models are implemented in real-world programs, variables that are highly vulnerable to adversarial attacks must be monitored for accuracy to ensure accurate predicting.

References

[1] T.N. Thomas, J. Gausman, S.R. Lattof, M.N. Wegner, A.D. Kearns, A. Langer, Improved maternal health since the ICPD: 20 years of progress, Contraception 90 (6) (2014) S32–S38.

[2] E. Kalipeni, J. Iwelunmor, D. Grigsby-Toussaint, Maternal and child health in Africa for sustainable development goals beyond 2015, Global Publ. Health 12 (6) (2017) 643–647.
[3] O.M. Campbell, W.J. Graham, Lancet Maternal Survival Series steering group, Strategies for reducing maternal mortality: getting on with what works, Lancet 368 (9543) (2006) 1284–1299.
[4] C.A. Moyer, P. Dako-Gyeke, R.M. Adanu, Facility-based delivery and maternal and early neonatal mortality in sub-Saharan Africa: a regional review of the literature, Afr. J. Reprod. Health 17 (3) (2013) 30–43.
[5] J. Chinkhumba, M. De Allegri, A.S. Muula, B. Robberstad, Maternal and perinatal mortality by place of delivery in sub-Saharan Africa: a meta-analysis of population-based cohort studies, BMC Publ. Health 14 (2014) 1014.
[6] K.S. Adde, K.S. Dickson, H. Amu, Prevalence and determinants of the place of delivery among reproductive age women in sub–Saharan Africa, PLoS One 15 (12) (2020), e0244875.
[7] H.V. Doctor, S. Nkhana-Salimu, M. Abdulsalam-Anibilowo, Health facility delivery in sub-Saharan Africa: successes, challenges, and implications for the 2030 development agenda, BMC Publ. Health 18 (1) (2018) 765.
[8] P.C. Rockers, M.L. Wilson, G. Mbaruku, M.E. Kruk, Source of antenatal care influences facility delivery in rural Tanzania: a population-based study, Matern. Child Health J. 13 (6) (2009) 879–885.
[9] A. Anyait, D. Mukanga, G.B. Oundo, F. Nuwaha, Predictors for health facility delivery in Busia district of Uganda: a cross sectional study, BMC Pregnancy Childbirth 12 (1) (2012) 1–9.
[10] M. Boah, A.B. Mahama, E.A. Ayamga, They receive antenatal care in health facilities, yet do not deliver there: predictors of health facility delivery by women in rural Ghana, BMC Pregnancy Childbirth 18 (1) (2018) 125.
[11] C.A. Moyer, A. Mustafa, Drivers and deterrents of facility delivery in sub-Saharan Africa: a systematic review, Reprod. Health 10 (2013) 40.
[12] T.M. Huda, M. Chowdhury, S. El Arifeen, M.J. Dibley, Individual and community level factors associated with health facility delivery: a cross sectional multilevel analysis in Bangladesh, PLoS One 14 (2) (2019), e0211113.
[13] E. Gitonga, F. Muiruri, Determinants of health facility delivery among women in Tharaka Nithi county, Kenya, Pan Afr. Med. J. 25 (2) (2016) 9.
[14] T.W. Kohi, L.T. Mselle, J. Dol, M.L. Aston, When, where and who? Accessing health facility delivery care from the perspective of women and men in Tanzania: a qualitative study, BMC Health Serv. Res. 18 (2018) 564.
[15] M.A. Bohren, E.C. Hunter, H.M. Munthe-Kaas, J.P. Souza, J.P. Vogel, A.M. Gülmezoglu, Facilitators and barriers to facility-based delivery in low-and middle-income countries: a qualitative evidence synthesis, Reprod. Health 11 (1) (2014) 71.
[16] M.E. Kruk, S. Kujawski, G. Mbaruku, K. Ramsey, W. Moyo, L.P. Freedman, Disrespectful and abusive treatment during facility delivery in Tanzania: a facility and community survey, Health Pol. Plann. 33 (1) (2018) 26–33.
[17] M. Koblinsky, I. Anwar, M.K. Mridha, M.E. Chowdhury, R. Botlero, Reducing maternal mortality and improving maternal health: Bangladesh and MDG 5, J. Health Popul. Nutr. 26 (2008) 280–294.
[18] M. Kyei-Nimakoh, M. Carolan-Olah, T.V. Mccann, Access barriers to obstetric care at health facilities in sub-Saharan Africa—a systematic review, Syst. Rev. 6 (2017) 110.
[19] V. Mochache, A. Lakhani, H. El-Busaidy, M. Temmerman, P. Gichangi, Correlates of facility-based delivery among women of reproductive age from the Digo community residing in Kwale, Kenya, BMC Res. Notes 11 (1) (2018) 715.
[20] F. Moshi, T. Nyamhanga, Understanding the preference for homebirth; an exploration of key barriers to facility delivery in rural Tanzania, Reprod. Health 14 (2017) 132.
[21] L.E. Cofie, C. Barrington, K. Singh, S. Sodzi-Tettey, S. Ennett, S. Maman, Structural and functional network characteristics and facility delivery among women in rural Ghana, BMC Pregnancy Childbirth 17 (1) (2017) 425.
[22] S.M. Haddad, R.T. Souza, J.G. Cecatti, Mobile technology in health (mHealth) and antenatal care-Searching for apps and available solutions: a systematic review, Int. J. Med. Inf. 127 (2019) 1–8.
[23] A. Feroz, S. Perveen, W. Aftab, Role of mHealth applications for improving antenatal and postnatal care in low and middle-income countries: a systematic review, BMC Health Serv. Res. 17 (1) (2017) 704.
[24] T. Tamrat, S. Kachnowski, Special delivery: an analysis of mHealth in maternal and newborn health programs and their outcomes around the world, Matern. Child Health J. 16 (5) (2012) 1092–1101.
[25] J. Early, C. Gonzalez, V. Gordon-Dseagu, L. Robles-Calderon, Use of mobile health (mHealth) technologies and interventions among community health workers globally: a scoping review, Health Promot. Pract. 20 (6) (2019) 805–817.
[26] C. Hategeka, H. Ruton, M.R. Law, Effect of a community health worker mHealth monitoring system on uptake of maternal and newborn health services in Rwanda, Glob. Health Res. Policy 4 (1) (2019).
[27] S. Agarwal, C. Glenton, T. Tamrat, N. Henschke, N. Maayan, M.S. Fønhus, G.L. Mehl, S. Lewin, Decision-support tools via mobile devices to improve quality of care in primary healthcare settings, Cochrane Database Syst. Rev. 7 (7) (2021).
[28] I.B. Mboya, M.J. Mahande, M. Mohammed, J. Obure, H.G. Mwambi, Prediction of perinatal death using machine learning models: a birth registry-based cohort study in northern Tanzania, BMJ Open 10 (10) (2020).
[29] B. Tesfaye, S. Atique, T. Azim, M.M. Kebede, Predicting skilled delivery service use in Ethiopia: dual application of logistic regression and machine learning algorithms, BMC Med. Inf. Decis. Making 19 (1) (2019) 209.
[30] A. Kwizera, N. Kissoon, N. Musa, O. Urayeneza, P. Mujyarugamba, A.J. Patterson, L. Harmon, J.C. Farmer, M.W. Dünser, J. Meier, Sepsis in resource-limited nations" task force of the surviving sepsis campaign, in: A Machine Learning-Based Triage Tool for Children with Acute Infection in a Low Resource Setting. Pediatric Critical Care Medicine, vol. 20, 2019, 12.
[31] K.J. Rittenhouse, B. Vwalika, A. Keil, J. Winston, M. Stoner, J.T. Price, M. Kapasa, M. Mubambe, V. Banda, W. Muunga, J.S.A. Stringer, Improving preterm newborn identification in low-resource settings with machine learning, PLoS One 14 (2) (2019).
[32] T. Tuti, A. Agweyu, P. Mwaniki, N. Peek, M. English, Clinical Information Network Author Group, An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from Kenya, BMC Med. 15 (1) (2017) 201.
[33] S. Sazawal, K.K. Ryckman, S. Das, R. Khanam, I. Nisar, E. Jasper, A. Dutta, et al., Machine learning guided postnatal gestational age assessment using new-born screening metabolomic data in South Asia and sub-Saharan Africa, BMC Pregnancy Childbirth 21 (1) (2021) 609.
[34] W. Ogallo, S. Speakman, V. Akinwande, K.R. Varshney, A. Walcott-Bryant, C. Wayua, et al., Identifying factors associated with neonatal mortality in sub-saharan Africa using machine learning, AMIA Annu. Symp. Proc. 2020 (2021) 963–972.
[35] V.V. Shukla, B. Eggleston, N. Ambalavanan, E.M. McClure, M. Mwenechanya, E. Chomba, C. Bose, et al., Predictive modeling for perinatal mortality in resource-limited settings, JAMA New Open 3 (11) (2020).
[36] I.R. Fulcher, A.R. Nelson, J.I. Tibaijuka, S.S. Seif, S. Lilienfeld, O.A. Abdalla, et al., Improving health facility delivery rates in Zanzibar, Tanzania through a large-scale digital community health volunteer programme: a process evaluation, Health Pol. Plann. 35 (10) (2020) 1–11.
[37] A. Fredriksson, I.R. Fulcher, A. Nelson, T. Li, Y.T. Tsai, S.S. Seif, R.N. Mpembeni, B. Hedt-Gauthier, Machine learning for maternal health: predicting delivery location in a community health worker program in Zanzibar, Front. Digit. Health 4 (2022), 855236.
[38] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, I.S. Kohane, Adversarial attacks on medical machine learning, Science 363 (6433) (2019) 1287–1289.
[39] S. Cresci, M. Petrocchi, A. Spognardi, S. Tognazzi, Adversarial machine learning for protecting against online manipulation, arXiv: 2111.12034, arXiv preprint (2021).
[40] A.K.M.I. Newaz, N.I. Haque, A.K. Sikder, M.A. Rahman, A.S. Uluagac, Adversarial attacks to machine learning-based smart healthcare systems, arXiv: 2010.03671, arXiv preprint (2020).
[41] S. Qiu, Q. Liu, S. Zhou, C. Wu, Review of artificial intelligence adversarial attack and defense technologies, Appl. Sci. 9 (5) (2019) 909.

[42] Z. Abaid, M.A. Kaafar, S. Jha, Quantifying the impact of adversarial evasion attacks on machine learning based android malware classifiers, in: IEEE 16th international symposium on network computing and applications (NCA), 2017, pp. 1–10.
[43] N. Papernot, P. McDaniel, I. Goodfellow, Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, arXiv: 1605.07277, arXiv preprint (2016).
[44] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Med. Inf. Decis. Making 16 (3) (2016) 197–208.
[45] A. Jadhav, D. Pramod, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset, Appl. Artif. Intell. 33 (10) (2019) 913–933.
[46] N.Z. Abidin, A.R. Ismail, N.A. Emran, Performance analysis of machine learning algorithms for missing value imputation, Int. J. Adv. Comput. Sci. Appl. 9 (6) (2018).
[47] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B 58 (1) (1996) 267–288.
[48] D.M. Hamby, A review of techniques for parameter sensitivity analysis of environmental models, Environ. Monit. Assess. 32 (1994) 135–154.
[49] D.G. Cacuci, M. Ionescu-Bujor, A comparative review of sensitivity and uncertainty analysis of large-scale systems—II: statistical methods, Nucl. Sci. Eng. 147 (3) (2004) 204–217.
[50] I.R. Fulcher, K. Hedt, S. Marealle, J. Tibaijuka, O. Abdalla, R. Hofmann, E. Layer, M. Mitchell, B. Hedt-Gauthier, Errors in estimated gestational ages reduce the likelihood of health facility deliveries: results from an observational cohort study in Zanzibar, BMC Health Serv. Res. 20 (1) (2020) 50.
[51] A. Saltelli, P. Annoni, How to avoid a perfunctory sensitivity analysis, Environ. Model. Software 25 (12) (2010) 1508–1517.