frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 24 September 2025
DOI 10.3389/frai.2025.1484299



OPEN ACCESS

EDITED BY
Tim Hulsen,
Rotterdam University of Applied Sciences,
Netherlands

REVIEWED BY
Elif Keles,
Northwestern Medicine, United States
Murali Dharan Rajagopalan,
BGS Global Hospital, India

*CORRESPONDENCE
Wudneh Ketema Moges
✉ wudnehketema@gmail.com

RECEIVED 22 August 2024
ACCEPTED 05 September 2025
PUBLISHED 24 September 2025

CITATION
Moges WK, Tegegne AS, Mitku AA, Tesfahun E and Hailemeskel S (2025) Deep learning for causal inference using low birth weight in midwife-led continuity care intervention in north Shoa zone, Ethiopia. *Front. Artif. Intell.* 8:1484299.
doi: 10.3389/frai.2025.1484299



# Deep learning for causal inference using low birth weight in midwife-led continuity care intervention in north Shoa zone, Ethiopia

Wudneh Ketema Moges[1,2,3]*, Awoke Seyoum Tegegne[1], Aweke A. Mitku[1,4], Esubalew Tesfahun[5] and Solomon Hailemeskel[6]

[1]Department of Statistics, College of Science, Bahir Dar University, Bahir Dar, Ethiopia, [2]Department of Statistics, College of Science, Debre Berhan University, Debre Berhan, Ethiopia, [3]Department of Data Science, College of Computing, Debre Berhan University, Debre Berhan, Ethiopia, [4]Global Change Institute (GCI), Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa, [5]Department of Public Health, College of Health Science, Debre Berhan University, Debre Berhan, Ethiopia, [6]Department of Midwifery, College of Health Science, Debre Berhan University, Debre Berhan, Ethiopia

**Introduction:** Low birth weight (LBW), under 2,500 g, poses health risks, though not always requiring treatment. Early detection of high-risk pregnancies enables preventive care, improving outcomes for mother and baby. This study aimed to establish cause-and-effect relationships using Causal Deep Learning (CDL) models that reduce bias and estimate heterogeneous treatment effects on LBW in the Midwife-Led Continuity Care (MLCC) intervention.

**Methods:** This study used a quasi-experimental study design (August 2019–September 2020) in North Shoa, Ethiopia, and enrolled 1,166 women divided into two groups: one receiving MLCC and the other receiving other professional groups for comprehensive antenatal/postnatal care. The dataset and code are provided in data availability section. Our model combines counterfactual convolutional neural networks to analyze time-based patterns and Bayesian Ridge regression to reduce bias in propensity scores. We use Counterfactual Regression with Wasserstein Distance (CFR-WASS) and Counterfactual Regression with Maximum Mean Discrepancy (CFR-MMD) to balance patient characteristics and improve counterfactual estimates of treatment effects. This approach strengthens causal insights into how MLCC interventions affect LBW outcomes.

**Result:** The Deep neural networks (DNN) model showed strong predictive accuracy for LBW, with 81.3% training and 81.4% testing performance, an area under the curve (AUC) of 0.88, enabling the reliable early identification of high-risk pregnancies. The study found a strong link between meconium aspiration syndrome (MAS) and LBW ($p$ = 0.002), but this does not mean MAS directly causes LBW. MAS likely results from fetal distress or other pregnancy complications that may independently affect LBW. While statistical associations exist, clinical causation remains unproven; therefore, the counterfactual analysis showed MLCC could help reduce LBW risk. CFR-WASS achieved high accuracy (84%) while the precision in heterogeneous treatment effect (PEHE = 1.006) and the average treatment effect (ATE = 0.24), and CFR-MMD PEHE of 1.02, ATE of 0.45, demonstrating potential for tailored treatment strategies. DNN and multilayer

perceptrons uniquely identified key neural weights and biases favoring normal birth weight while suppressing LBW predictions, offering interpretable insights for clinical risk assessment.

**Conclusion:** The CFR-WASS/CFR-MMD model strengthens LBW prediction by identifying crucial factors like MAS and healthcare access, while accurate PEHE and ATE estimates support data-driven prenatal care and targeted interventions for healthier outcomes.

KEYWORDS

**causal deep learning, low birth weight, precision in estimating heterogeneous treatment effects, average treatment effect, midwife-led continuity care**

## Introduction

Low birth weight (LBW), defined as a birth weight below 2,500 grams, is a major global public health issue linked to higher neonatal mortality and long-term adverse health outcomes (Desta, 2019; Endalamaw et al., 2018). Despite existing interventions aimed at reducing LBW, there remains a critical gap in understanding how these strategies differentially affect subpopulations of pregnant women, particularly in low-resource settings (Park et al., 2020; Koivu et al., 2023).

Antenatal care (ANC) has been widely recognized as a key factor in reducing LBW risk. Studies consistently show that inadequate ANC, often defined as fewer than four visits during pregnancy, is strongly associated with higher LBW incidence. For example, (Roslina, 2020) found that mothers receiving fewer than four ANC visits had a 1.9 times greater risk of delivering LBW infants compared to those with adequate care. Similarly, (Sunarni et al., 2018) demonstrated that regular ANC attendance significantly improves birth weight outcomes, reinforcing the need for targeted maternal healthcare policies.

This gap relates to the concept of heterogeneous treatment effects. Heterogeneous treatment effects occur when the effectiveness of an intervention varies across different groups of individuals (Kent et al., 2018). In the context of LBW prevention, reduction-specific Midwife-Led Continuity Care (MLCC) is highly effective for one sub-group of pregnant women. It has a minimal or negative effect on another subgroup (Sandall et al., 2024). Unfortunately, most studies on LBW prevention have focused on average treatment effects (ATE) and neglected the potential for heterogeneous treatment effects. This limitation hinders the development of precisely targeted interventions to conduct the current study, which contributes to filling a critical knowledge or awareness gap regarding the adverse treatment effect.

In healthcare research, observational studies estimate causal effects, specifically the impact of MLCCs or interventions on medical outcomes. However, quasi-experimental studies are the gold standard for establishing cause-and-effect relationships, and practical constraints such as ethical, logistical, or financial limitations may make them infeasible. In such situations, researchers rely on quasi-experimental studies to explore causal effects. These studies address potential bias by considering confounders, although both approaches rely on untestable assumptions (Zawadzki et al., 2023). Deep learning (DL) is a subset of machine learning (ML) that uses deep neural networks (DNNs) to recognize patterns in large, complex datasets, achieving state-of-the-art results in fields such as computer vision and health sciences (Keles and Bagci, 2023). In neonatology, DL has proven revolutionary, particularly in tasks such as survival analysis, neuroimaging, and diagnosis of conditions such as retinopathy of prematurity (Keles and Bagci, 2023). Furthermore, DL models are complex networks that learn independently without human intervention. These models have multiple layers, enabling them to process information without explicit human guidance (Taye, 2023).

Deep learning is preferred over the existing methods for causal inference in LBW due to its ability to model complex and non-linear relationships. DL models, such as neural networks, can capture intricate interactions and confounding factors, providing more accurate causal effect estimates (Fan et al., 2023). Frameworks such as Counterfactual Regression with Wasserstein Distance (CFR-WASS) and Counterfactual Regression with Maximum Mean Discrepancy (CFR-MMD) integrate causal inference techniques, enhancing robustness and generalizability (Meinshausen, 2018; Shalit et al., 2017). While the existing methods are somehow more interpretable, DL's superior predictive performance and flexibility make it a powerful tool for identifying causal factors in LBW studies, going beyond prediction to uncover actionable insights (Uauy et al., 2013).

Recent research highlights the integration of causal inference with DL to enhance model robustness, interpretability, and generalizability. This approach addresses limitations in existing DL, which may capture spurious correlations and lack interpretability (Jiao et al., 2024). By integrating causal effects into DL models, analysts have enhanced predictive performance and the interpretability of results in complex domains such as electronic health records (Ghosh et al., 2018). The combination of ML and causal inference methods has also shown promise in system dynamics modeling, enabling better forecasting and understanding of complex interactions across various disciplines (Koch et al., 2025). To facilitate adoption, researchers have developed frameworks and tutorials for implementing DL-based causal inference methods, focusing on observational causal estimation and extending causal inference to settings with non-linear confounding and diverse data types (Koch et al., 2025). This emerging field offers significant potential for advancing our understanding of cause-and-effect relationships in complex systems. Therefore, DL models augmented with causal inference techniques are better equipped to address

Abbreviations: ATE, Average treatment effect; CATE, Conditional average treatment; CCNN, Counterfactual convolutional neural network; CDL, Causal deep learning; CFR-MMD, Counterfactual regression with maximum mean discrepancy; CFR-WASS, Counterfactual Regression with Wasserstein distance; DNN, Deep neural network; FFDNN, Feed-forward deep neural network; ITE, Individual treatment effect; LBW, Low birth weight; MAS, Meconium aspiration syndrome; MLCC, Midwife-led continuity care; NBW, Normal birth weight; PEHE, Precision of estimation heterogeneous effect; PIH, Pregnancy-induced hypertension.

confounding variables and biases, resulting in more precise and dependable predictions. This approach helps to detect targeted interventions to prevent LBW, leading to better health outcomes for mothers and babies, and it also helps to connect causal inference and DL within quasi-experimental settings (Zhang et al., 2022; Samek et al., 2021). Causal Deep Learning (CDL) leverages partial causal knowledge among some and not necessarily all variables of interest and quantitatively characterizes the functional form among variables of interest and decision-makers (Berrevoets et al., 2024). Despite strong evidence supporting MLCC, diverse stakeholder interests and power dynamics hinder its implementation (Simmelink et al., 2025). This study was to address these challenges by leveraging CDL to evaluate MLCC's impact on LBW outcomes.

## Related work

Some existing studies are as follows: We used an ML approach to predict the weight range of infants in studies conducted in Belihuloya, Balangoda, Sri Lanka. The study was conducted in the United Arab Emirates (Khan et al., 2022) to estimate infant birth weight and LBW using ML algorithms. The authors conducted their study in Shanghai, China, using an ML approach to estimate fetal birth weight in high-risk pregnancies (Moreira et al., 2019).

Studies in the USA (Lu et al., 2019) investigated fetal weight at varying gestational ages using an ML approach. Studies conducted in Mexico (Campos Trujillo et al., 2020) predict early fetal weight using a support vector machine (SVM). Furthermore, the study conducted in China (Tao et al., 2021) used hybrid data from electronic medical records with the B-ultrasonic examinations of pregnant women to build a predicted birth weight classifier based on extended short-term memory networks.

Previous studies have employed ML and DL algorithms to predict LBW but have not sufficiently addressed causal inference. For instance, research in Iran (Arayeshgari et al., 2023) compared multiple ML models, including decision trees, random forests, artificial neural networks (ANNs), SVM, and logistic regression, to predict LBW risk factors. Similarly, another study in Iran evaluated eight ML and DL algorithms (XGBoost, LightGBM, and K-nearest neighbors) for LBW prediction but focused solely on predictive accuracy rather than causal effects (Alam et al., 2023). In the United States, ML approaches have been applied to LBW prediction. Yet, these studies also lacked rigorous causal analysis, such as estimating ATE or addressing the precision in heterogeneous treatment effect (PEHE) (Jiao et al., 2024).

A critical gap in these studies is their reliance on associative models rather than causal frameworks, which limits their utility for policy and intervention design. Recent advancements in CDL and doubly robust methods (Bayesian additive regression trees for propensity score estimation) have improved bias reduction and PEHE estimation in observational health data (Jiao et al., 2024; Mbogu, 2023). While ML and DL have been applied to neonatal outcomes, most studies focus on associative predictions rather than causal inference. For instance, (Keles and Bagci, 2023) systematically reviewed 106 AI studies in neonatology and found that primary applications included survival analysis and diagnosis, but few addressed causal relationships or heterogeneous treatment effects. This gap highlights the need for methods such as CDL to move beyond correlation and quantify intervention impacts, such as MLCC on LBW. Therefore, this study aimed to establish cause-and-effect relationships using CDL models that reduce bias and estimate heterogeneous treatment effects on LBW in the MLCC intervention.

The novelty of this work lies in its innovative application of CDL models, specifically CFR-WASS and CFR-MMD, to estimate heterogeneous treatment effects of MLCC on LBW outcomes in a low-resource setting. Unlike previous studies focusing solely on predictive accuracy, this research integrates counterfactual analysis with DL to reduce bias and provide interpretable causal insights. The study uniquely combines causal inference with DNNs to improve propensity score estimation, enabling tailored intervention strategies. Additionally, it offers robust metrics such as PEHE, ATE, and individualized treatment effect (ITE), advancing precision in maternal healthcare decision-making.

Propensity Score Matching (PSM) offers valuable insights for improving maternal healthcare by creating comparable groups from observational data. This approach helps clinicians determine which care strategies prove most effective for particular patient populations by matching women with similar medical histories and risk factors. For health policymakers, PSM serves as a powerful tool to assess real-world program impacts, such as evaluating community health worker initiatives by comparing health outcomes between equivalent groups who did and did not receive the MLCC. The method generates practical evidence to expand successful programs and modify less effective ones, particularly important in settings with limited healthcare resources where data-driven decisions. While PSM provides crucial evidence when clinical trials aren't feasible, our advanced CDL methods overcome PSM's constraints by modeling intricate relationships in maternal health data, leading to more nuanced policy recommendations and clinical guidelines. Our CDL framework builds on PSM by addressing its limitations, capturing complex relationships to further improve precision in maternal health strategies (Yu and Kang, 2019).

# Methods and participants

## Data sources

A quasi-experimental study was conducted between August 2019 and September 2020 in the North Shoa Zone in the Amhara Regional State of Ethiopia. This region is home to over two million people, with approximately 2,393,877 individuals residing within its boundaries. Among these residents, 1,207,839 are males, and 1,186,038 are females.

There are a total of nine hospitals in the region. One hospital is a referral center specifically equipped to provide comprehensive emergency obstetric care. The North Shoa Zone boasts 95 health centers. These centers serve as essential points of access to primary healthcare services. In addition to hospitals and health centers, the region has 389 health posts. These health posts are strategically distributed across rural and urban areas, ensuring that even remote communities can access basic healthcare services. Overall, the North Shoa Zone's healthcare infrastructure strives to address its population's diverse healthcare needs, emphasizing maternal and child health, emergency care, and community-based services.

A total of 1,166 mothers visiting prenatal and antenatal care clinics during the study period were included. Four primary hospitals in the study area, Shoa Robit, Ataye, Mehal Meda, and

Alem Ketema Enat Hospital, were randomly selected using a two-stage stratified cluster sampling technique. These hospitals serve both urban and rural populations and provide delivery services. Samples were equally distributed, and participants were selected using systematic random sampling with an interval of two. Shoa Robit and Ataye hospitals were designated as intervention sites offering MLCC, while Mehal Meda and Alem Ketema Enat hospitals served as control sites. Eligible pregnant women were approached and enrolled until the target sample size was achieved.

## Data collection

Midwives recorded participants' baseline characteristics, including socio-demographics and obstetric, gynecologic, medical, and surgical histories, using a standard tool via face-to-face interviews and maternal antenatal cards. An independent, blinded data collector from the birth registry collected post-birth outcomes. Intervention exposure and continuity of care data were obtained from medical records and postnatal interviews. To avoid the Hawthorne effect, healthcare providers were blinded to outcome data. Eight midwife data collectors and four supervisors underwent a three-day training program for data collection and extraction.

## Eligibility criteria

The study included pregnant women who were less than 24 weeks' gestational age at their first antenatal care visit, had a singleton pregnancy, and were classified as low obstetric risk. Women with multiple pregnancies, those planning to seek care from a different provider, or those with a history of medical or obstetric complications were excluded from the study.

## Quasi-experimental setup

### Treatment group (MLCC)

Antenatal care plays a pivotal role in preventing LBW, with the effectiveness often influenced by the model of care provided. MLCC is increasingly recognized for its positive impact on birth outcomes, including a reduced risk of LBW. This model fosters a strong, trusting relationship between a woman and her consistent midwife or small team of midwives throughout pregnancy, birth, and the postnatal period (Moges et al., 2025). This continuity facilitates early and comprehensive risk identification, allowing for prompt interventions such as nutritional counseling, vigilant monitoring for conditions such as pre-eclampsia, and timely referrals for complications, all of which directly mitigate LBW risk (Sandall et al., 2024). Furthermore, MLCC promotes health education and supports physiological pregnancy and birth processes. By avoiding potentially harmful interventions such as routine episiotomies, elective labor inductions without medical indication, or unnecessary cesarean sections, MLCC helps maintain optimal conditions for fetal growth (Mayberry et al., 2017). This approach reduces interruptions to a natural pregnancy, which can negatively affect birth weight. Studies indicate that MLCC protects against preterm birth and LBW, especially for at-risk women, by improving their access to and engagement with community-based care.

### Control (other professional)

In contrast, antenatal care models led by other professional groups, such as obstetricians or those involving fragmented standard care, while essential for high-risk pregnancies, may sometimes face challenges in optimizing factors related to LBW prevention for all women. Obstetrician-led care is critical for managing complex medical conditions and severe complications that directly threaten fetal growth and contribute to LBW. However, in lower-risk pregnancies, a more medicalized approach might lead to higher intervention rates without a corresponding benefit in LBW reduction compared to MLCC (Voon et al., 2017).

Continuous care is crucial for preventing LBW. When pregnant women see different providers, it breaks down trust and causes inconsistent health advice, making it harder to spot complications early and leading to poor adherence to medical advice. This lack of consistent care can increase the risk of LBW, so clinical practices should focus on building strong, ongoing relationships between women and their healthcare providers (Fernandez Turienzo et al., 2021).

### Variables in the study

This study analyzed binary outcome variables, categorizing newborns into two groups: LBW (≤2,499 grams) and normal birth weight (NBW) (≥2,500 grams). This research framework focused on the causal relationship between MLCC and other professional groups while controlling demographic characteristics, obstetric history, medical factors, and neonatal outcomes influenced LBW risk across these care models (Reza and Salma, 2024), as shown in Figure 1.

## Variable selection

In this study, we initially examined a comprehensive set of over 200 variables encompassing demographic, obstetric, medical, and neonatal factors to investigate their association with LBW, as shown in Figure 1 below. We employed CML techniques combined with feature selection methods to identify the most predictive features while minimizing redundancy and multicollinearity (Moges et al., 2025).

First, we applied univariate analysis to assess the preliminary relevance of each variable. Next, we utilized regularized regression to penalize non-influential predictors, shrinking their coefficients to zero and retaining only the most significant ones. We implemented recursive feature elimination to further refine our selection, which iteratively removes the least important features while optimizing model performance (Moges et al., 2025).

## Causal deep learning algorithms

This study employed TensorFlow 2 and PyTorch to develop CDL models for estimating heterogeneous treatment effects across subgroups. Custom neural networks were designed for specific causal inference tasks, while Scikit-learn handled existing ML. Data processing relied on NumPy and Pandas, with visualizations generated

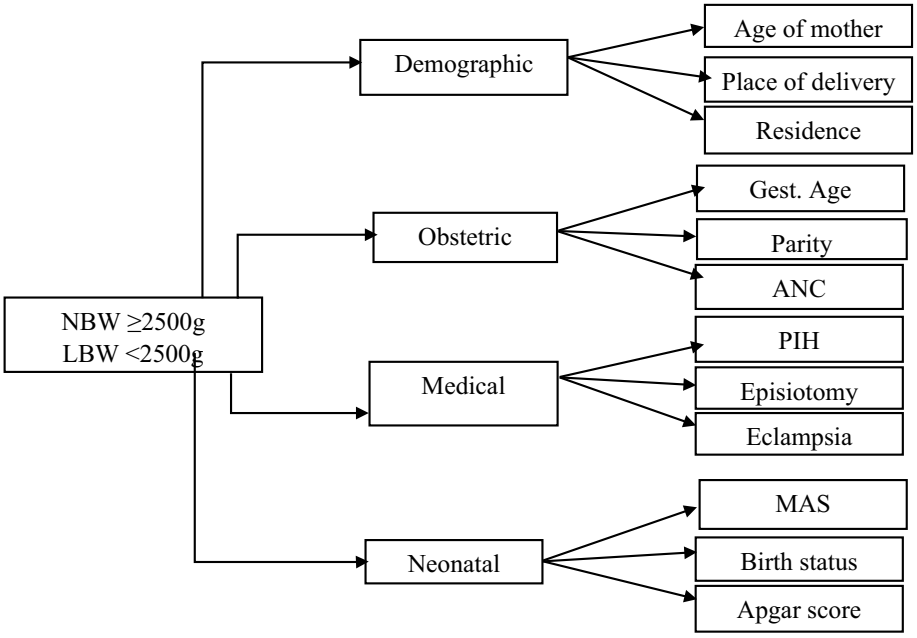


FIGURE 1
Key factors associated with low birth weight. ANC, antenatal care visit; PIH, pregnancy-induced hypertension; MAS, meconium aspiration syndrome; Gest.Age, gestational age.

using Matplotlib and Seaborn. Bootstrapping ensured reproducibility, and causal analysis was supported by PyWhy and CausalML.

These models adapt well to non-linear confounding and dynamic factors, addressing the limitations of traditional linear approaches (Pattanayak et al., 2017). The study investigates the use of CDL methods, including DNNs, Counterfactual Convolutional Neural Network (CCNNs), Bayesian Ridge, Bagging Regressor, Treatment Agnostic Representation Network (TARNet), Balancing Neural Networks, CFR-WASS, Causal Effect Variational Autoencoder, and CFR-MMD, to estimate propensity scores in causal inference (Whata and Chimedza, 2022; Ramachandra, 1803).

These advanced CDL models significantly improve treatment effect estimation from real-world clinical data. DNNs and CCNNs excel at analyzing complex patient characteristics for precise treatment comparisons. Bayesian methods incorporate medical expertise into analyses, while ensemble approaches such as Bagging Regressor produce stable results. TARNet and balancing-focused networks (CFR-WASS/MMD) rigorously control for confounding factors in quasi-experimental studies. Most innovatively, the Causal Effect Variational Autoencoder predicts how individual patients would respond to different treatments, enabling truly personalized care recommendations. Together, these methods provide clinicians with more reliable evidence about treatment effectiveness while accounting for real-world data limitations (Keles and Bagci, 2023).

## Counterfactual convolutional neural networks

The PEHE and ATE for CCNNs depend on factors like architecture, training data, and hyperparameters (Kong et al., 2022). While CCNNs excel in image-based tasks, adapting them for causal inference requires careful design. To estimate PEHE and ATE using CCNNs, we have created an architecture that takes covariates and MLCC as input, predicting LBW; this is well-suited for handling sequential data like pregnancy stages and their influence on LBW.

## Bayesian ridge

We have trained our model on the data to estimate PEHE and ATE using Bayesian Ridge, incorporating covariates and MLCC as features. PEHE compares predicted outcomes for MLCC and other professional groups. At the same time, ATE is computed based on average outcomes for MLCC and other professional groups. Therefore, the mathematical expression for Bayesian ridge regression can be represented as follows:

$$y = Xw + \in$$

Where: $y$ represents the target variable, $X$ is the design matrix (features), $w$ represents the weight vector (coefficients), and $\in$ represents the noise.

## Counterfactual regression with Wasserstein distance

The CFR-WASS improves causal effect estimation by balancing covariate distributions between treatment groups, addressing selection bias, and non-overlapping support (Shalit et al., 2017). Unlike conventional methods like PSM, it minimizes distributional discrepancies using the Wasserstein distance, enhancing accuracy in heterogeneous treatment effects. CFR-WASS excels in

quasi-experimental data, particularly for complex relationships and confounding variables, providing reliable Conditional Average Treatment Effect (CATE) estimates. Its robustness makes it especially valuable for evaluating interventions like MLCC compared to standard care models.

## Counterfactual regression with maximum mean discrepancy

Counterfactual Regression (CFR) estimates ITE by predicting potential outcomes under different treatments, such as MLCC. While useful for quasi-experimental data, CFR faces selection bias when comparing MLCC groups to other care providers. CFR-MMD overcomes this by incorporating Maximum Mean Discrepancy to balance covariate distributions in representation space, reducing bias (Shalit et al., 2017). This approach improves ITE accuracy, particularly in complex scenarios such as MLCC evaluation, where covariate balance is crucial for reliable causal inference. CFR-MMD thus strengthens traditional CFR by addressing key limitations in quasi-experimental analysis (Kallus, 2020).

The CFR-MMD improves causal effect estimation by reducing selection bias through distributional alignment. The Gromov-Wasserstein Information Bottleneck framework enhances its precision in evaluating MLCC interventions compared to standard care models (Shi et al., 2019).

## Model training and evaluation

In CDL, rigorous evaluation metrics are essential for validating model performance in estimating treatment effects. The PEHE quantifies accuracy in individual-level effect estimation, with lower values indicating better performance (Lin et al., 2020). The ATE measures bias in population-level effect estimation, where values closer to zero reflect unbiased estimation. Mean Squared Error (MSE) and R-squared ($R^2$) assess predictive accuracy, though they alone cannot guarantee correct causal identification (Kiriakidou and Diou, 2022). These metrics must be evaluated collectively, as models may achieve strong prediction (high $R^2$) while failing to recover true causal relationships (high ATE error). Recent methodological work emphasizes the necessity of combining these metrics with robustness checks and out-of-sample validation to ensure strong causal inference, particularly when applying CDL methods to high-stakes domains like clinical decision-making. The optimal model should simultaneously minimize PEHE and ATE error while maintaining reasonable predictive performance (MSE, $R^2$), with preference given to methods demonstrating stability across different experimental conditions (Huang, 2022).

Beyond these causal-specific metrics, the existing classification measures, accuracy, precision, recall, and area under the curves (AUC), can offer additional insights, particularly when evaluating propensity score models or binary outcomes. Accuracy indicates overall correctness but may be unreliable in imbalanced datasets, such as those with rare treatments. Precision (the proportion of true positives among predicted positives) and recall (the ability to capture all true positives) are especially useful in clinical settings where false treatment recommendations or missed interventions carry significant consequences. AUC evaluates a model's ability to distinguish between treated and control groups, with higher values suggesting better separation. However, while these metrics help assess model reliability, they do not directly validate causal effects and should always be paired with causal-specific evaluations such as PEHE and ATE. Recent methodological work emphasizes the necessity of combining these metrics with robustness checks and out-of-sample validation to ensure strong causal inference, particularly when applying CDL methods to high-stakes domains, such as clinical decision-making. The optimal model should simultaneously minimize PEHE and ATE error while maintaining reasonable predictive performance (MSE, $R^2$), with preference given to methods that demonstrate stability across different experimental conditions.

## Learning process of artificial neural networks

Artificial neural networks, particularly multilayer perceptron's (MLPs), excel at identifying complex patterns in clinical data, such as predicting pregnancy risks or birth outcomes. By analyzing relationships between variables such as maternal health indicators and fetal growth, ANNs can uncover subtle, non-linear associations that existing statistical methods might miss. For clinicians, this means more accurate risk stratification and personalized care plans, for example, flagging high-risk pregnancies for LBW or preterm birth based on nuanced interactions between factors such as pregnancy-induced hypertension (PIH) and ANC adherence (Nahatkar et al., 2025).

Figure 2 shows a schematic representation of the mathematical model of an artificial neuron, i.e., a processing element, highlighting input $X_i$, weights ($w_0, w_1, w_{n-1}$, and $w_n$), constant [bias $(b)$], $\sum$ is the summation function, f is the activation function, and Out (y) is the output signal.

This Figure shows how a simple neural network processes medical data to support clinical decisions. Inputs (such as PIH) are assigned importance weights, combined into a weighted sum, and transformed through a step function to produce an output. For clinicians, this mirrors how we intuitively weigh multiple risk factors such as PIH, prenatal death, or meconium aspiration syndrome (MAS) to assess a patient's risk of complications. The model quantifies this decision-making process, helping to standardize predictions for outcomes such as LBW or preeclampsia based on the combined influence of key clinical variables.

Furthermore, based on the above Figure 2, an ANN is a computational model inspired by the brain's structure, made up of an input layer, one or more hidden layers, and an output layer (Bland et al., 2020). A DNN is a type of ANN with many hidden layers, allowing it to handle more complex learning tasks. Essentially, all DNNs are ANNs, but not all ANNs are DNNs; only those with multiple hidden layers qualify as DNNs (Goodfellow et al., 2016). In our work, since we use multiple hidden layers, we refer to our model as a DNN rather than a basic ANN.

## Estimating the ATE/CATE, deep learning estimation

Deep learning methods estimate CATE to reveal how MLCC differentially impacts birth outcomes across patient subgroups. ATE

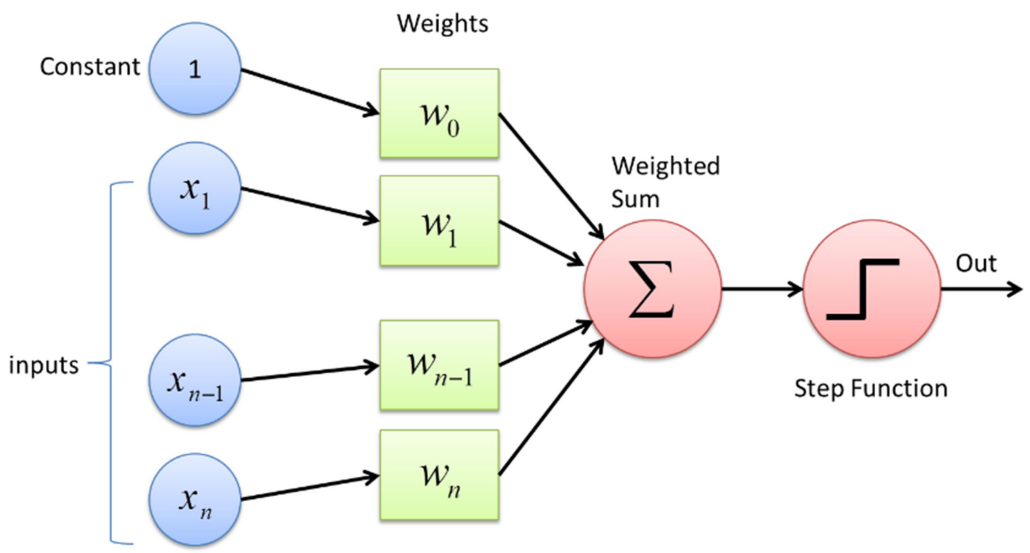


FIGURE 2
Neural network node diagram (Pang et al., 2020), each node represents a neuron in the figure. The input data flows into the neural network from the input layer. Specifically, the input data values are connected to all neurons in the first layer. Then, each neuron in a layer passes its output to all neurons in the subsequent layer. Finally, the network's output is on the image's right side. In regression tasks, this output corresponds to the dependent variable that needs to be estimated.

quantifies MLCC's average causal effect on LBW compared to other professionals, guiding targeted maternal interventions (Curth et al., 2024).

It provides an overall estimate of the MLCC effect across the entire population.

$$ATE = E\left[Y_i(1) - Y_i(0)\right] = E\left[\tau_i\right]$$

Where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes, which means NBW and LBW had the unit MLCC or did not receive the treatment, respectively.

The CATE is defined as:

$$CATE = E\left[Y(1) - Y(0) | X = x\right]$$

Where: X is the set of selected covariates, and $x \in X$. The ATE measures an intervention's overall impact across a population, guiding evidence-based clinical decisions. In contrast, PEHE assesses how accurately models predict ITE responses, enabling personalized care strategies when PEHE values are sufficiently low. Together, these metrics help clinicians balance population-level recommendations with patient-specific interventions (Kent et al., 2018; Ling et al., 2023).

The distinction between ITE and ATE is crucial for clinical interpretation. The ATE represents the average difference in outcomes between a treated group and a control group, providing an overall measure of a treatment's effectiveness across a population. In clinical practice, the ATE informs decisions about whether a treatment is beneficial on average for a broad patient population, typically guiding policy-level decisions, treatment guidelines for general conditions, and public health interventions (assessing the average impact of allocation of mothers for MLCC and other professionals). However, the ATE can mask significant heterogeneity in treatment response, meaning some individuals may benefit greatly, some may experience no effect, and others may even be harmed. This is where ITE becomes vital. The ITE, on the other hand, quantifies the specific effect of a treatment for a single individual, taking into account their unique characteristics and comorbidities. In clinical practice, ITE is increasingly sought for personalized medicine, enabling clinicians to tailor treatment decisions to individual patients. This approach allows clinicians to predict who is most likely to benefit from a particular therapy, who might experience adverse effects, and who may respond better to alternative treatments (Gopalkrishnan, 2020). Recent advancements in machine learning and causal inference are enabling better estimation of ITEs, facilitating more nuanced clinical decision-making, and optimizing patient outcomes by moving beyond a one-size-fits-all approach to treatment.

## Results

When investigating heterogeneous treatment effects, it is essential to account for confounding variables that may impact both MLCC and LBW. CDL models can integrate these variables, yielding more reliable estimates of the actual MLCC effect.

Several antenatal factors show significant associations with LBW outcomes. Iron/folic acid supplementation demonstrates a protective effect, with unsupplemented mothers showing higher LBW rates (95%, CI: 0.34–0.67, $p = 0.007$). The timing of the first antenatal care visit matters significantly; mothers with delayed initiation have increased LBW risk ($p = 0.044$). PIH emerges as a strong risk factor, with affected mothers having 58% higher LBW rates (7.45% versus 4.72%, $p = 0.036$). While nutrition counseling coverage was high (>94%), the counterintuitive finding of higher LBW among counseled mothers ($p = 0.0275$) warrants further investigation. The allocation of

TABLE 1 Association between antenatal care and low birth weight outcomes, in north Shoa zone, Amhara Region, Ethiopia.

| Variables | Category | Birth Weight, *n*, %, (95% CI) | | | |
|---|---|---|---|---|---|
| | | ≥2,500 g | <2,500 g | Chi-square | *p*-value |
| Maternal Age | < 20 years | 86,16.92, (45.3–48.4) | 106,16.11, (14.3–16.4) | 0.340 | 0.044 |
| | 20–29 years | 339,66.73, (55.2–67.4) | 437,66.41, (8.3–10.3) | | |
| | ≥30 years | 83,16.33, (42.1–53.2) | 115,17.47, (12.2–15.4) | | |
| Residence | Urban | 397, 78.15, (23.3–33.7) | 533,81.01, (0.81–0.90) | 1.446 | 0.229 |
| | Rural | 111,21.85, (26.3–45.2) | 125,18.9, (0.77–0.82) | | |
| Folic acid /Iron | Given | 483,95.1,(74.9–78.7) | 639,97.12, (1.4–2.8) | 3.265 | 0.0071 |
| | No given | 25,4.9, (23.5–28.3) | 19,2.88, (0.34–0.67) | | |
| Antenatal Care Visit | First ANC visit | 36,7.08, (6.7–7.1) | 44,6.68, (2.4–4.5) | 4.160 | 0.044 |
| | 2 to 3 visits | 209,41.2, (41.2–41.7) | 275,41.79, (12.3–14.7) | | |
| | ≥4 visit | 263,51.77, (51.5–51.7) | 339,51.52, (21.3–26.6) | | |
| Nutrition Counseling During Pregnancy | Yes | 482, 94.9, (94.9–96.2) | 633,96.2, (34.8–42.4) | 1.192 | 0.0275 |
| | No | 26,5.12, (3.8–5.2) | 25,3.79, (0.34–0.56) | | |
| Allocated group of mothers | Other groups | 250,49.22, (49.2–50.8) | 334,50.75, (15.2–16.1) | 0.275 | 0.0054 |
| | Midwife-led | 258,50.78, (49.3–50.8) | 324,49.24, (13.5–15.4) | | |
| Pregnancy-Induced Hypertension | No | 484,95.27, (92.6–95.3) | 609,92.55, (45.8–58.6) | 3.62 | 0.036 |
| | Yes | 24,4.72, (4.7–7.5) | 49,7.45, (1.3–1.7) | | |

mothers (MLCC compared to other groups) shows a small but significant difference in LBW distribution ($p = 0.0054$). These findings showed that targeted micronutrient supplementation, early antenatal care initiation, and proper management of hypertensive disorders could effectively reduce LBW incidence, as shown in Table 1.

The postnatal outcomes reveal several significant associations with LBW. Newborns with MAS showed markedly higher LBW rates [7.3% (0.34–0.78)] versus 3.14% (3.2–7.3), $p = 0.002$, indicating this complication nearly doubles LBW risk. Low Apgar scores (≤7 at 5 min) were significantly more common among LBW infants [27.5% (2.5–3.4)] versus 22.64% (27.5–27.8), $p = 0.034$. Vacuum-assisted deliveries also showed higher LBW prevalence [11.09% (4.6–5.8)] versus 8.27% (8.3–11.1), $p = 0.045$. Postnatal care patterns differed significantly ($p = 0.03$), with LBW infants more likely to receive only one visit (44.8% versus 41.92%). These findings suggest that LBW infants face greater neonatal complications and require more intensive postnatal monitoring, particularly after instrumental deliveries. The increased MAS risk specifically highlights the vulnerability of LBW newborns to birth-related complications (Table 2).

## Performance metrics of the causal deep learning model for birth weight classification

The DNN model attained a training accuracy of 81.3% and a testing accuracy of 81.4%. When analyzing the dataset composition, NBW conditions accounted for 5.1% of the training and 5.4% of the testing sets. Conversely, LBW constituted 94.9% of the training set and 94.6% of the testing set. Remarkably, the classification accuracy for both normal and LBW conditions exceeded 80% in both datasets. In the training set, 28 NBW cases were correctly classified, resulting in a sensitivity of 86.8%. However, 139 NBW instances were misclassified. For LBW, 636 cases were accurately classified, yielding a specificity of 97.8%. Unfortunately, 14 LBW instances were misclassified. In the testing sample, 8 NBW conditions were accurately classified, achieving 82.9% sensitivity. However, 54 NBW instances were misclassified. For LBW, 276 cases were accurately classified, resulting in a specificity of 96.2%. Regrettably, 11 LBW instances were misclassified. These metrics provide valuable insights into the performance across different birth weight categories in training and testing scenarios presented in Table 3.

Figure 3 presents the normalized importance of key predictors of LBW identified through the feature engineering process analysis. MAS emerged as the strongest predictor, followed by perinatal mortality and gestational age category, with PIH, previous surgery, mother's allocation, and vacuum-assisted delivery showing progressively lower predictive importance. Our analysis initially identified the top 10 variables linked to LBW. However, after validation with ensemble methods including Random Forest and XGBoost feature importance rankings along with stability selection, only seven key predictors remained consistently significant. These final variables were chosen based on statistical strength, biological relevance, and agreement across multiple causal machine learning models, balancing interpretability and generalizability, as shown in Figure 3. This ranking is associational, not causal, and guides the subsequent causal estimation of ATE, PEHE, and ITE (Moges et al., 2025).

The receiver operating characteristic (ROC) curves depict the performance of the model for both values of the dependent variable. Notably, all data points lie above the diagonal, signifying effective classification. Furthermore, the AUC, calculated from both the training and testing samples, the AUC quantifies the overall performance of the model; an AUC of 1.0 indicates a perfect model, while an AUC of 0.5 suggests random guessing. Therefore, in this study, the model's ROC curve yielded an AUC of 0.88, indicating strong classification performance, which indicates better model

TABLE 2 association between postnatal care and low birth weight outcomes, in north Shoa zone, Amhara Region, Ethiopia.

| Variables | Category | Birth weight, *n*, %, (95% CI) | | | |
|---|---|---|---|---|---|
| | | ≥2,500 g | <2,500 g | Chi-square | *p*-value |
| Initiation of breastfeeding | After 1 h | 121,23.82, (23.8–26.3) | 173,26.29, (17.6–18.3) | 0.930 | 0.335 |
| | Within 1 h | 387,76.18, (73.7–76.2) | 485,73.708, (1.9–2.3) | | |
| Meconium aspiration (newborn health Status) | No | 492,96.85, (92.7–96.9) | 610,92.7, (73.4–80.5) | 9.496 | 0.002 |
| | Yes | 16,3.14, (3.2–7.3) | 48,7.3, (0.34–0.78) | | |
| Apgar score ≤7 at 5 min (newborn health Status) | No | 393,77.36, (22.6–27.5) | 477,72.49, (26.4–29.6) | 3.589 | 0.034 |
| | Yes | 115,22.64, (27.5–27.8) | 181,27.5, (2.5–3.4) | | |
| Vacuum-assisted delivery | No | 466,91.73, (88.9–91.3) | 585,88.91, (19.4–20.5) | 2.576 | 0.045 |
| | Yes | 42,8.27, (8.3–11.1) | 73,11.09,(4.6–5.8) | | |
| Postnatal Care | 1 visit | 213,41.92, (37.7–46.3) | 295,44.8, (41.1–48.6) | 1.020 | 0.03 |
| | 2 visit | 163,32.08, (28.1–36.3) | 200,30.39, (26.9–34.1) | | |
| | 3 visit | 119,23.43, (19.8–27.3) | 148,22.49, (19.3–25.9) | | |
| | 4 visit | 13,2.55, (1.4–4.3) | 15,2.27, (1.3–3.7) | | |

TABLE 3 Classification accuracy for low birth weight prediction in training and testing datasets.

| Sample | Observed | Predicted | | Percent Correct |
|---|---|---|---|---|
| | | No | Yes | |
| Training | No | 28 | 139 | 86.8% |
| | Yes | 14 | 636 | 97.8% |
| | Overall Percent | 5.1% | 94.9% | 81.3% |
| Testing | No | 8 | 54 | 82.9% |
| | Yes | 11 | 276 | 96.2% |
| | Overall Percent | 5.4% | 94.6% | 81.4% |

performance. This substantial AUC underscores the model's high classification accuracy rate, given in Figure 4.

## Learning mechanism of deep neural networks

This DNN and MLP model uses seven input variables such as MAS, perinatal death, PIH, vacuum baby in need of resuscitation, gestational age category, mother allocation, and previous surgery to predict two output outcomes (LBW = 0) or (NBW = 1). The model's single hidden layer, comprising five nodes [H(1:1) to H(1:5)], introduces crucial nonlinearity, allowing it to capture complex relationships within the data that a simple linear model could not. Connections between layers are defined by positive and negative synaptic weights, indicating the influence of each input on hidden layer nodes and, subsequently, the final output, as shown in Figure 5. The study revealed a significant association between MAS and LBW, as evidenced by the neural network's strong positive weight (0.728) connecting MAS to hidden neuron H(1:2). However, this statistical association does not imply direct causation. MAS typically occurs secondary to fetal distress or other perinatal complications (placental insufficiency or intrauterine hypoxia), which themselves are established risk factors for LBW. The model's inhibitory weight (−0.345) between perinatal mortality and H(1:3) further suggests complex mediating pathways, where adverse perinatal outcomes may influence birth weight through multiple biological mechanisms rather than through simple direct effects. These findings emphasize that while MAS serves as a clinically useful predictor in the model, it likely represents a marker of underlying pathological processes that independently contribute to restricted fetal growth, rather than functioning as a direct causal agent of LBW. The model's architecture supports this interpretation, with hidden neurons H(1:1) and H(1:3) showing positive weights (0.165 and 0.166, respectively) for NBW outcomes, while nearly all hidden neurons exhibit negative weights for LBW predictions. So, this captures a nuanced relationship without asserting causal directionality between the observed variables.

Overall, the network architecture transforms input features through hidden layers to predict birth outcomes, with connection weights quantifying each variable's influence. MAS and gestational age emerge as predictors, demonstrating the model's ability to identify clinically significant risk factors for LBW.

This interpretation showed the DNN's ability to capture complex, non-linear relationships between predictors and outcomes (Table 4).

$$p(y) = (Bias) + (0.310 * input1) - (0.155 * input\ 2) + \ldots + (-0.261)$$

Figure 6's comparative density plot for CATE estimation offers key insights into the impact of MLCC on individual LBW outcomes. The red curve, representing prediction errors, peaks near zero, suggesting the model generally aligns with true effects. However, its spread showed that the challenge of precisely estimating ITEs. The blue curve illustrates the CATE, showing the diverse benefits individuals might gain from MLCC; its broader distribution emphasizes significant variability in individual responses. Ideally, the green curve, depicting the true impact, would reveal a bimodal distribution, indicating distinct subgroups: those who benefit substantially from MLCC and those who experience minimal or no effect. This collective view underscores the necessity of considering heterogeneity and subgroup differences to refine causal inference and tailor MLCC interventions for maximum impact.

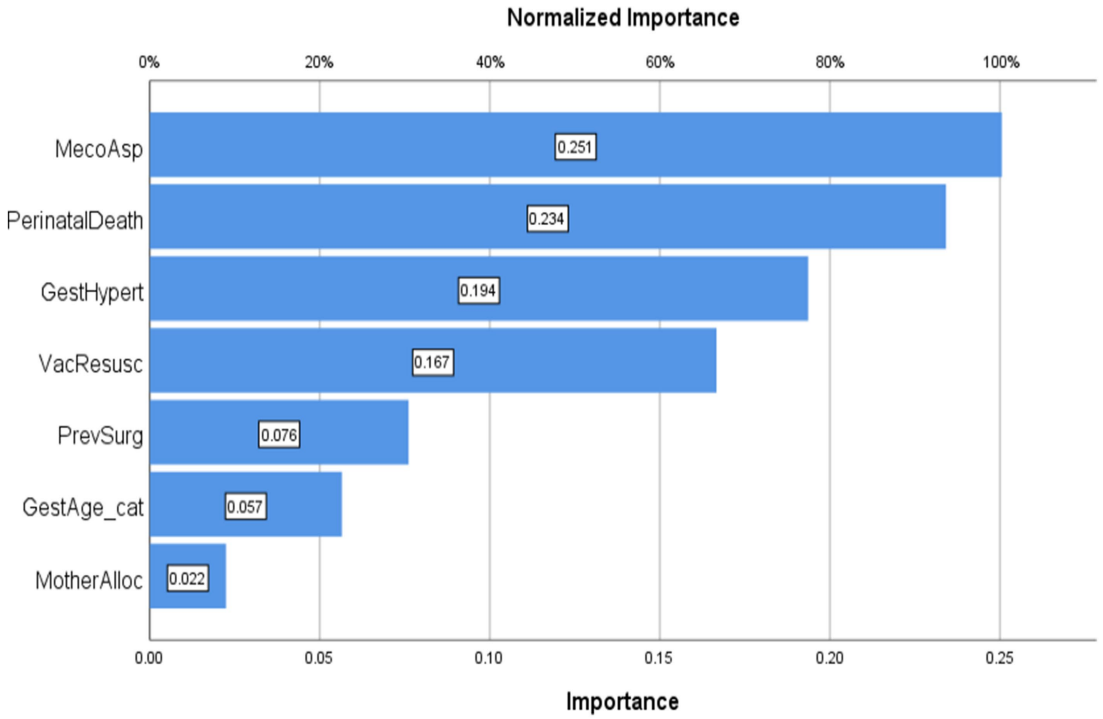


FIGURE 3
Normalized importance of variables. MecoAsp, meconium aspiration; PerinatalDeath, perinatal death; GestAge_cat, gestational age category; GestHypert, pregnancy-induced hypertension; VacResusc, vacuum baby in need of resuscitation; PrevSurg, previous surgery; MotherAlloc, mother allocation.

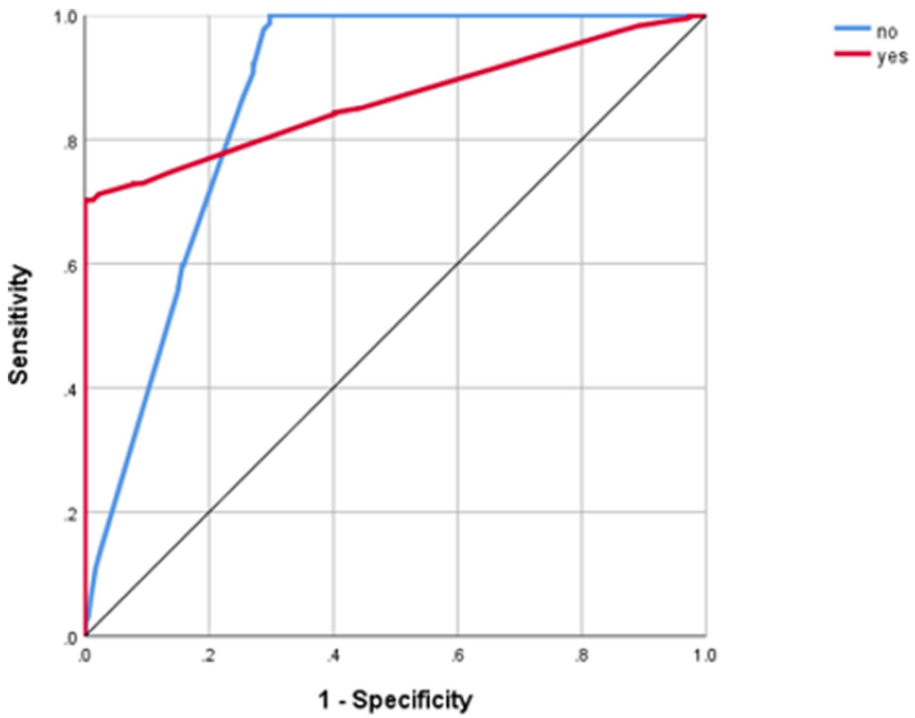


FIGURE 4
Power of receiver operating characteristic curves.

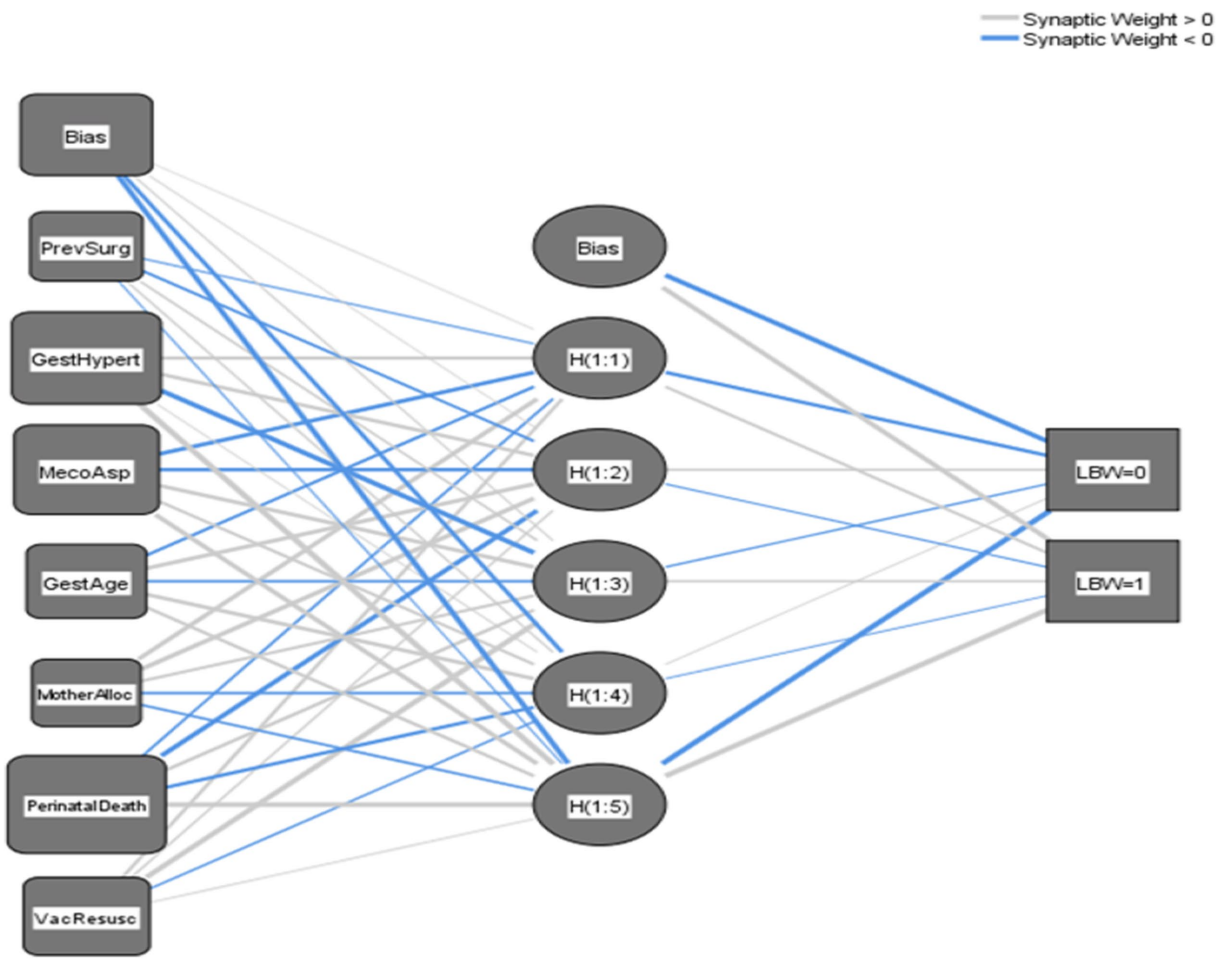


FIGURE 5
Navigating the neural network unraveling the mysteries of a three-layer multilayer perceptron. MecoAsp, meconium aspiration; PerinatalDeath, perinatal death; GestAge_cat, gestational age category; GestHypert, pregnancy-induced hypertension; VacResusc, vacuum baby in need of resuscitation; PrevSurg, previous surgery; MotherAlloc, mother allocation.

## Performance comparison of CDL models for PEHE, MSE, and ATE estimation in LBW analysis

In a comparative analysis of CDL models, CFR-WASS emerges as the top performer, demonstrating superior PEHE (lowest PEHE of 0.35) and strong classification performance across the board (84% accuracy, 82% precision, 85% recall, 0.88 AUC). This robust performance suggests its potential for accurately identifying individuals most likely to benefit from specific medical interventions, thus enabling more personalized treatment strategies. While CFR-MMD also performs well in causal effect estimation (PEHE 0.36, lowest ATE Bias 0.13), CFR-WASS combined strength in both causal and classification metrics makes it particularly promising for clinical applications where both accurate predictions and precise causal inference are critical. For scenarios with moderate-sized datasets where computational efficiency is a priority, Bayesian Ridge Regression and Bayesian Neural Networks offer a practical balance of competitive performance (PEHE 0.42, ATE Bias 0.15, and AUC 0.84–0.86) and significantly faster training times. This makes them viable options for rapid prototyping or deployment in resource-constrained medical environments (Shalit et al., 2017).

Conversely, simpler models such as Lasso Regression show limitations in causal effect estimation (higher ATE Bias 0.23), potentially due to their linear constraints. Hybrid methods, such as Sequential Feature Selection-KNN (PEHE 0.55, AUC 0.75) and Convolutional-KNN (PEHE 0.48, AUC 0.80), generally underperformed, indicating that their presumed advantages for smaller sample sizes might not hold in this context. The notable poor performance of Generative Adversarial Net for Individualized Treatment Effects (PEHE 0.60, ATE Bias 0.30, and AUC 0.71) is attributed to training instability and data inefficiency with the dataset size of 1,166, showing their unsuitability for direct causal inference in such settings. The consistent minimal precision-recall gaps among top-performing models further underscore their balanced predictive capabilities, which are vital for reliable medical decision support, as shown in Table 5.

Given the evaluation metrics in Table 5, the next step involves estimating the causal effects of MLCC on LBW compared to other care models. This will entail quantifying the PEHE to gauge individual-level causal accuracy, the ATE error to assess population-level bias,

TABLE 4 Parameter estimates hidden layers and output layers.

| Predictor | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Hidden Layer 1 | | | | | Output Layer | |
| | | H(1:1) | H(1:2) | H(1:3) | H(1:4) | H(1:5) | Normal birth weight | Low birth weight |
| Input Layer | Bias | 0.310 | −0.155 | 0.370 | 0.093 | −0.261 | | |
| | Gestational age Category | 0.087 | −0.022 | 0.415 | 0.085 | 0.471 | | |
| | Mother Allocation | 0.352 | 0.072 | 0.227 | −0.281 | 0.261 | | |
| | Perinatal Death | −0.049 | −0.029 | −0.345 | −0.060 | −0.181 | | |
| | Meconium Aspiration | 0.284 | 0.728 | −0.536 | 0.760 | −0.173 | | |
| | Pregnancy-induced hypertension | 0.144 | −0.034 | 0.249 | 0.115 | −0.002 | | |
| | Previous surgery | 0.399 | −0.045 | 0.051 | 0.314 | −0.221 | | |
| | Vacuum baby in need of resuscitation | 0.292 | −0.067 | 0.017 | 0.056 | −0.250 | | |
| Hidden Layer 1 | Bias | | | | | | 0.152 | 0.853 |
| | H(1:1) | | | | | | 0.165 | −0.150 |
| | H(1:2) | | | | | | 0.004 | −0.219 |
| | H(1:3) | | | | | | 0.166 | −0.197 |
| | H(1:4) | | | | | | 0.281 | −0.193 |
| | H(1:5) | | | | | | 0.278 | 0.144 |

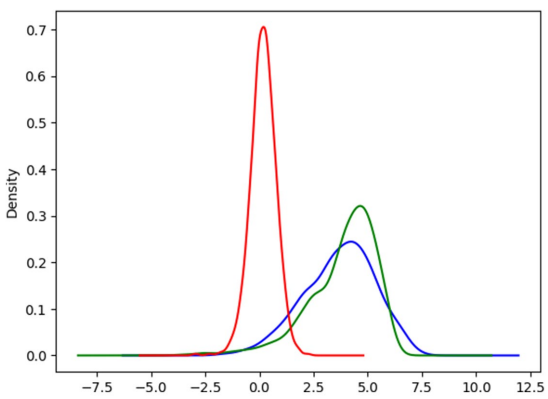


FIGURE 6
Comparative density plots for CATE estimations. Red indicates the estimated Error CATE, Blue indicates the individualized CATE, and green indicates the individualized CATE True.

and the ITE for personalized causal insights for each mother. Additionally, we will incorporate classification measures such as accuracy, precision, recall, and AUC to further evaluate the model's performance.

For clinicians, CFR-WASS, the $\sqrt{\epsilon_{PEHE}}$, and the ATE measures are vital for effective decision-making. A low PEHE of 1.006 ± 0.03 signifies the model's excellent ability to predict ITE, allowing for highly personalized interventions. Conversely, a low ATE of 0.24 ± 0.21 with high variability suggests that while there might be a small average benefit across a population, the effect on individual patients can differ significantly. Therefore, prioritizing a low PEHE is crucial for clinicians to tailor interventions precisely to each patient, moving beyond generalized averages to truly optimize care. Therefore, the proposed CFR-WASS achieved good performance and outperformed state-of-the-art models, and the model tuning or ensemble methods may enhance performance (Shi et al., 2019).

The CFR-MMD's precise estimates (ITE 0.34 ± 0.12, ATE 0.24 ± 0.01) help clinicians identify which mothers would benefit most from interventions while assessing overall treatment impact. Though results are consistent, real-world validation remains important. This approach enables targeted care for high-risk pregnancies while guiding population-level decisions about resource allocation.

CEVAE's higher $\sqrt{\epsilon_{PEHE}}$ (3.21 ± 0.32) and ATE (1.23 ± 0.23) indicate broader variability in its predictions, suggesting better capture of individual patient differences but with less precision. While this helps identify nuanced treatment responses, the wider ranges mean clinical decisions should be cautious, prioritizing high-risk cases where personalized benefits outweigh uncertainty.

The CCNN model shows moderate precision ($\sqrt{\epsilon_{PEHE}}$ 2.25 ± 0.25, ATE 2.65 ± 1.45), making it suitable for identifying general treatment trends but requiring cautious interpretation for individual cases. In contrast, Bayesian Ridge delivers more reliable

TABLE 5 Evaluation of causal deep learning models ($n$ = 1,166, 80% training, 20% testing).

| Methods | PEHE ↓ | ATE Bias ↓ | MSE ↓ | $R^2$ ↑ | Training TIME | Accuracy | Precision | Recall | AUC |
|---|---|---|---|---|---|---|---|---|---|
| Counterfactual Convolutional Neural Network | 0.43 | 0.17 | 1.11 | 0.81 | Very slow | 0.78 | 0.75 | 0.80 | 0.82 |
| Bayesian Ridge Regression | 0.42 | 0.15 | 1.08 | 0.82 | Fast | 0.80 | 0.77 | 0.82 | 0.84 |
| Lasso Regression | 0.42 | 0.23 | 1.14 | 0.75 | Slow | 0.76 | 0.73 | 0.78 | 0.79 |
| 1. Convolutional Neural Network-K-Nearest Neighbors | 0.48 | 0.20 | 1.15 | 0.78 | Fast | 0.77 | 0.74 | 0.79 | 0.80 |
| 2. Sequential Feature Selection-K-Nearest Neighbors | 0.55 | 0.26 | 1.30 | 0.72 | Slow | 0.72 | 0.70 | 0.74 | 0.75 |
| 3. Bagging Regressor | 0.45 | 0.18 | 1.12 | 0.80 | Moderate | 0.79 | 0.76 | 0.81 | 0.83 |
| 4. Treatment Agnostic Representation Network | 0.39 | 0.18 | 1.16 | | Moderate | 0.80 | 0.78 | 0.82 | 0.75 |
| 5. Bayesian Neural Network | 0.42 | 0.15 | 1.08 | 0.81 | Slow | 0.81 | 0.78 | 0.83 | 0.86 |
| 6. Counterfactual Regression-Wasserstein Distance | 0.34 | 0.11 | 1.02 | 0.83 | Moderate | 0.84 | 0.82 | 0.85 | 0.88 |
| 7. Conditional Variational Autoencoder | 0.52 | 0.24 | 1.25 | | Moderate | 0.74 | 0.72 | 0.76 | 0.77 |
| 8. Generative Adversarial Net for Individualized Treatment Effects | 0.60 | 0.30 | 1.40 | 0.68 | Extreme | 0.68 | 0.65 | 0.70 | 0.71 |
| 9. Counterfactual Regression-Maximum Mean Discrepancy | 0.36 | 0.13 | 1.16 | | Moderate | 0.83 | 0.81 | 0.84 | 0.87 |
| Deep neural network | 0.37 | 0.13 | 1.04 | | Slow | 0.82 | 0.80 | 0.83 | 0.87 |

↓ = shows decrease in value, ↑ = shows increase in value.

estimates (PEHE 1.12 ± 0.021, ATE 1.32 ± 0.75), supporting both personalized and population-level decisions. Clinically, this means Bayesian Ridge is better suited for guiding interventions, while CCNN may help screen broader risk patterns, though both benefit from further refinement for high-stakes prenatal care.

These findings indicate that our proposed model can extract deep, representative, and discriminative features related to the assigned MLCC, leading to improved ITE estimation performance. The Bayesian ridge stands out due to its high PEHE and reasonable ATE estimation. The bagging regressor performs well in ATE estimation but has moderate precision. Consider the trade-offs between PEHE and ATE estimation when choosing the best method for your specific use case.

In conclusion, the proposed CFR-WASS and CFR-MMD models demonstrated superior performance in terms of $\sqrt{\epsilon_{PEHE}}$, outperforming state-of-the-art models in estimating causal effects for LBW outcomes. Based on their CFR-WASS and CFR-MMD outperform other CDL models in estimating PEHE, ATE, and ITE, as evidenced by lower Mean ± Standard Error values in comparisons (Table 6).

A feed-forward deep neural network (FFDNN) facilitates causal inference by modeling intricate, non-linear relationships between

TABLE 6 Causal deep learning algorithm for estimating (mean ± standard error) on maternal and neonatal dataset.

| Methods | $\sqrt{\epsilon_{PEHE}}$ | $\epsilon_{ITE}$ | $\epsilon_{ATE}$ |
|---|---|---|---|
| Counterfactual Convolutional Neural Network | 2.25 ± 0.25 | 2.15 ± 1.23 | 2.65 ± 1.45 |
| Bayesian Ridge Regression | **1.12 ± 0.021** | 2.45 ± 0.75 | 1.32 ± 0.75 |
| Lasso Regression | 6.65 ± 0.45 | 5.25 ± 0.43 | 0.92 ± 0.06 |
| Convolutional Neural Network-K-Nearest Neighbors | 3.54 ± 0.52 | 2.4 ± 0.3 | 0.85 ± 0.35 |
| Sequential Feature Selection-K-Nearest Neighbors | 2.9 ± 0.41 | 2.6 ± 0.02 | 0.75 ± 0.25 |
| Bagging Regressor | 5.35 ± 1.47 | 4.14 ± 0.27 | 5.35 ± 1.32 |
| Treatment Agnostic Representation Network | **1.23 ± 0.52** | 1.53 ± 0.49 | 0.65 ± 0.16 |
| Bayesian Neural Network | 1.21 ± 0.12 | 1.09 ± 0.13 | 0.59 ± 0.01 |
| Counterfactual Regression-Wasserstein Distance | **1.006 ± 0.03** | **0.25 ± 0.01** | **0.24 ± 0.21** |
| Conditional Variational Autoencoder | 3.21 ± 0.32 | 2.15 ± 0.64 | 1.23 ± 0.23 |
| Generative Adversarial Net for Individualized Treatment Effects | 2.44 ± 0.08 | 1.24 ± 0.56 | 0.67 ± 0.14 |
| Counterfactual Regression-Maximum Mean Discrepancy | **1.012 ± 0.001** | 0.34 ± 0.12 | 0.45 ± 0.01 |
| Deep Neural Network | **1.45 ± 0.051** | 2.3 ± 0.32 | 0.78 ± 0.63 |

Bold values indicate the algorithm with the best performance (lowest mean) for each causal estimation metric. The mean shows the performance, while the standard error reflects its precision.

treatment, such as MLCC, and LBW, all while accounting for confounding factors. The network's structure allows it to generate counterfactual predictions by estimating potential outcomes for each observation under both MLCC and other professional scenarios. Through its internal layers, the FFDNN can identify varied treatment effects across different groups, offering predictions for both ITE and ATE effects. Its ability to approximate complex functions helps address the core challenge of causal inference: that only one potential outcome is ever observed for a given subject. Careful regularization and architectural design are crucial for the FFDNN to yield dependable causal predictions, rather than just correlational associations.

To solve a binary classification problem, we combine sigmoid output units with maximum likelihood. A sigmoid output unit has 2 components; one is which uses a linear layer to compute.

$z = w * h + b$, and then it uses an activation function to convert $z$ into a probability

$$Z_1 = w_{11} + w_{12} + w_{13} + w_{14} + w_{15} + w_{16} + w_{17} + W_0 + \theta_1$$

$$Z_1 = -0.738 + 0.568 + (-0.245) + (-0.8149) + (-0.897) + (-0.581) + 0.1489 + 0.99 + 0.156 = -1.413$$

Now the sigmoid function

$$f(input_1) = \frac{1}{1 + e^{-(input_1)}}$$

$$f(z) = \frac{1}{1 + e^{-(-1.413)}} = 0.1956, \text{ the predicted value of LBW}$$

$$Z_2 = w_{11} + w_{22} + w_{33} + w_{44} + w_{55} + w_{66} + w_{77} + \theta_2$$

$$Z_2 = 0.1229 + (-0.1538) + (-0.1694) + 0.3538 + 0.3494 + 0.34939 + (-0.3857) + 0.556 + 0.9875 = 2.01009.$$

$$f(input_2) = \frac{1}{1 + e^{-(input_2)}}$$

The logistics function $f(z) = \frac{1}{1 + e^{-2.01009}} = 0.882$, predicts an 88.2% probability of NBW when mothers receive MLCC, well above the 0.5 classification threshold. This demonstrates MLCC's strong protective effect against LBW, as the model consistently associates MLCC adherence with higher probabilities of normal birth outcomes. The results quantitatively confirm that structured ANC significantly reduces LBW risk, showing the MLCC's clinical importance. These findings underscore the need to expand MLCC access to improve neonatal health outcomes, as shown in Figure 7.

## Discussion

Low birth weight babies face health complications, but not all require interventions (Chen et al., 2013); the impact of LBW was associated with infant mortality and long-term health issues. The prediction of LBW at birth is based on the analysis of different characteristics of newborn babies and mothers. Some characteristics classified as the most important features include MAS, perinatal death, PIH, vacuum babies in need of resuscitation, gestational age category, mother allocation, and previous surgery. This study indicates that higher maternal weight was associated with larger birth weights in babies, which is consistent with (Alabbad et al., 2024). Our study found significant differences in folic acid/iron supplementation (97.1% versus 95.1%, $p = 0.007$) and PIH (7.5% versus 4.7%, $p = 0.036$) between LBW and normal birth weight groups. These findings align with current evidence showing iron-folate supplementation reduces

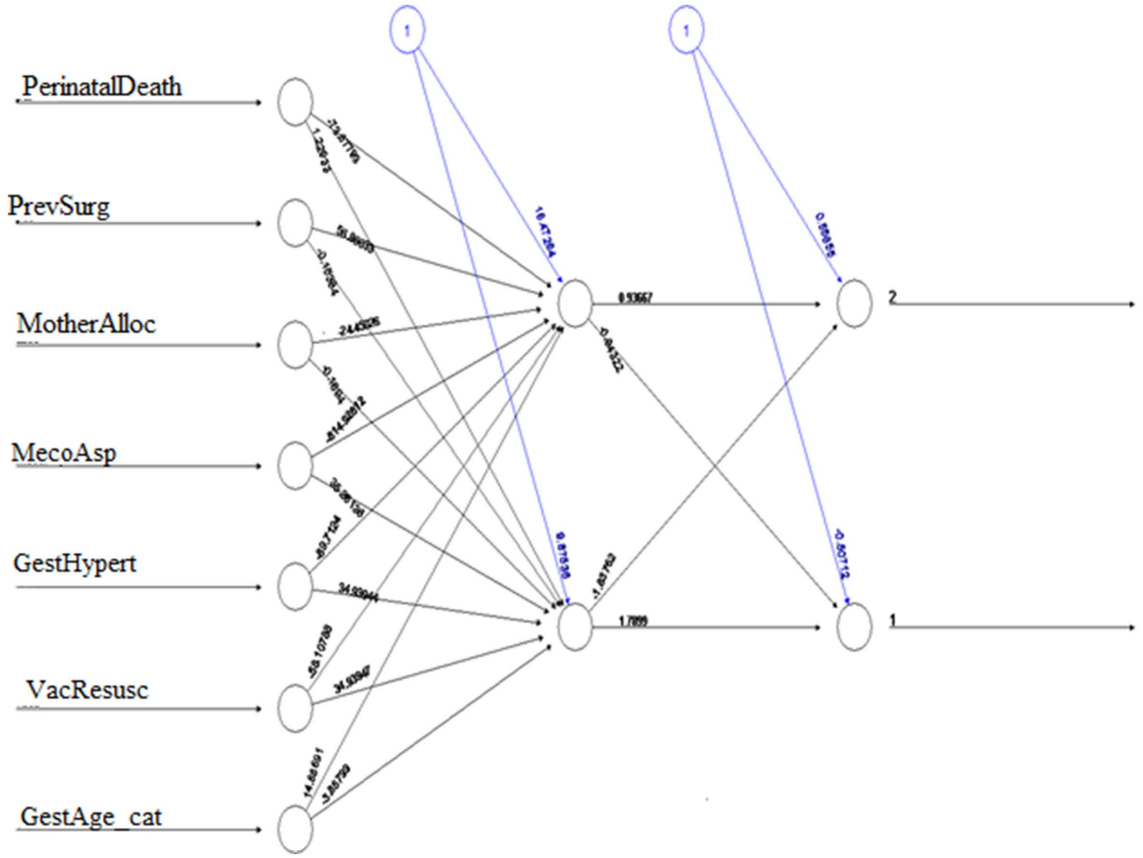


FIGURE 7
Feed-forward deep neural network. MecoAsp, meconium aspiration; PerinatalDeath perinatal death; GestAge_cat, gestational age category; GestHypert, pregnancy-induced hypertension; VacResusc, vacuum baby in need of resuscitation; PrevSurg, previous surgery; MotherAlloc, mother allocation.

LBW risk by 19% (0.81, 0.71–0.93) (Zenebe et al., 2021), while PIH increases LBW risk 2–3 fold (Yang et al., 2022). These preventable risk factors appear to play a significant role in low birth weight outcomes, underscoring the need for enhanced prenatal care strategies.

Our findings, supported by CDL feature selection methods, identified MAS as the strongest predictive factor for adverse neonatal outcomes compared to other features (Moges et al., 2025). However, while MAS is highly associated with complications in LBW infants, it does not causally precede LBW. Instead, existing literature suggests that MAS is a consequence of intrapartum events rather than a direct cause of LBW. The MAS was a significant cause of respiratory distress in newborns, occurring in about 10–15% of infants born through meconium-stained amniotic fluid (Uniyal et al., 2021). Studies have shown that MAS is more common in term babies and those with LBW (Uniyal et al., 2021). However, it can also affect post-mature and small-for-date infants and those weighing over 2,500 g (Jain et al., 2020). Critically, recent research showed that when MAS develops in infants with LBW, it is associated with a significantly increased risk of severe outcomes and mortality. Although LBW does not cause meconium passage, the physiological vulnerabilities of LBW infants, such as immature lungs, compromise their ability to cope with MAS, leading to a higher incidence of complications such as birth asphyxia, hypoxic–ischemic encephalopathy, seizures, septicemia, and persistent pulmonary hypertension (Jain et al., 2020). Consequently, MAS substantially contributes to neonatal morbidity and mortality, with birth asphyxia being a common cause of death, particularly in vulnerable groups such as LBW infants. The WHO underscores the importance of universal, high-quality perinatal care, including vigilant monitoring and appropriate resuscitation for all newborns, irrespective of gestational age or birth weight, to mitigate adverse neonatal outcomes and address the heightened risks faced by LBW infants with MAS. Early diagnosis and prompt treatment are crucial for improving outcomes (Uniyal et al., 2021; Widiyaningrum et al., 2020).

The findings of this study align with recent research that showed the importance of maternal and neonatal interventions in reducing LBW. The significant association between folic acid/iron supplementation and reduced LBW ($p = 0.0071$) is consistent with studies demonstrating the critical role of micronutrients in improving birth outcomes (Hunter et al., 2023; Johnson, 2022). Similarly, the protective effect of MLCC ($p = 0.0054$) corroborates evidence from recent trials showing that continuity of care models, particularly MLCC approaches, significantly reduce adverse neonatal outcomes (Fikre et al., 2023; Mose et al., 2023). The association between adequate ANC visits and reduced LBW ($p = 0.044$) further supports global recommendations emphasizing the importance of regular ANC visits in improving maternal and neonatal health (World Health Organization, 2021; Dandona et al., 2022).

The findings of this study indicate that ANC visits have a significant effect on the risk of LBW, which contrasts with a previous study conducted in Adwa General Hospital, Northern Ethiopia, which

reported no significant association between ANC attendance and LBW among term newborns (Gebregzabiherher et al., 2017; Hailu and Kebede, 2018). This discrepancy may be attributed to differences in study populations, quality of ANC, or healthcare infrastructure. While our results align with global evidence emphasizing the protective role of ANC in reducing LBW (Neupane et al., 2023; Katiso et al., 2020; Wachamo et al., 2019), further context-specific research is needed to explore these variations and optimize maternal care strategies.

Similarly, a 2022 systematic review published in Frontiers in Public Health demonstrated that increased ANC utilization was associated with a 20–30% reduction in LBW incidence, particularly in low- and middle-income countries where access to quality ANC remains a challenge (Engdaw et al., 2023). Another study in (Kassaw et al., 2023) emphasized that ANC visits facilitate early detection and management of conditions such as hypertension and anemia, which are known contributors to LBW. Furthermore, a 2020 WHO multi-country analysis reiterated that at least eight ANC contacts, as per the updated WHO guidelines, further reduce LBW risks by ensuring continuous maternal health monitoring and interventions. The findings highlight that ANC visits are a key modifiable factor in reducing LBW. These results align with existing research and support the need for stronger ANC policies and programs worldwide.

The identification of PIH ($p = 0.036$) and MAS ($p = 0.002$) as significant risk factors for LBW aligns with existing literature, which highlights their detrimental impact on fetal growth and neonatal outcomes (Gomez-Lumbreras et al., 2024; Adugna et al., 2025). However, unlike some prior studies, this study did not find a significant association between ANC attendance and LBW ($p > 0.05$), contrasting with evidence suggesting that adequate ANC reduces LBW risk through early detection and management of complications such as PIH (Tekeba et al., 2024). Additionally, while some studies report protective effects of urban residence and timely breastfeeding initiation against LBW, our findings showed no significant associations (residence: $p = 0.229$; breastfeeding initiation: $p = 0.335$). These discrepancies may stem from variations in healthcare access, ANC quality, or population characteristics, underscoring the need for context-specific interventions to optimize maternal and neonatal health outcomes (Basile Ibrahim et al., 2022; D'Hollander et al., 2025). In this study, we conducted a methodology scoping review, which identified DL causal predictive modeling for MLCC, with the main differences between the methods being the source of data from which the causal effects are estimated. We identified that when the causal effects required for the predictions were fully estimated from the quasi-experimental data, methods were available for predictions under MLCC. We developed a guide for the predictive analysis of PEHE in a quasi-experimental study. Predictive precision of heterogeneity treatment analysis aims at MLCC effects (Lin et al., 2021).

Significant efforts have recently been made to utilize ML techniques for causal inference problems. One notable application is estimating heterogeneous treatment effects. These efforts aim to enhance our understanding and improve outcomes in various domains (Athey and Imbens, 2016), propensity score modeling, and neighbor matching for ITE. DL, a subset of AI, is crucial in estimating MLCC effects (Ren et al., 2023; Davidson and Boland, 2021).

The findings align with (Simmelink et al., 2025), who showed the role of leadership and collaborative efforts in successfully implementing MLCC. Our results further demonstrate that MLCC, when supported by robust policies and interdisciplinary collaboration, can significantly reduce adverse neonatal outcomes such as LBW.

The DNN has been employed to estimate heterogeneous treatment effects within the causal inference framework. DL's ability to handle complex confounding factors is valuable for understanding LBW and enhancing outcomes (Koch et al., 2025). In this paper, we built a DNN classifier, Propensity Net, for propensity score-based matching to estimate ITE and ATE (Ramachandra, 1803).

The results show that CFR-WASS achieves the best performance with the highest accuracy (84%), precision (82%), recall (85%), and AUC (0.88), along with the lowest PEHE (0.34) and ATE bias (0.11). This aligns with recent findings by (Shalit et al., 2017), who demonstrated that Wasserstein-based methods excel in causal inference by effectively balancing covariate distributions. Similarly, the strong performance of CFR-MMD (AUC = 0.87) supports (Håkansson et al., 2020) work on distribution matching for unbiased treatment effect estimation. In contrast, simpler models such as Lasso Regression (AUC = 0.79) and Generative Adversarial Net (AUC = 0.71) underperform, consistent with (Koch et al., 2024), who showed their limitations in handling complex causal relationships. These findings reinforce the superiority of advanced CDL methods in precision medicine applications. The high AUC scores($\geq$0.86) for top models validate their discriminative power, supporting their use in precision healthcare applications.

The DNN model's performance (81.3% accuracy) aligns with findings from (Keles and Bagci, 2023), who reported that DL models in neonatology achieve high accuracy (95% for Retinopathy of Prematurity diagnosis) but often lack interpretability. Our use of neural network weight analysis (e.g., H(1:2) for MAS) addresses this limitation by providing clinically actionable insights, a direction recommended for future AI applications in neonatal care (Keles and Bagci, 2023).

In a DNN, parameter estimates serve as independent variables. These estimates typically correspond to weights and biases associated with neuron connections. In DNN architecture, a single hidden layer is utilized. Models with additional layers did not perform well (Montesinos López et al., 2022; Hassoun, 1995). Hidden layers play a crucial role in capturing non-linear patterns within the data. Without hidden layers, the DNN behaves similarly to a linear regression model, unable to detect nonlinearity (Hussain et al., 2019). In our chosen model, the hidden layer consists of five nodes (neurons). Each node represents a specific combination of input features. The DNN exhibits nonlinearity because the effects at each node vary. Some independent variables have positive effects for one set of observations while having adverse effects for another set. This dynamic behavior results in mean scores near zero, reflecting the intricate interplay of variables (Alzubaidi et al., 2021).

The MLCC model demonstrates better clinical outcomes than standard care, with lower rates of medical interventions (epidurals, forceps delivery, episiotomies) and higher rates of natural births and patient satisfaction (Sandall et al., 2016). Patient outcomes under MLCC were assessed by clinicians using standardized protocols, ensuring reliable validation. A structured medical records system was essential for tracking care continuity, enabling consistent evaluations and data-driven improvements in maternal and neonatal health (Hailemeskel et al., 2022).

## Strengths, limitations, and future work

This study contributes to the growing body of literature on causal inference using DNN. Its primary strength lies in addressing causal inference through the potential outcome framework, building and

optimizing custom DL models for causal estimation, and adapting these models to predict PEHE effects on LBW. DL models offer significant advantages, such as automatically extracting relevant features from data, reducing the need for manual feature selection, and effectively capturing non-linear relationships. However, the study focuses on quasi-experimental designs, which inherently face limitations due to confounding and uncontrolled variables. While DL models excel in handling complex data, they struggle with complex confounding structures and often lack interpretability, making it challenging to understand the underlying causal mechanisms.

Quasi-experimental designs are particularly prone to selection bias due to non-random assignment, leading to imbalances between MLCC and other professional groups and potentially compromising the validity of results. Additionally, unmeasured confounders can obscure causal relationships, further complicating accurate effect estimation. To address these challenges, we recommend future research to explore Double/Debiased Machine Learning (DML), a state-of-the-art algorithm that provides unbiased, root-n-consistent estimators for ATE, heterogeneous treatment effects, and their confidence intervals. DML enhances adjustments for non-linear confounding relationships, offering a more robust approach to causal inference in complex datasets. By integrating DML, future studies can improve the accuracy and reliability of causal estimates, advancing the intersection of DL and causal inference in healthcare and beyond.

## Conclusion

In this study, we employed DL causal inference techniques, such as CCNN, CFR-WASS, causal effect variational autoencoder, and balancing neural network, for measuring the effectiveness of PEHE, ITE, and ATE of LBW predictions for capturing more complex patterns and relationships of the given data. The analysis revealed that MAS was the strongest predictor, but other factors such as gestational age and perinatal mortality also played a role.

In this study, the DNN model delivered reliable results, reaching 81.3% accuracy on the training set and 81.4% on the test set, indicating stable predictive performance. Its impressive AUC score of 0.88 further validates its ability to accurately predict LBW. The analysis of the hidden layer identified $H(1:1)$ for the allocation of the mother, $H(1:2)$ for the allocation of the mother, $H(1:3)$ for perinatal death, and $H(1:3)$ for MAS with positive and negative influences on LBW, respectively. Therefore, this indicated that, hidden layer provided insights into the specific influences of various factors on LBW.

CFR-WASS outperformed all other models, achieving the highest accuracy (84%), precision (82%), recall (85%), and AUC (0.88). It also had the lowest errors, with a PEHE of 0.34 and ATE bias of 0.11. In predicting LBW, CFR-WASS maintained strong performance with a PEHE of 1.006 and an ATE of 0.24, surpassing competing methods. Both CFR-WASS and CFR-MMD effectively estimated causal effects, showing that their potential to enhance maternal and neonatal healthcare interventions, particularly by evaluating the impact of ANC visits on LBW risk. In Addition, the FFDNN model, using a sigmoid function, predicted a higher probability of 0.882 of NBW for newborns whose mothers followed MLCC compared to a lower probability of 0.1956 for LBW, reinforcing the importance of adequate ANC in improving birth outcomes. These findings showed that a critical role of MLCC in reducing LBW, particularly in resource-limited settings. By ensuring consistent antenatal monitoring, timely folic acid/iron supplementation, and adherence to ANC visits, MLCC models demonstrate promise in improving maternal and neonatal outcomes. Future research should prioritize cost-effectiveness analyses and implementation strategies to scale MLCC programs, informing policies aimed at LBW prevention and maternal–infant health equity.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving humans were approved by Ethical approval was obtained from the Institutional Review Board of Bahir Dar University, Ethiopia, Ethical Review Board referencing number RCS/1412/2022. All experiments adhered to relevant guidelines and regulations, including the Declaration of Helsinki. In the context of the current investigation, informed consent has been waived due to the use of secondary data. The Board of Bahir Dar University waived informed consent, Institutional Review Board (IRB). Researchers were provided with the patient's ID number and other essential factors related to the experiment. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

## Author contributions

WM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AT: Conceptualization, Validation, Visualization, Writing – review & editing. AM: Methodology, Software, Supervision, Writing – review & editing. ET: Data curation, Writing – review & editing. SH: Data curation, Writing – review & editing.

## Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Adugna, M., Asmare, K., and Wondim, A. (2025). Meconium aspiration syndrome and associated factors among neonates admitted at the neonatal intensive care unit at Northwest Ethiopia comprehensive specialized hospitals, Northwest Ethiopia 2023. *BMC Pediatr.* 25:167. doi: 10.1186/s12887-024-05181-4

Alabbad, D. A., Ajibi, S. Y., Alotaibi, R. B., Alsqer, N. K., Alqahtani, R. A., Felemban, N. M., et al. (2024). Birthweight range prediction and classification: a machine learning-based sustainable approach. *Mach. Learn. Knowledge Extract.* 6, 770–788. doi: 10.3390/make6020036

Alam, M.A., Zaman Sajib, M. R. U., Rahman, F., Ether, S., Hanson, M., Sayeed, A., et al. (2023). Implication of big data analytics, artificial intelligence, machine learning and deep learning in the public health sector of Bangladesh: a scoping review protocol Open Science Framework.

Alzubaidi, L., Zhang, J., Humaidi, AJ., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* 8, 1–74. doi: 10.1186/s40537-021-00444-8

Arayeshgari, M., Najafi-Ghobadi, S., Tarhsaz, H., Parami, S., and Tapak, L. (2023). Machine learning-based classifiers for the prediction of low birth weight. *Healthcare Inform. Res.* 29, 54–63. doi: 10.4258/hir.2023.29.1.54

Athey, S., and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* 113, 7353–7360. doi: 10.1073/pnas.1510489113

Basile Ibrahim, B., Interrante, J. D., Fritz, A. H., Tuttle, M. S., and Kozhimannil, K. B. (2022). Inequities in availability of evidence-based birth supports to improve perinatal health for socially vulnerable rural residents. *Children* 9:1077. doi: 10.3390/children9071077

Berrevoets, J., Kacprzyk, K., Qian, Z., and van der Schaar, M. (2024). Causal deep learning: encouraging impact on real-world problems through causality. Foundations and trends®. *Signal Process.* 18, 200–309. doi: 10.1561/2000000123

Bland, C., Tonello, L., Biganzoli, E., Snowdon, D., Antuono, P., Lanza, M., et al. (2020). Advances in artificial neural networks. *Adv. Artif. Neural Netw.* 2020:119. doi: 10.1136/bmj.m2922

Campos Trujillo, O., Perez-Gonzalez, J., and Medina-Bañuelos, V.. (2020). "Early prediction of weight at birth using support vector regression". In VIII Latin American conference on biomedical engineering and XLII National Conference on biomedical engineering: proceedings of CLAIB-CNIB 2019, October 2-5, 2019, Cancún, México. Springer,

Chen, Y., Li, G., Ruan, Y., Zou, L., Wang, X., and Zhang, W. (2013). An epidemiological survey on low birth weight infants in China and analysis of outcomes of full-term low birth weight infants. *BMC Pregnancy Childbirth* 13, 1–9. doi: 10.1186/1471-2393-13-242

Curth, A., Peck, R. W., McKinney, E., Weatherall, J., and van der Schaar, M. (2024). Using machine learning to individualize treatment effect estimation: challenges and opportunities. *Clin. Pharmacol. Therap.* 115, 710–719. doi: 10.1002/cpt.3159

D'Hollander, C. J., McCredie, V. A., Uleryk, E. M., Kucab, M., le, R. M., Hayosh, O., et al. (2025). Breastfeeding support provided by lactation consultants: a systematic review and meta-analysis. JAMA. *Pediatrics* 179, 508–520. doi: 10.1001/jamapediatrics.2024.6810

Dandona, R., Majumder, M., Akbar, M., Bhattacharya, D., Nanda, P., Kumar, G. A., et al. (2022). Assessment of quality of antenatal care services in public sector facilities in India. *BMJ Open* 12:e065200. doi: 10.1136/bmjopen-2022-065200

Davidson, L., and Boland, M. R. (2021). Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes. *Brief. Bioinform.* 22:bbaa369. doi: 10.1093/bib/bbaa369

Desta, M. (2019). "Low birth weight and adverse perinatal outcomes," in *Childbirth* (Alphen aan den Rijn, The Netherlands: IntechOpen).

Endalamaw, A., Engeda, E. H., Ekubagewargies, D. T., Belay, G. M., and Tefera, M. A. (2018). Low birth weight and its associated factors in Ethiopia: a systematic review and meta-analysis. *Ital. J. Pediatr.* 44, 1–12. doi: 10.1186/s13052-018-0586-6

Engdaw, G. T., Tesfaye, A. H., Feleke, M., Negash, A., Yeshiwas, A., Addis, W., et al. (2023). Effect of antenatal care on low birth weight: a systematic review and meta-analysis in Africa, 2022. *Front. Public Health* 11:1158809. doi: 10.3389/fpubh.2023.1158809

Fan, Z., Kernan, K. F., Sriram, A., Benos, P. V., Canna, S. W., Carcillo, J. A., et al. (2023). Deep neural networks with knockoff features identify nonlinear causal relations and estimate effect sizes in complex biological systems. *GigaScience* 12:giad044. doi: 10.1093/gigascience/giad044

Fernandez Turienzo, C., Rayment-Jones, H., Roe, Y., Silverio, S. A., Coxon, K., Shennan, A. H., et al. (2021). A realist review to explore how midwifery continuity of care may influence preterm birth in pregnant women. *Birth* 48, 375–388. doi: 10.1111/birt.12547

Fikre, R., Gubbels, J., Teklesilasie, W., and Gerards, S. (2023). Effectiveness of midwifery-led care on pregnancy outcomes in low-and middle-income countries: a systematic review and meta-analysis. *BMC Pregnancy Childbirth* 23:386. doi: 10.1186/s12884-023-05664-9

Gebregzabiherher, Y., Haftu, A., Weldemariam, S., and Gebrehiwet, H. (2017). The prevalence and risk factors for low birth weight among term newborns in Adwa general hospital, northern Ethiopia. *Obstet. Gynecol. Int.* 2017, 1–7. doi: 10.1155/2017/2149156

Ghosh, I., Sanyal, M. K., and Jana, R. (2018). "An integrated approach incorporating nonlinear dynamics and machine learning for predictive analytics and delving into causal interaction" in Intelligent engineering informatics: Proceedings of the 6th international conference on FICTA. ed. Z. Sheng (Boca Raton, FL, USA: Springer).

Gomez-Lumbreras, A., Vilaplana-Carnerero, C., Lestón Vázquez, M., Vedia, C., Morros, R., and Giner-Soriano, M. (2024). Treatment of hypertension during pregnancy: a cohort of pregnancy episodes from the SIDIAP database, Catalonia, Spain. *Front. Pharmacol.* 15:1346357. doi: 10.3389/fphar.2024.1346357

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning 1. Cambridge: MIT Press.

Gopalkrishnan, R. (2020). Advances in deep generative modeling for clinical data. Basel, Switzerland: Massachusetts Institute of Technology.

Hailemeskel, S., Alemu, K., Christensson, K., Tesfahun, E., and Lindgren, H. (2022). Midwife-led continuity of care improved maternal and neonatal health outcomes in north Shoa zone, Amhara regional state, Ethiopia: a quasi-experimental study. *Women Birth* 35, 340–348. doi: 10.1016/j.wombi.2021.08.008

Hailu, L. D., and Kebede, D. L. (2018). Determinants of low birth weight among deliveries at a referral Hospital in Northern Ethiopia. *Biomed. Res. Int.* 2018, 1–8. doi: 10.1155/2018/8169615

Håkansson, S., Lindblom, V., Gottesman, O., and Johansson, FD. (2020). Learning to search efficiently for causally near-optimal treatments. *Adv. Neural Inf. Proces. Syst.* 33, 1333–1344. doi: 10.1111/aogs.14039

Hassoun, M. H. (1995). Fundamentals of artificial neural networks. Stuttgart, Germany: MIT press.

Huang, Y. (2022). "Robust causal learning for the estimation of average treatment effects." In the 2022 International Joint Conference on Neural Networks (IJCNN). IEEE.

Hunter, P. J., Muthiani, Y., Näsänen-Gilmore, P. K., Koivu, A. M., Pörtfors, P., Bastola, K., et al. (2023). A modular systematic review of antenatal interventions to address undernutrition during pregnancy in the prevention of low birth weight. *Am. J. Clin. Nutr.* 117, S134–S147. doi: 10.1016/j.ajcnut.2023.01.024

Hussain, H. I., Kamarudin, F., Thaker, H. M. T., and Salem, M. A. (2019). Artificial neural network to model managerial timing decision: non-linear evidence of deviation from target leverage. *Int. J. Comput. Intell. Syst.* 12, 1282–1294. doi: 10.2991/ijcis.d.191101.002

Jain, K., Kothari, R. B., Mhaske, S. N., Misal, G. B. (2020). To study the complications of meconium aspiration syndrome & their outcome about Gestational age & Birth Weight in tertiary health care Centre. *VIMS Health Sci. J.* 7, 47–49. doi: 10.46858/vimshjs.7203

Jiao, L., Wang, Y., Liu, X., Li, L., Liu, F., Ma, W., et al. (2024). Causal inference meets deep learning: a comprehensive survey. *Research* 7:0467. doi: 10.34133/research.0467

Johnson, J. (2022). Maternal micronutrient status as a risk factor for perinatal mortality in a Tanzanian randomized controlled trial. *Afr. J. Reprod. Health* 26, 38–48. doi: 10.29063/ajrh2022/v26i7.5

Kallus, N. (2020). Deepmatch: balancing deep covariate representations for causal inference using adversarial training. In international conference on machine learning. PMLR.

Kassaw, A., Kassie, Y. T., Kefale, D., Azmeraw, M., Arage, G., Asferi, W. N., et al. (2023). Pooled prevalence and its determinants of stunting among children during their critical period in Ethiopia: a systematic review and meta-analysis. *PLoS One* 18:e0294689. doi: 10.1371/journal.pone.0294689

Katiso, N. A., Kassa, G. M., Fekadu, G. A., Kidanemariam Berhe, A., and Muche, A. A. (2020). Prevalence and determinants of low birth weight in Ethiopia: a systematic review and meta-analysis. *Adv. Public Health* 2020, 1–13. doi: 10.1155/2020/7589483

Keles, E., and Bagci, U. (2023). The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review. *NPJ Digit. Med.* 6:220. doi: 10.1038/s41746-023-00941-5

Kent, D. M., Steyerberg, E., and Van Klaveren, D. (2018). Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*:k4245. doi: 10.1136/bmj.k4245

Khan, W., Zaki, N., Masud, M. M., Ahmad, A., Ali, L., Ali, N., et al. (2022). Infant birth weight estimation and low birth weight classification in United Arab Emirates using machine learning algorithms. *Sci. Rep.* 12:12110. doi: 10.1038/s41598-022-14393-6

Kiriakidou, N., and Diou, C. (2022). "An evaluation framework for comparing causal inference models." In Proceedings of the 12th Hellenic Conference on Artificial Intelligence.

Koch, B. J., Park, D. E., Hungate, B. A., Liu, C. M., Johnson, J. R., and Price, L. B. (2024). Predicting sepsis mortality in an era of pandrug-resistant E. coli through modeling. *Commun. Med.* 4:278. doi: 10.1038/s43856-024-00693-7

Koch, B. J., Sainburg, T., Geraldo Bastías, P., Jiang, S., Sun, Y., and Foster, J. G. (2025). A primer on deep learning for causal inference. *Sociol. Methods Res.* 54, 397–447. doi: 10.1177/00491241241234866

Koivu, A. M., Näsänen-Gilmore, P. K., Hunter, P. J., Muthiani, Y., Isojärvi, J., Heimonen, O., et al. (2023). Antenatal interventions to address harmful behaviors and psychosocial risk factors in the prevention of low birth weight. *Am. J. Clin. Nutr.* 117, S148–S159. doi: 10.1016/j.ajcnut.2022.11.028

Kong, W., Qiu, M., Li, M., and Xuanyu, J. (2022). "Causal graph convolutional neural network for emotion recognition." in IEEE Transactions on Cognitive and Developmental Systems.

Lin, A., Lu, J., Xuan, J., Zhu, F., and Zhang, G. (2020). A causal Dirichlet mixture model for causal inference from observational data. *ACM Trans. Intellig. Syst. Technol.* 11, 1–29. doi: 10.1145/3379500

Lin, L., Sperrin, M., Jenkins, D. A., Martin, G. P., and Peek, N. (2021). A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn. Progn. Res.* 5, 1–16. doi: 10.1186/s41512-021-00092-9

Ling, Y., Upadhyaya, P., Chen, L., Jiang, X., and Kim, Y. (2023). Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: methodology review and benchmark. *J. Biomed. Inform.* 137:104256. doi: 10.1016/j.jbi.2022.104256

Lu, Y., Zhang, X., Fu, X., Chen, F., and Wong, K. K. L. (2019). Ensemble machine learning for estimating fetal weight at varying gestational age. *Proc. AAAI Conf. Artif. Intell.* 33, 9522–9527. doi: 10.1609/aaai.v33i01.33019522

Mayberry, L. J., Avery, M. D., Budin, W., and Perry, S. (2017). Improving maternal and infant outcomes by promoting normal physiologic birth on hospital birthing units. *Nurs. Outlook* 65, 240–241. doi: 10.1016/j.outlook.2017.02.007

Mbogu, H.M., (2023). Improving outcomes in machine learning and data-driven learning systems using structural causal models.

Meinshausen, N. (2018). Causality from a distributional robustness point of view. In the 2018 IEEE data science workshop (DSW). IEEE.

Moges, W. K., Tegegne, A. S., Mitku, A. A., Tesfahun, E., and Hailemeskel, S. (2025). Causal machine learning models for predicting low birth weight in midwife-led continuity care intervention in north Shoa zone, Ethiopia. *BMC Med. Inform. Decis. Mak.* 25:64. doi: 10.1186/s12911-025-02917-9

Montesinos López, O. A., Montesinos López, A., and Crossa, J. (2022). "Fundamentals of artificial neural networks and deep learning" in Multivariate statistical machine learning methods for genomic prediction. ed J. Abdulqader (Cham, Switzerland: Springer), 379–425.

Moreira, M.W. L., Rodrigues, J., Furtado, V., Mavromoustaki, C., Kumar, N., Woungang, I., et al. (2019). Fetal birth weight estimation in high-risk pregnancies through machine learning techniques. In ICC 2019–2019 IEEE International Conference on Communications (ICC). 2019. IEEE.

Mose, A., Fikadu, Y., Zewdie, A., Haile, K., Shitu, S., Wasie Kasahun, A., et al. (2023). Pregnant women's perception of midwifery-led continuity care model in Ethiopia: a qualitative study. *BMC Womens Health* 23:304. doi: 10.1186/s12905-023-02456-3

Nahatkar, S., et al (2025). "Collating random forest classifier and artificial neural networks for the risk detection of maternal health." In the 2025 International Conference on Computational, Communication and Information Technology (ICCCIT). IEEE.

Neupane, S., Scott, S., Piwoz, E., Kim, S. S., Menon, P., and Nguyen, P. H. (2023). More is not enough: high quantity and high quality antenatal care are both needed to prevent low birthweight in South Asia. *PLoS Glob. Public Health* 3:e0001991. doi: 10.1371/journal.pgph.0001991

Pang, B., Nijkamp, E., and Wu, Y. N. (2020). Deep learning with tensorflow: a review. *J. Educ. Behav. Stat.* 45, 227–248. doi: 10.3102/1076998619872761

Park, J. J. H., Harari, O., Siden, E., Zoratti, M., Dron, L., Zannat, N. E., et al. (2020). Interventions to improve birth outcomes of pregnant women living in low-and middle-income countries: a systematic review and network meta-analysis. *Gates Open Res.* 3:1657. doi: 10.12688/gatesopenres.13081.2

Pattanayak, S., Pattanayak, J. S., and John, S. (2017). Pro deep learning with tensorflow. Washington, DC, USA: Springer.

Ramachandra, V. (1803). Deep learning for causal inference. *ArXiv preprint* 149:2018.

Ren, Y., Wu, D., Tong, Y., López-DeFede, A., and Gareau, S. (2023). Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: establishment of benchmarking key machine learning models with data rebalancing strategies. *J. Med. Internet Res.* 25:e44081. doi: 10.2196/44081

Reza, T. B., and Salma, N. (2024). Prediction and feature selection of low birth weight using machine learning algorithms. *J. Health Popul. Nutr.* 43:157. doi: 10.1186/s41043-024-00647-8

Roslina, E. (2020). Relationship between antenatal care and low birth weight (LBW) incidence in Indonesia: secondary data analysis of IDHS 2017. *Indian J. Public Health Res. Dev.* 11:1179.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Muller, K. R. (2021). Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109, 247–278. doi: 10.1109/JPROC.2021.3060483

Sandall, J., Soltani, H., Gates, S., Shennan, A., and Devane, D.Cochrane Pregnancy and Childbirth Group (2016). Midwife-led continuity models versus other models of care for childbearing women. *Cochrane Database Syst. Rev.* 2016:4667. doi: 10.1002/14651858.CD004667

Sandall, J., Fernandez, C., Turienzo, Devan, D., Soltani, H., Gillespie, P., et al. (2024). Midwife continuity of care models versus other models of care for childbearing women. *Cochrane Database Syst. Rev.* 2024:CD004667. doi: 10.1002/14651858.CD004667.pub6

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. in the international conference on machine learning. PMLR.

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Adv. Neural Inf. Proces. Syst.* 32, 1–44. doi: 10.48550/arXiv.1905.12741

Simmelink, R., Neppelenbroek, E., Pouwels, A., van der Lee, N., Pajkrt, E., Ziesemer, K. A., et al. (2025). Understanding how midwife-led continuity of care can be implemented and under what circumstances: a realist review. *BMJ Open* 15:e091968. doi: 10.1136/bmjopen-2024-091968

Sunarni, N., Noviati, E., Kurniawan, R., and Mulyati, N. (2018). Kelengkapan Antenatal Care (ANC) Dengan Kejadian BBLR. *Jurnal Kesehatan STIKES Muhammadiyah Ciamis* 5, 42–51. doi: 10.52221/jurkes.v5i2.35

Tao, J., Yuan, Z., Sun, L., Yu, K., and Zhang, Z. (2021). Fetal birthweight prediction with measured data by a temporal machine learning method. *BMC Med. Inform. Decis. Mak.* 21, 1–10. doi: 10.1186/s12911-021-01388-y

Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers* 12:91. doi: 10.3390/computers12050091

Tekeba, B., Techane, M. A., Workneh, B. S., Zegeye, A. F., Gonete, A. T., Ahmed, M. A., et al. (2024). Determinants of preterm birth among reproductive age women in sub-Saharan Africa: evidence from the most recent demographic and health survey data-2019-2022. *PLoS One* 19:e0305810. doi: 10.1371/journal.pone.0305810

Uauy, R., Corvalan, C., Casanello, P., and Kuzanovic, J. (2013). "Intervention strategies for preventing low birthweight in developing countries: Importance of considering multiple interactive factors" in Maternal and child nutrition: The first 1,000 days (Washington, DC, USA: Karger Publishers), 31–52.

Uniyal, P., Kalra, B., and Wasim, S. (2021). A study of profile of meconium aspiration syndrome in relation with birth weight and gestational age of newborns and their immediate outcome. *Int. J. Contemp. Pediatr.* 8, 1021–1026. doi: 10.18203/2349-3291.ijcp20212041

Voon, S. T., Lay, J. T. S., San, W. T. W., Shorey, S., and Lin, S. K. S. (2017). Comparison of midwife-led care and obstetrician-led care on maternal and neonatal outcomes in Singapore: a retrospective cohort study. *Midwifery* 53, 71–79. doi: 10.1016/j.midw.2017.07.010

Wachamo, T. M., Bililign Yimer, N., and Bizuneh, A. D. (2019). Risk factors for low birth weight in hospitals of north Wello zone, Ethiopia: a case-control study. *PLoS One* 14:e0213054. doi: 10.1371/journal.pone.0213054

Whata, A., and Chimedza, C. (2022). Evaluating uses of deep learning methods for causal inference. *IEEE Access* 10, 2813–2827. doi: 10.1109/ACCESS.2021.3140189

Widiyaningrum, A. N., Pamungkasari, E. P., and Murti, B. (2020). Meta-analysis: amniotic meconium and low birth weight as predictors of asphyxia in newborns. *J. Matern. Child Health* 5, 590–600. doi: 10.26911/thejmch.2020.05.05.12.

World Health Organization (2021). WHO antenatal care recommendations for a positive pregnancy experience. Nutritional interventions update: Zinc supplements during pregnancy. Geneva, Switzerland: World Health Organization.

Yang, L., Wang, W., Mao, B., Qiu, J., Guo, H., Yi, B., et al. (2022). Maternal folic acid supplementation, dietary folate intake, and low birth weight: a birth cohort study. *Front. Public Health* 10:844150. doi: 10.3389/fpubh.2022.844150

Yu, M. H., and Kang, D. R. (2019). Improving causal inference in observational studies: propensity score matching. *Cardiovasc. Prevent. Pharmacotherapy* 1, 57–62. doi: 10.36011/cpp.2019.1.e6

Zawadzki, R. S., Grill, J. D., and Gillen, D. L.and for the Alzheimer's Disease Neuroimaging Initiative (2023). Frameworks for estimating causal effects in observational settings: comparing confounder adjustment and instrumental variables. *BMC Med. Res. Methodol.* 23:122. doi: 10.1186/s12874-023-01936-2

Zenebe, A., Tegegne, K. T., Bifato, B., and Assefa, A. A. (2021). Association between iron and folic acid supplementation and birth weight in Ethiopia: systemic review and Meta-analysis. *Bull. Natl. Res. Cent.* 45:135. doi: 10.1186/s42269-021-00595-x

Zhang, Y.-F., Zhang, H., Lipton, Z. C., Li, L. E., and Xing, E. P. (2022). Exploring transformer backbones for heterogeneous treatment effect estimation. *ArXiv preprint* 2202. doi: 10.48550/arXiv.2202.01336