BMC Infectious Diseases

RESEARCH

Open Access

Data-driven machine learning algorithm model for pneumonia prediction and determinant factor stratification among children aged 6–23 months in Ethiopia

Addisalem Workie Demsash[1*], Rediet Abebe[1], Wubishet Gezimu[2], Gemeda Wakgari Kitil[2], Michael Amera Tizazu[1], Abera Lambebo[1], Firomsa Bekele[3], Solomon Seyife Alemu[4], Mohammedamin Hajure Jarso[4], Geleta Nenko Dube[2], Lema Fikadu Wedajo[3], Sanju Purohit[5,6] and Mulugeta Hayelom Kalayou[7]

Abstract

Introduction Pneumonia is the leading cause of child morbidity and mortality and accounts for 5.6 million under-five child deaths. Pneumonia has a significant impact on the quality of life, the country's economy, and the survival of children. Therefore, this study aimed to develop data-driven predictive model using machine learning algorithms to predict pneumonia and stratify the determinant factors among children aged 6–23 months in Ethiopia.

Methods A total of 2035 samples of children were used from the 2016 Ethiopian Demographic and Health Survey dataset. Jupyter Notebook from Anaconda Navigators was used for data management and analysis. Important libraries such as Pandas, Seaborn, and Numpy were imported from Python. The data was pre-processed into a training and testing dataset with a 4:1 ratio, and tenfold cross-validation was used to reduce bias and enhance the models' performance. Six machine learning algorithms were used for model building and comparison, and confusion matrix elements were used to evaluate the performance of each algorithm. Principal component analysis and heatmap function were used for correlation detection between features. Feature importance score was used to identify and stratify the most important predictors of pneumonia.

Results From 2035 total samples, 16.6%, 20.1%, and 24.2% of children had short rapid breath, fever, and cough respectively. The overall magnitude of pneumonia among children aged 6–23 months was 31.3% based on the 2016 EDHS report. A random forest algorithm is the relatively best performance model to predict pneumonia and stratify its determinates with 91.3% accuracy. The health facility visits, child sex, initiation of breastfeeding, birth interval, birth weight, husbands' education, women's age, and region, are the top eight important predictors of pneumonia among children with important scores of more than 5% to 20% respectively.

Conclusions Random forest is the best model to predict pneumonia and stratify its determinant factors. The implications of this study are profound for advanced research methodology, tailored to promote effective health interventions such as lifestyle modification and behavioral intervention, based on individuals' unique features, specifically

*Correspondence:
Addisalem Workie Demsash
addisalemworkie599@gmail.com
Full list of author information is available at the end of the article

for stakeholders to take proactive childcare interventions. The study would serve as pioneering evidence for future research, and researchers are recommended to use deep learning algorithms to enhance prediction accuracy.

Keywords Data-driven, Prediction Model, Pneumonia, Children, Machine learning

Introduction

Pneumonia is an infectious disease caused by viruses and bacteria that commonly affect the respiratory organs [1]. Though pneumonia can cause mild to severe illness in people of all ages, it is the most significant infectious disease and cause of death in children worldwide [2]. Globally, pneumonia is the leading cause of child death and accounts for 5.6 million under-five child deaths [3]. The annual incidence of clinical pneumonia in under-five children is approximately 152 million globally [4]. Approximately 156 million cases of pneumonia occurred among children, of which 151 million episodes of pneumonia are found in developing countries, and 1.2 million pneumonia cases last in death [5]. Pneumonia combined with malaria, diarrhea, and HIV/ADIS cases accounted for 1 million child deaths around the world [3].

In most developing countries, the mortality rates of children range from 60 to 100 per 1000 live births, and 20% of child deaths are caused by pneumonia [6]. Nearly 172 under-five child deaths from a total of 1,000 live births occur in sub-Saharan African countries [7]. Pneumonia and its complications are very common in LMICs. For instance, a systematic review and meta-analysis of pneumonia among under-five children in East Africa is 34% [5]. A study from Nigeria states that 54.3% of children have pneumonia case and heart failure complication [8]. Based on the national surveillance report of Uganda, the incidence of severe pneumonia is 108 per 100,000 children under five years in 2022 [9]. In Burkina Faso, 21% of children aged 12 months, and 15% of children aged 1–4 years were confirmed for severe pneumonia from a total of 17% of pneumonia cases identified from a total of 2843 patients [10].

In Ethiopia, pneumonia is a major risk factor for child morbidity and mortality. An estimated 3,370 children encounter pneumonia annually, and over 40,000 children's deaths are affected by pneumonia [11]. More than pneumonia infections and associated deaths, pneumonia has a significant clinical and economic burden, on the access and utilization of healthcare resources [12]. Pneumonia infection affects the life expectancy of the populations [13], and significantly lowers the quality of life [14]. Studies from different regions of Ethiopia show that the prevalence of pneumonia is 30% in the Gamo zone of the southern region [7], 18.5% in Northwest Amhara region [15], 18.1% of caprine pleuropneumonia case in Gambela region [16], and 285 children are admitted for severe pneumonia in the Tigray region [17].

Among the factors that affect pneumonia infectious and prevalence, nonexclusive breastfeeding [7], age of children [18], vaccine inaccessibility and incompleteness, environmental conditions [19], air pollution, food insecurity, undernutrition, and micronutrient deficiency [20] are significant risk factors for pneumonia [21, 22]. Parents' educational status, wealth status [23], place of residency, place of delivery, birth interval, birth order, age of children, and vaccination during and after pregnancy are also predictors for pneumonia infection among children [24]. Following the World Health Organization's Integrated Management of Childhood Illness recommendations [25], Ethiopia introduced community-based pneumonia screening and treatment to control and minimize the risk of pneumonia [7, 26]. Moreover, child vaccinations [27], nutritional supplementations [28], environmental hygiene, and sanitation services [29] are provided to reduce the risk of pneumonia.

Regardless of the efforts made to reduce pneumonia infections, the problem and risks of pneumonia and its complications among children are increasing. For instance, a study shows that children less than 24 months have a higher frequency of complications associated with pneumonia [8]. This shows that pneumonia is a serious public health problem among young children [30].

Even if many studies have been done about pneumonia and its associated factors, little is known and investigated about pneumonia case prediction, pneumonia development after respiratory tract infection [31], risk factors stratification, and pneumonia prediction using a large dataset [32]. Additionally, the previous studies are affected by recall bias, have temporal relationships as data is not managed using advanced data management techniques such as machine learning techniques, and lack representativeness as data have been acquired from selected institutions or communities. Thus, the methods often fail to provide specific insights, which can lead to ineffective decision-making in healthcare settings, unable to capture hidden patterns and complex relationships between features [33]. Moreover, the traditional analysis methods rely on assumptions, have a static nature in model building, and are not able to integrate diverse data sources, which potentially provides misleading results [34]. Thus, employing machine learning algorithms to

develop a predictive model for pneumonia and determinant factors stratification among children is crucial for understanding the prediction magnitude, and determinant factors of pneumonia.

Recently, machine learning algorithms have been increasingly growing research methods to investigate large amounts of health data generated from various health institutions. The algorithms are important for generating and discovering insights from a large amount of dataset [32, 35]. The need for a data-driven machine learning algorithm model for pneumonia prediction and risk factor stratification among children aged 6–23 months is critical due to the high prevalence of pneumonia in this vulnerable age group, which significantly contributes to child morbidity and mortality. The development of tailored algorithms that account for unique demographic, environmental, and healthcare factors is essential to effectively address the lack of specific insights in traditional analysis reports [36]. Leveraging machine learning can enhance accuracy for problem detection and provide intervention strategies, ultimately reducing the burden of pneumonia through timely and targeted healthcare resources [37].

Moreover, this research can inform public health policies and improve health outcomes by identifying risk factors through guiding preventive measures and optimizing resource allocation in healthcare systems. The findings would provide evidence for policymakers and stakeholders to take interventions, and work on the predictive factors. Moreover, this research would serve as input and state-of-the-art evidence for further similar research initiatives. Therefore, this study aimed to develop a data-driven predictive model of pneumonia and determinant factors stratification among children aged 6–23 months in Ethiopia.

Methods

Data source

The dataset was accessed from the 2016 Ethiopian Demographic and Health Survey (EDHS) survey report, available from the DHS program website (https://dhsprogram.com). The Ethiopian Public Health Institute collected the survey data in collaboration with the Central Statistical Agency cross-sectionally.

Study design and setting

A cross-sectional study design is used for this study as long as the data is collected cross-sectionally by the Ethiopian Public Health Institute. The data was collected across nine regions and two city administrations of Ethiopia. Thus, this study used nationally representative data to study pneumonia among children aged 6–23 months in Ethiopia. Ethiopia is located in the Horn of Africa, bordered by Eritrea in the north, Kenya in the south, Sudan in the west, and Djibouti in the east.

Sampling techniques and procedures

The sampling frame used for the 2016 EDHS is a frame of all Census Enumeration Areas (EAs) created for the 2016 Ethiopia Population and Housing Census and conducted by the Central Statistical Agency. A two-stage stratified cluster sampling technique was used. First, the regions were stratified into urban and rural areas. In the second stage, a household listing operation was used as a sampling frame for household selection. Finally, a fixed number of households were selected in each cluster, and samples of EAs were selected independently in each stratum.

Study populations

All living children were the source population, and all sampled living children aged 6–23 months were the study population. Children aged 6–23 months are more likely vulnerable to pneumonia incidence and severity, and many vaccines are administered in this period to prevent pneumonia infection. Thus, children aged 6–23 months are included in this study. Based on the set of measurement items of pneumonia and the data management process, the children aged 6–23 months have more sufficient data as compared with children aged below 6 and above 23 months, to provide valid and generalizable reports. Thus, children under 6 months and above 23 months were excluded to maximize the validity and generalizability of the findings. Details about the methodology of the data source, sampling procedure, and source population were presented in the 2016 EDHS report from the Measure DHS website [38].

Study variables

Dependent variable

Pneumonia among children aged 6–23 months.

Independent variables

Socio-demographic characteristics such as region, residency, wealth status, educational status of mothers, husbands' education, mothers' age, sex, and age of children were used as sociodemographic characteristics of the study participants. Place of delivery, health facility visits, breastfeeding initiation, birth interval, birth order, ANC visit, working status, and media exposure was also an independent predictor used to develop a data-driven predictive model for pneumonia infection among children aged 6–23 months in Ethiopia.

Operationalizations

Pneumonia among children

Pneumonia is a dependent variable in this study. The cough, fever, and short or rapid breaths among children are used as symptoms of pneumonia. Thus, children had pneumonia if the children had either one of the symptoms based on the verbal report of mothers two weeks before the survey [39].

Birth interval

The period between two successive live births is a birth interval. For this study, a birth interval of < 33 months between two consecutive live births is a *short birth interval*, whereas a birth interval of 33 and above is an *optimum birth interval* [40, 41].

ANC visits

These pregnant women who had visited health facilities for ANC services at most three times are considered inadequate. Otherwise, the women had adequate ANC visits [42, 43].

Media exposure

If the mothers had access to either radio or television or both, then the mothers had media exposure; and if mothers did not have any means of media access, then the mothers had no media exposure [44].

Breastfeeding initiation

The baby has early breastfeeding initiation If the child was provided the mother's breast milk within 1 h after birth. Otherwise, late initiation of breastfeeding [34].

Data management and analysis

The dataset was downloaded in STATA format, and the data cleaning and labeling were done. Jupyter Notebook in Anaconda Navigator was used for data management and model building. Important libraries such as Pandas, NumPy, and Seaborn were imported from Python software. Python packages such as matplotlib and sklearn were used to import important supervised machine-learning algorithms for data analysis and visualization. The missing values, outliers, noise, and duplication are very common in the dataset and were checked in the preprocessing stage of the machine learning concept [45, 46]. The data types of each feature are converted to numeric data types for ease of outlier detection. Missing values, outliers' detection, and duplication were detected using isnull(), the threshold Z score, and duplicated() function respectively. Accordingly, missing values were not detected in the original data set. The threshold Z score values were computed using the mean and variance of each feature. Thus, the Z score value for each feature ranges from 3 and − 3, which indicates the absence of outliers in the dataset. Then all the features which have not any missing values, outliers, and duplications are retained for further correlation analysis and final feature selections.

Correlation detection and feature selection

For data-driven predictive model development for pneumonia, the variable was checked for their dependency and correlation before the feature selection process to keep the loss of important variables twice. Managing and removing the correlated features in the dataset is important for the reduction of high dependency between features, and improve the model performance [45, 46], and crucial to focus on the most informative features [47]. In this study, a heatmap() function was built for correlation analysis by importing the seaborn library from Python. A variable that correlated with other variables was dropped from the dataset. The variables that insignificantly correlated were kept for the final important feature selection and prediction of pneumonia and to stratify the risk factors among children aged 6–23 months using a more accurate algorithm. In the heatmap function, features with a correlation level between 0 and 0.5 are considered insignificantly correlated, while features with a correlation greater than 0.5 are deemed significantly correlated. Additionally, a principal component analysis (PCA) was used for the dimensionality reduction technique that can reveal relationships among variables by transforming them into a set of uncorrelated variables [48]. The PCA divides the dataset into different components and each component shows how many variations exist in the original dataset. The higher variance of a component indicates that the variables in a component are more likely represented and show the absence of significant redundancy or correlation in the dataset.

Methods for Feature Importance Score Measurement

The feature importance score (FIS) is a technique used to evaluate and order the contribution of each feature in a model to determine their prediction power for pneumonia among children [49]. It helps to understand which features are most influential in making predictions, enabling better model interpretation, feature selection, and insights into the data. A Gini impurity is used to calculate FIS considering how much a feature impurity is reduced, and it explains the contribution of each feature to the prediction for individual instances. In the FIS measurement, variables that have relatively higher Gini Impurity scores are considered as the most important or first important feature, and the next important features are stratified based on their order of FIS values.

Data split and model selection

The datasets were divided into training and testing datasets considering the K-fold cross-validation technique for bias reduction and model performance enhancement. K-fold cross-validation divides the dataset into nearly K-equal subsets or folds. The process partitions the dataset into K folds for the validation set, and the remaining K- 1 folds for training. The process iterated repeated K times until the validation set is ensured by either one of each fold. However, the choice of K is crucial, as smaller values can lead to high bias and larger values can increase computational cost. To control such problems, a maximum of tenfold cross-validation technique was considered. A total of 2035 records were included for model development and prediction of pneumonia among children aged 6–23 months. To ensure the normalization of the data, sklearn.preprocessing packages were imported from Python, and the training and testing dataset was fitted using the MinMaxScaler() function.

From the previous research reports, various machine learning algorithms such as Naïve Bayes, logistic regression, multilayer perceptron, random forest, decision tree classifier, support vector machine, and K-nearest neighbor random forest are used for predictive model development for different health parameters among various study populations [34, 50, 51]. In this study, six machine-learning algorithms are considered, and details about each algorithm are stated as follows.

Naïve Bayes

Naïve Bayes is a probabilistic classifier based on Bayes'theorem, which assumes independence among predictors [34]. It is widely used in data classification due to its simplicity and speed, making it suitable for large datasets [52]. However, its main limitation lies in the assumption of feature independence, which is often unrealistic in practice. Additionally, it can struggle with small datasets and issues related to zero probabilities.

Logistic Regression

Logistic regression is a statistical method used to predict binary outcomes by modeling the probability that a given input belongs to a particular category [53]. It is easy to implement and interpret, providing probabilities for class membership, which adds significant insight into the model's predictions [52]. However, logistic regression assumes a linear relationship between the features. This makes it less suitable for complex relationships, limiting its applicability in certain aspects.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and merges them to enhance predictive accuracy. By reducing the risk of overfitting associated with individual trees, Random Forest is robust and can handle various data types effectively [37]. It also accommodates missing values and provides insights into feature importance. However, it can be computationally intensive and slower than other models, especially with large datasets, and its complexity may make it less interpretable compared to single decision trees.

Decision Tree Classifier

The Decision Tree Classifier is a model that makes decisions based on feature values by creating a tree-like structure of branches [54]. This approach is straightforward to understand and interpret, requiring minimal data preprocessing, as it can handle both numerical and categorical data. However, decision trees are prone to overfitting, particularly when they become too complex. They are also sensitive to small variations in the training data, which can lead to significant changes in the resulting model structure [55].

Support Vector Machine

Support Vector Machine is a powerful supervised learning model that finds the hyperplane that best separates classes in a high-dimensional space. It is particularly effective when there is a clear margin of separation between classes and is robust against overfitting in high-dimensional scenarios [35]. Despite its strengths, support vector machines can be less effective on very large datasets and require careful tuning of parameters, such as kernel choice, to achieve optimal performance.

K-Nearest Neighbor

K-Nearest Neighbor is a non-parametric classification method that assigns a class to an instance based on the majority class among its k-nearest neighbors in the feature space [56]. Its simplicity and intuitive nature make it easy to implement, and it does not require a training phase, as it merely stores the dataset [37]. However, the K-Nearest Neighbor is computationally expensive during prediction, as it calculates distances to all training samples. It is also sensitive to the choice of distance metric and the value of k, and its performance can degrade with high-dimensional data due to the curse of dimensionality.

Model building and evaluation

Model evaluation

The confusion matrix model was used to determine the algorithm's performance [57]. The confusion matrix elements such as true positive, false positive, true negative, and false negative, and receiver operators' curve (ROC) were also used for model evaluation based on sensitivity,

and specificity relationships. The ROC is based on probability, the area under the ROC curve (AUC) is crucial to representing the degree or measure of separability, and it's important to differentiate the model's ability to predict the classes. Hence, the algorithm with higher AUC values shows the most accurate algorithm [58]. Moreover, the accuracy, precision, recall, and f1-measure were used to determine the machine learning algorithms'performance. Accuracy measures the overall correctness of the model to predict pneumonia cases by calculating the ratio of true results to the total instances of pneumonia. Precision is vital in pneumonia prediction because high precision means that the model is reliable in identifying actual pneumonia cases, and vital when false positives can lead to unnecessary treatments and healthcare costs. The model's ability to identify all pneumonia cases, reflects how well it captures true positives measured by the recall and is vital to detect missing pneumonia cases (false negative), specifically in medical diagnosis, and decline child morbidity due to pneumonia [59]. F1 Measure provides a balance between precision and recall, making it particularly useful in pneumonia prediction where both false positives and false negatives carry significant risks. It helps in understanding the balance between false positives and false negatives, ensuring a comprehensive evaluation of model performance for pneumonia prediction [60]. The mathematical relationship of the performance measurement indicators is described as follows:

$$\text{Accuracy} = \frac{TP + TN}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \quad (1)$$

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{TP}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F1} - \text{score} = 2 * \frac{Precision * Recall}{\text{Precision} + \text{Recall}} \quad (4)$$

The machine learning workflow for pneumonia prediction begins with data aquation and preprocessing, which includes cleaning the dataset, removing null values, and resampling. A random dataset is taken to create a training and testing dataset. The process involves cross-validation and hyperparameter tuning to optimize model performance. Various machine learning models are then employed, followed by performance evaluation. The final step produces predictions regarding pneumonia among children, confirming the presence or absence of pneumonia in each class. The overall methodology flow of the study is presented in Fig. 1.

Result

Descriptive result

Sociodemographic characteristics of the study participants

A total of 2035 study participants were included for analysis. Nearly, three-tenth (31.0%) of children's mothers were under 25–29 years. Nine hundred seventy-seven (48.0%), and seven out of ten (69.0%) of children's fathers and mothers had no formal education respectively. The majority (43.5) of the respondents were from the Oromia region, and 40.5% were Muslim religious followers. Nearly three-fourths (67.9%) and 44.6% of children's parents had no media exposure and were under poor wealth status respectively. The majority (90.2%) of household heads were male, and 73.5% of the respondents did not work during the survey (Table 1).

The characteristics of children and mothers

Most (90.4%) of children were found feeding breast during the interview period, and 77.7% initiated their breastfeeding within one hour after birth. 17.9% and 52.6% of children had short birth intervals and less than five birth orders. The majority (52.6%) of children were female, and 56.1% of the children had received tetanus injections more than once before birth. The majority of mothers gave birth at home (67.5%), had inadequate ANC visits (67.4%), had no health facility visits in the last 12 months (51.3%) before the survey, and never attended school in their life (69.0%) till the interview period (Table 2).

Magnitude of pneumonia, fever, cough, and short/rapid breath among children

As shown in Fig. 2, 16.6%, 20.1%, and 24.2% of children aged 6–23 months had short or rapid breath, fever, and cough respectively. The overall magnitude of pneumonia among children aged 6–23 months is 31.3% (95% CI: 29.29%– 33.32%) based on the 2016 EDHS data report.

Correlation matrix and dimensionality reduction

The correlation function analysis indicates the presence of a significant correlation (90%) between birth order and women's age, and 80% between wealth index and media exposure in the given dataset. The yellow color indicates the correlation of each predictor with itself with the value of 100%. The dark and light blue colors show the relationship between each predictor

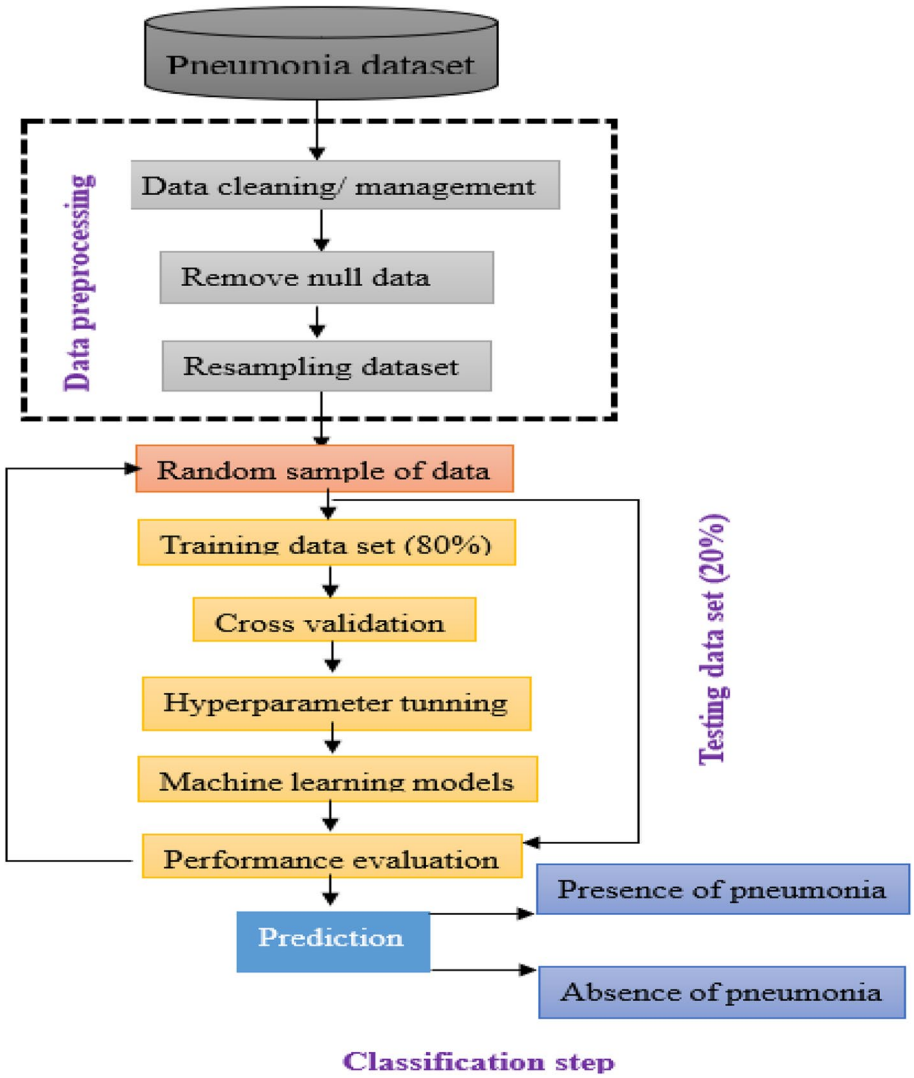

Fig. 1 The workflow of the entire machine methodology in this study

with values ranging from 0 to 0.4 which indicates the absence of a significant correlation between the predictors (Fig. 3).

Moreover, the principal component analysis before the dimensionality reduction shows a 54% and 17% variance for Component 1 and Component 2 respectively. This means that principal component 1 is approximately 3.2 times more significant to explain the total variance in the original dataset than principal component 2. Additionally, 54% shows the need for dimensionality reduction in the dataset. After the dimensionality reduction, principal component 1 counted 64% power to explain the total variance, and underly the structure of the original dataset is 9 times well repressed by principal component 1 as compared with principal component 2 (Fig. 4).

Data imbalance detection and management

Data imbalance usually occurs in the dataset, showing unequal class representation. The data imbalance leads biased model that performs poorly on underrepresented classes. The data imbalance is detected through class distribution analysis using bar plot visualization. After the data imbalance is detected, a synthetic minority over-sampling technique (SMOTE) is used to generate synthetic samples for the minority class to ensure that the data is balanced between classes. Accordingly, a 21.9% synthetic sample was generated for the minority class (class 0), and a similar number of samples was reduced from the overfitted class (class 1) using SMOTE (Fig. 5).

Table 1 Sociodemographic characteristics of the study participants, 2016 EDHS data

Variable	Category	Frequency (n)	%
Age of mothers (year)	15–19	11	.5
	20–24	323	15.9
	25–29	630	31.0
	30–34	577	28.3
	35–39	334	16.4
	40–44	126	6.2
	45–49	35	1.7
Husbands' educational status	No education	977	48.0
	Primary	858	42.2
	Secondary	136	6.7
	Higher	64	3.1
Mothers' educational status	No education	1404	69.0
	Primary	551	27.1
	Secondary	53	2.6
	Higher	27	1.3
Region	Tigray	149	7.3
	Afar	19	.9
	Amhara	383	18.8
	Oromia	885	43.5
	Somali	78	3.8
	Benishangul	22	1.1
	SNNPR	449	22.0
	Gambela	4	.2
	Harari	4	.2
	Addis Adaba	35	1.7
	Dire Dawa	7	.3
Media access and exposure	No	1382	67.9
	Yes	654	32.1
Respondents currently working	No	1496	73.5
	Yes	540	26.5
Wealth status	Poor	907	44.6
	Middle	464	22.8
	Rich	664	32.6
Religion	Orthodox	681	33.4
	Catholic	26	1.3
	Protestant	460	22.6
	Muslin	824	40.5
	Traditional and others	44	2.3
Sex of household head	Male	1835	90.2
	Female	200	9.8

Data-driven model performance comparison

The data-driven model developments were done considering six machine learning algorithms such as support vector machine, K-nearest neighbors, random forest, decision tree classifier, Gaussian Naïve Bayes, and logistic regression algorithms. Based on predictive model development, a random forest algorithm had relatively high performance to predict pneumonia among children aged 6–24 months compared to other data-driven machine learning algorithms, with an AUC value of 91.8%. Furthermore, the same line in the curve shows, that logistic regression and support vector machines demonstrate 70.3% of AUC to predict pneumonia in children 6 to 24 months of age. Whereas, the Gaussian Naïve Bayes shows the lowest AUC value of 68%, a relatively low-performance algorithm to predict pneumonia among children aged 6–24 months (Fig. 6).

Confusion matrix

To evaluate the performance of the included machine learning algorithm; accuracy, precision, recall, F1-score, and the confusion matrix elements were used for the data-driven predictive models. Accordingly, the random forest algorithm stands out with relatively high accuracy (83.3%), high precision (87.3%), high recall (77.2%), and high F1 score (93.7%). Based on all performance measure metrics, the decision tree algorithm shows the second-best algorithm next to the random forest with 81.4% of accuracy, 84.3% of precision, 76.8% of recall, and 91.1% of F1 score. With an accuracy of 61.1%, precision of 64.3%, recall of 49.7%, and F1-score of 69.3%, the decision Naïve Bayes algorithm demonstrates relatively the least performance algorithm. Its low recall indicates that the algorithm struggles to identify positive instances effectively, leading to more false negatives (Table 3).

Determinant factor stratification for pneumonia among children

A random forest algorithm was used to stratify determinants of pneumonia among children aged 6–12 months. Based on the algorithm's report, all the included predictors have their impact on pneumonia infection with FIS value range from 1.25% to 20%. Importantly, region, women's age, husbands' education, birth weight, birth interval, initiation of breastfeeding, child six, and health facility visits were stratified as the top seven important predictors of pneumonia with an important score value of 5% to 20% (Fig. 7).

Prediction of pneumonia

The final step of the machine learning model is prediction of pneumonia among children aged 6–23 months. Accordingly, the predicted value of pneumonia among children aged 6–23 months is presented along with its actual value in the confusion matrix model. In the confusion matrix, the actual values for children with pneumonia and without pneumonia are represented by 1 (No) and 0 (Yes) compared to against predicted values. Of the total of 1007 children with pneumonia cases, 81.6% of

Table 2 Characteristics of the children and mothers, 2016 EDHS data

Variable	Category	Frequency (n)	%
Currently breastfeeding	No	195	9.6
	Yes	1840	90.4
Initiation of breastfeeding	Early	1581	77.7
	Late	454	22.3
Preceding birth interval	Short	364	17.9
	Optimal	622	30.5
	Larger	1050	51.6
Birth order	< 5	1070	52.6
	> = 5	965	47.4
Sex of children	Male	954	46.9
	Female	1081	53.1
Number of tetanus injections before birth	No injection	894	43.9
	One and more injection	1141	56.1
Place of delivery	Home	1374	67.5
	Health facility	662	32.5
Health facility visits in the last 12 months	No	1044	51.3
	Yes	991	48.7
Respondents ever attended school	No	1404	69.0
	Yes	632	31.0
ANC visits	Inadequate	1373	67.4
	Adequate	663	32.6

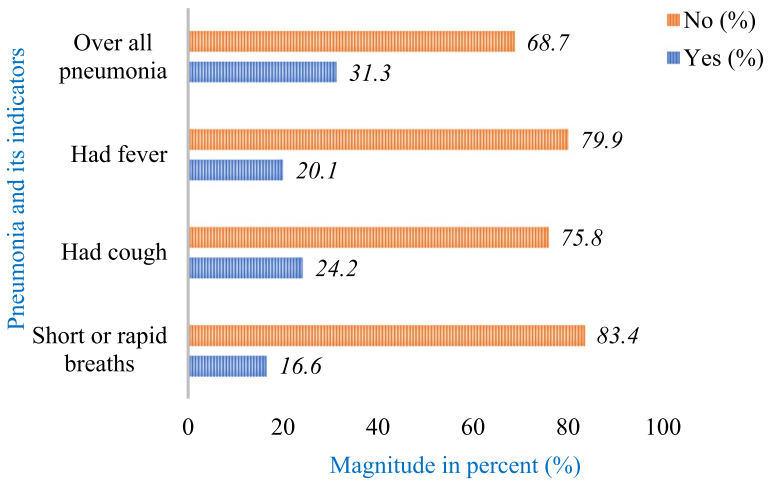

Fig. 2 Magnitude of pneumonia, fever, cough, and short/rapid breath among children aged 6–23 months in Ethiopia using the 2016 EDHS dataset

children were correctly predicted as positive (True Positives), while 18.6% of children were incorrectly predicted as negative or not confirmed for pneumonia (False Negatives). Additionally, from a total of 1028 children without pneumonia cases, 70.1% of children were accurately predicted as negatives (True Negatives), and 29.9% of children without actual pneumonia cases were misclassified as positive or confirmed for pneumonia (False Positives). The classification report shows 81.4% and 70.1% precision for children with pneumonia and without pneumonia cases, indicating that the model is relatively reliable in predicting pneumonia cases. Recall values show that about 73% of actual children with pneumonia cases and, 79.3% of children without pneumonia cases were correctly identified. The F1 scores for both classes highlighted the presence of a good balance between precision

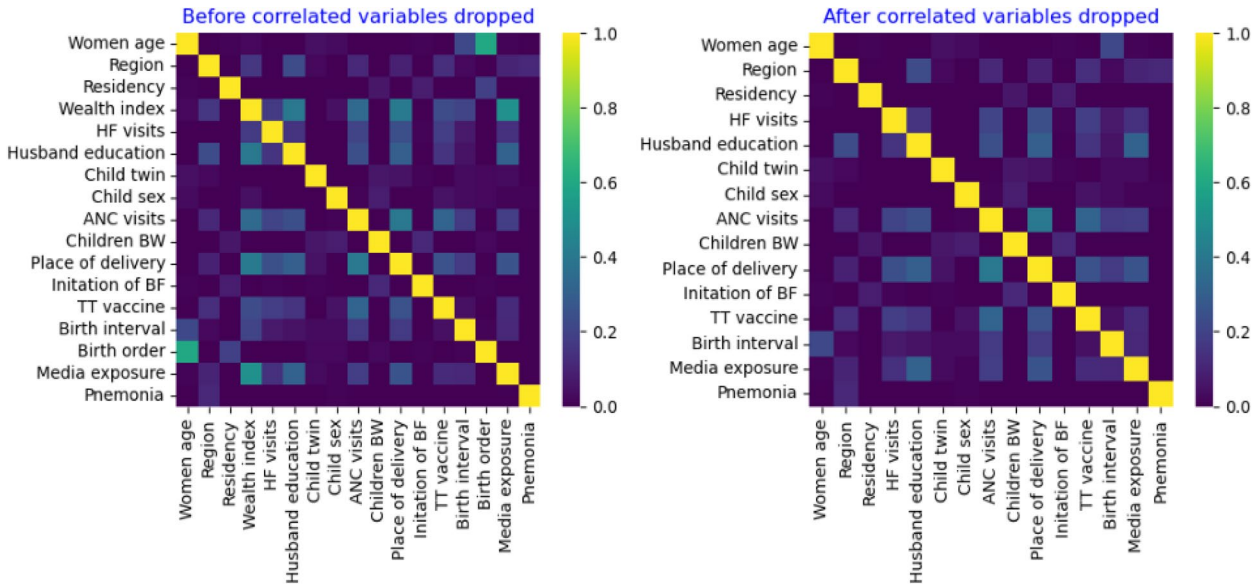

Fig. 3 Correlation detection within the dataset; Where BF = breast feeding, BI = Birth Interval, ANC = Antenatal care, BW = Birth weight

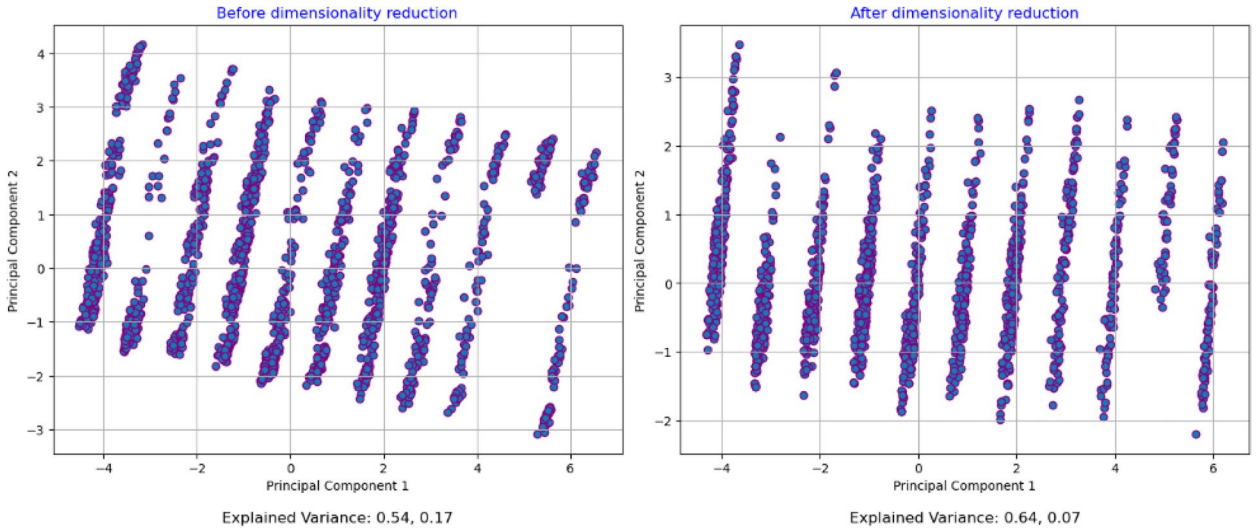

Fig. 4 The principal component analysis before and after dimensionality reduction

and recall with nearly three-fourths (75%). Overall, the model shows 75.6% accuracy in the prediction of pneumonia among children aged 6–23 months, suggesting a reasonably acceptable predictive model for pneumonia case prediction among children aged 6–23 months (Fig. 8).

Discussion

In this study, six machine learning algorithms were built and compared using various performance measurement indicators using a dataset accessed from the 2016 EDHS report. The various predictors such as sociodemographic, maternal, and child health characteristics were included for pneumonia prediction among children aged 6–23 months. A total of 2035 observations were considered for predictive model development. The algorithms were built and compared using these input variables once the dataset was split into training and testing data sets. A K-fold cross-validation technique was applied to enhance the model performance and reduce predicting bias. The k-fold cross-validation method worked with computed average values for enhancing model performance and is

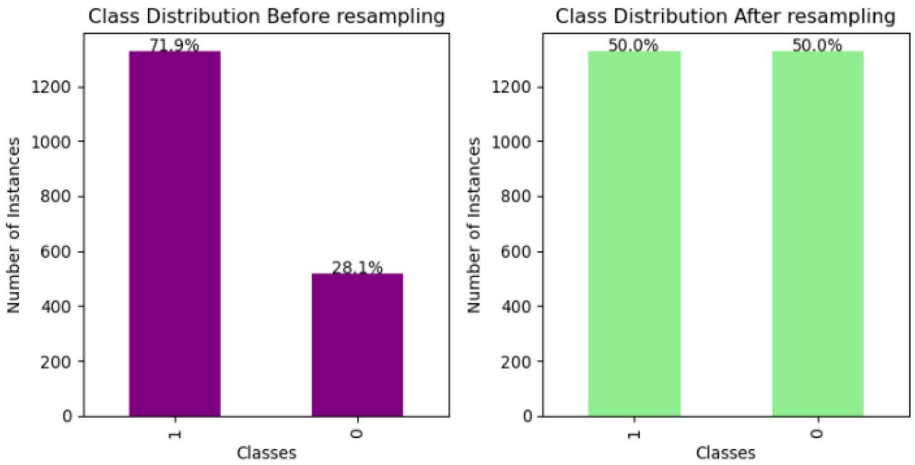

Fig. 5 The data imbalance detection and management using SMOTE

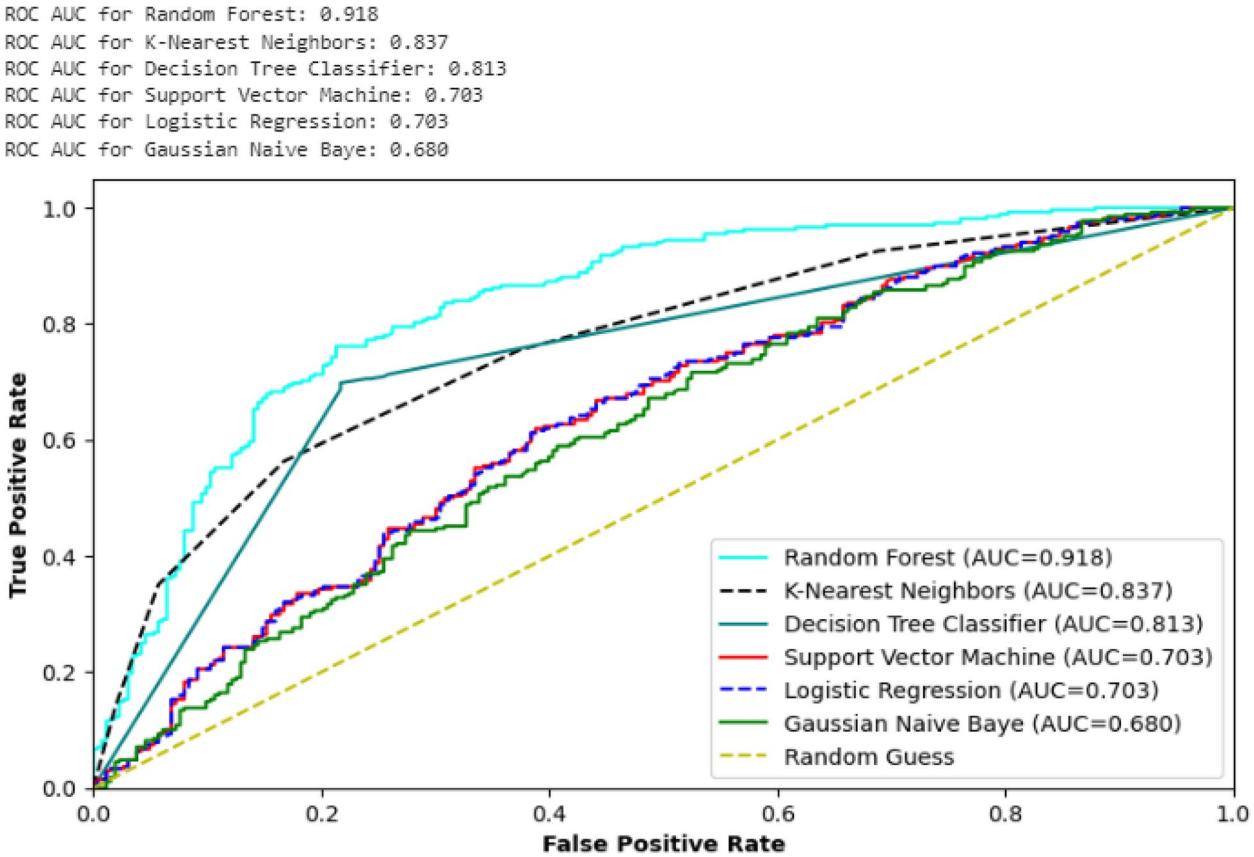

Fig. 6 Data-driven model comparison of included machine learning algorithms

vital to providing an optimized model specifically when there is a limited dataset [37].

Based on the dataset used for this study, a random forest has a relative high-performance score as compared with other algorithms with an AUC value of 91.8%. Additionally, the performance measurement metrics of the random forest algorithm have relatively higher scores than others' algorithm performance metrics. The 83.3% of accuracy, 87.3% of precision, 77.2% of recall, and 93.7% of F1 measures of the random forest algorithm show a

Table 3 Machine learning algorithms' model performance and comparisons with confusion matrix

Algorithms	Accuracy	Precision	Recall	F1-score
Support vector machine	65.9%	68.2%	58.3%	73.7%
Gaussian Naïve Bayes	61.1%	64.3%	49.7%	69.3%
Random forest	83.3%	87.3%	77.2%	93.7%
K-nearest neighbors	76.6%	85.2%	62%	88.2%
Logistic regression	65.7%	67.9%	58.3	73.8%
Decision tree classifier	81.4%	84.3%	76.8%	91.1%

relatively good model to predict pneumonia among children aged 6–23 months. Thus, based on overall performance metrics the random forest algorithm is marked as a top-scoring algorithm for pneumonia prediction and risk factors stratification. The performance of the random forest in this study is consistent with previous machine learning studies such as the prediction of acute respiratory infection [51], prediction of childhood vaccination [50], mental health prediction [61], and prediction of under-five child death [34, 62]. Moreover, the accuracy and ROC value of the random forest algorithm in the current study are consistent with the accuracy and ROC values of a study done about the early detection of cardiovascular disease [37]. Moreover, the performance of the random forest algorithm in this study shows the relative performance as compared with the included algorithms rather than the general context of machine learning algorithms. This is because the performance of the algorithm might be affected by the data nature, sample size, and the characteristics of the predictive model.

In this study, 31.3% of children aged 6–23 months had pneumonia based on fever, cough, and rapid breath symptoms assessment reports. After data normalization and standardization have been done, a 3.2% discrepancy is observed from 31.3% to 28.15%. Additionally, after data imbalance between the classes is detected, a total of 50% of children aged 6–23 months have been confirmed for pneumonia cases. This finding is consistent with studies conducted in various geographical areas such as Ethiopia [3, 18]. The finding is higher than studies done in India [63], Ethiopia [64], and Nigeria [65]. This discrepancy might be due to the study subjects' variation. The nationwide data might create significant variation in the magnitude of pneumonia as compared with single location data. Moreover, the measurement items including the case definition of pneumonia among children might be a reason for the discrepancy [5].

In the optimized predictive model, from a total of 1007 children with pneumonia cases, which is approximately 49.48% of total children, 81.6% of children were correctly predicted as positive for pneumonia with a predictive accuracy of 75.6%. The figure of 81.6% represents the true magnitude of pneumonia among children and highlights a significant public health concern, particularly in LMICs, which underscores the urgent need for effective diagnostic strategies and timely

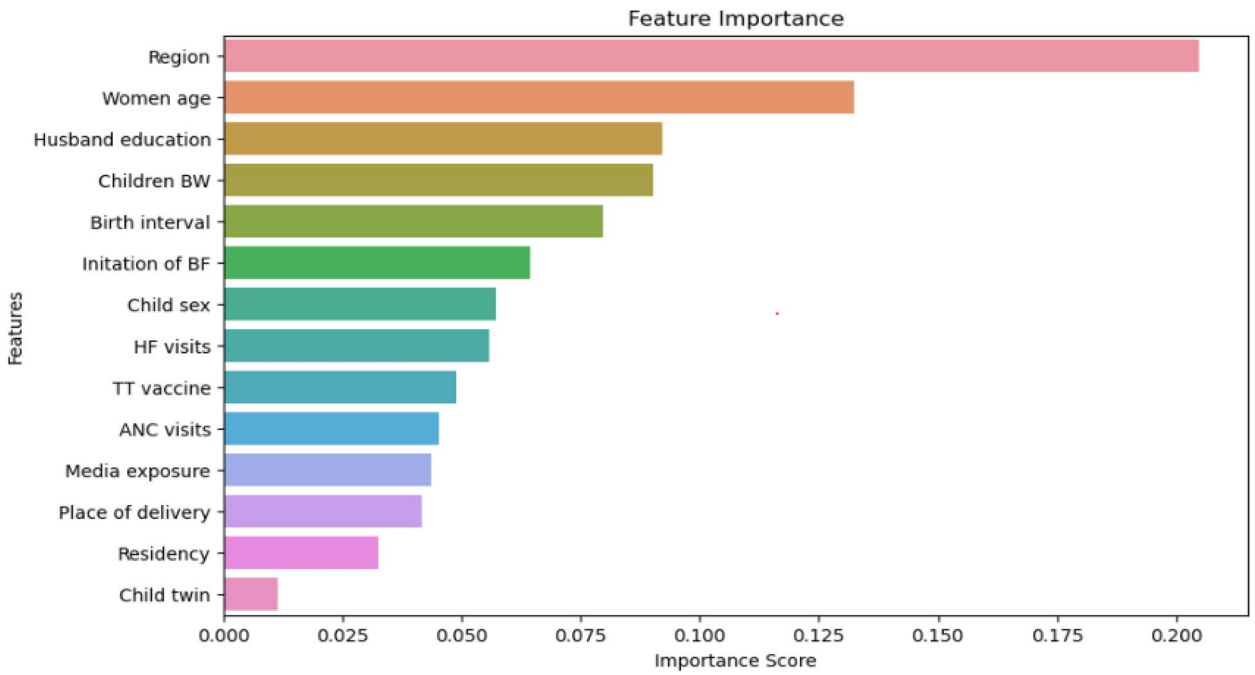

Fig. 7 Important feature stratification based on random forest algorithms

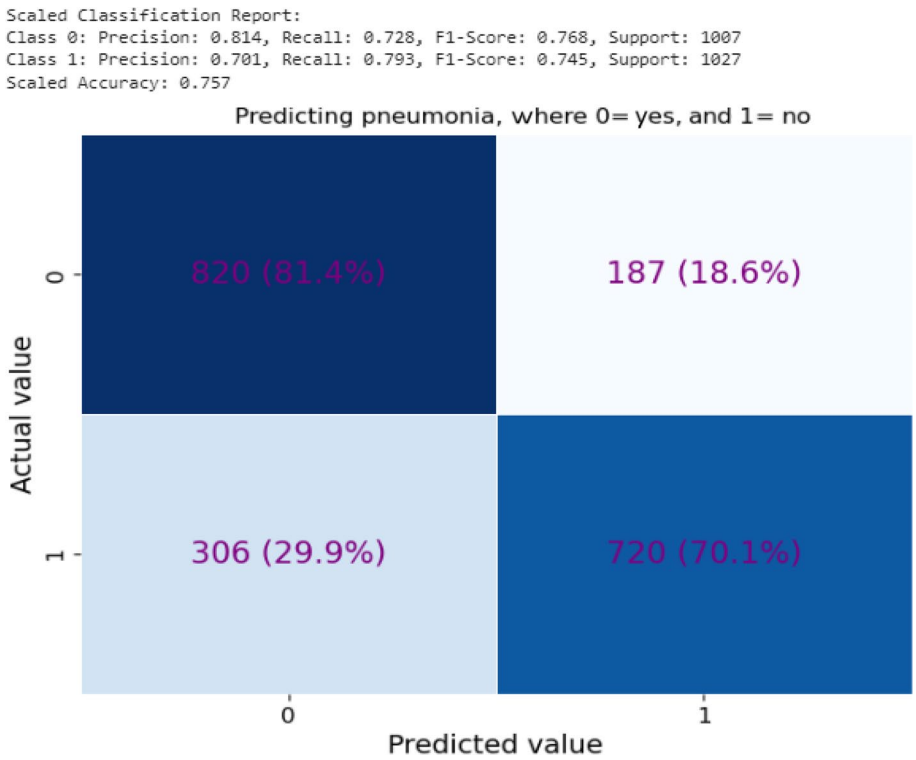

Fig. 8 The actual and predictive values of pneumonia among children aged 6–23 months in Ethiopia, using the 2019 EDHS dataset

interventions [66]. The pneumonia cases become increasing in magnitude which necessarily indicate the preparedness to tackle the problem effectively. The fact that nearly one-fifth of children may not be identified as having pneumonia emphasizes the importance of training healthcare providers to recognize symptoms accurately, investing in improved diagnostic tools, and raising community awareness about pneumonia [67]. Addressing these issues is crucial to ensure that vulnerable populations receive the necessary care for ultimate child health outcome improvements in the region.

Approximately 18.6% and 29.9% of children aged 6-23 months were incorrectly predicted as negative and positive for pneumonia. This shows that there is still room that needs improvement, particularly in reducing false negative and false positive pneumonia cases among children aged 6-23 months. The false negative prediction report may compromise the chance of children to get an effective treatment and medication. Whereas, the false positive also brings resource wastage in pneumonia case diagnosis and treatment, specifically in LMICs where medical resources and budgets are significantly inefficient. Minimizing this false negative and false positive prediction is crucial in pneumonia prediction among children aged 6-23 months using more robust and deep learning algorithms.

From the determinant factors of pneumonia among children aged 6–23 months, the region is identified and stratified as the most important or the first determinant factor for pneumonia cases with a 20% feature important score report. The geographical variation in health service access [68], socioeconomic differences [69], seasonal variations, and population density in regions [70] might be reasons that make region is the most determinant factor for pneumonia among children. Maternal age is a second important determinant factor for pneumonia among children with approximate important score values of 13%. This finding is congruent with primary studies conducted in Indonesia [71], and Ethiopia [72, 73]. This might be due to the mothers' different experiences in child health care, understanding and knowledge they have towards pneumonia and determinant factors, and socioeconomic challenges that can affect the child's immune system development [74]. Maternal age might also influence birth outcomes such as prematurity, and premature infants are at higher risk for respiratory infections and health complications, including pneumonia [75, 76]. Moreover, maternal age influences child health as mothers may face challenges like inadequate nutrition and lower health literacy, leading to increased vulnerability to pneumonia [77]. Husbands' education is a third important predictor for pneumonia among children with

an important score value of 10%. The education level of fathers can significantly impact children's health outcomes, including the risk of pneumonia. Educated fathers can foster better health practices, improve the economy, and encourage proactive healthcare in reducing the risk of respiratory infections including pneumonia in children [78]. The educative husbands reduce the likelihood of exclusive decision-making power in maternal and child health [79].

Childbirth weight is an important predictor of pneumonia among children with an approximated important score value of 9%. Similar to the previous findings, this study identified that birth weight is an important predictor of pneumonia among children aged 6–23 months [80–82]. Circuitously, it could be due to prematurity, a cause of low birth weight, which can predispose the child to respiratory problems due to lung immaturity [83]. Likewise, it could be linked with malnutrition, which causes low birth weight and might lead to pneumonia due to compromised immunity [64, 80, 84].

Birth interval is an important predictor for pneumonia among children with an approximated important score value of 7.7%. This is because birth intervals, specifically short birth and long birth intervals have negative effects on mothers' and children's health in LMICs [85]. The biological, social, or environmental components might have effects on why birth interval could be a predictor for pneumonia [86]. For instance, high birth interval may lead children to a greater risk of pneumonia due to increased exposure to infections, reduced parental resources, and potential delays in healthcare access [83, 87–89]. Breastfeeding initiation is an important predictor for pneumonia among children aged 6–23 months with an important score value of 7.5%. Evidence shows that lack of breastfeeding in the first six months of life is leading exposure to pneumonia in under-five children [90, 91]. This finding is also consistent with the previous studies conducted in Bangladesh [92], and Ethiopia [7, 93]. Timely breastfeeding is a critical determinant in reducing the risk of pneumonia in infants, provides essential nutrients, supports immune function, and minimizes exposure to harmful pathogens [94]. Nonetheless, a study based on 2019 EDHS data reveals that many women experience delayed breastfeeding at the national level [95], and a systematic review report in Ethiopia reveals that breastfeeding initiation is below the World Health Organization's recommendation [96].

The disparity in pneumonia infection by sex of children can be attributed to a combination of biological and social factors, including variations in immune response and differences in exposure to risk factors, which necessitate a comprehensive understanding of how gender influences health outcomes in children [97]. Moreover, research indicates that boys are generally more susceptible to pneumonia than girls, possibly due to inherent immunological differences that affect their vulnerability to infections; this pattern suggests the need for gender-sensitive approaches in health interventions targeted at preventing and managing pneumonia in pediatric populations [83, 98, 99].

Health facility visit within 12 months before the survey is an important predictor for pneumonia among children with an important score value of 7.53%. Within 24 h of birth, the maternal and child health complication is common in LMICs such as Ethiopia [100, 101]. Even though, health facilities are crucial to preventing, early diagnosis and treatment, and management of pneumonia through sharing information and receiving postnatal maternal and child health services receiving, many women might be challenged by finance and distance [68].

The TT vaccine is administered during pregnancy to protect both the mother and the newborn from tetanus, a severe bacterial infection. While the TT vaccine does not directly prevent pneumonia, it plays a role in reducing the risk factors associated with the disease by improving neonatal health, enhancing maternal health, and promoting health-seeking behavior [102].

The relationship between media exposure and pneumonia infection among children is multifaceted [103]. Media serves as a powerful tool for disseminating information about pneumonia, its causes, symptoms, prevention, and treatment [104]. Public health campaigns that utilize mass media platforms; such as television, radio, social media, and print media; can effectively raise awareness about pneumonia and promote preventive measures like vaccination, breastfeeding, and proper hygiene practices [105, 106]. By strategically using media to educate the public, counter misinformation, and reach underserved populations, we can reduce the incidence and impact of pneumonia, ultimately improving health outcomes for all.

Antenatal care serves as a vital opportunity to educate mothers, identify risk factors, and establish a foundation for ongoing care that can significantly reduce the risk of pneumonia in newborns and young children [107]. However, to maximize the benefits of antenatal visits in preventing pneumonia, it is essential to address barriers to access and ensure the quality of care provided [108, 109].

The relationship between media exposure and pneumonia infection among children is multifaceted [103]. Media serves as a powerful tool for disseminating information about pneumonia, its causes, symptoms, prevention, and treatment [104]. Public health campaigns that utilize mass media platforms; such as television, radio, social media, and print media; can effectively raise awareness about pneumonia and promote preventive

measures like vaccination, breastfeeding, and proper hygiene practices [105, 106]. By strategically using media to educate the public, counter misinformation, and reach underserved populations, we can reduce the incidence and impact of pneumonia, ultimately improving health outcomes for all.

The place of delivery is a critical determinant of neonatal and infant health outcomes, including the risk of pneumonia. Deliveries in healthcare facilities are generally associated with better health outcomes due to higher quality care, better infection control, and access to essential newborn care [110]. In contrast, home deliveries, particularly in low-resource settings, can increase the risk of pneumonia due to factors such as poor sanitation, lack of skilled care, and increased exposure to indoor air pollution [83]. Moreover, long-term hospitalization due to birth may also explore pneumonia infection and facilitate its severity among children [111]. Place of residency and child twin are also stratified as important predictors of pneumonia among children aged 6–23 months with important score values of 2.5%, and 1% respectively.

The place of residency is identified as a determinant factor of pneumonia among children aged 6–23 months. Mothers' and caregivers' place of residency often challenges maternal and child health, which is a critical factor in limiting access to healthcare and providing inadequate living conditions, which increases children's vulnerability to respiratory infections [112]. A study in Brazil shows that urban areas, characterized by high levels of air pollution, can exacerbate respiratory issues in children [113]. Conversely, rural areas may lack adequate healthcare facilities, which can delay treatment and worsen health outcomes. Furthermore, overcrowded living conditions can facilitate the spread of infections, making children more susceptible to respiratory illnesses. Generally, residency is key determinant for healthcare service access and delivery [68]. study show that timely access to healthcare is essential for reducing pneumonia-related morbidity and mortality in young children [114].

Being a twin is also a determinant factor of pneumonia in children. Twins often share a close living environment, which increases their exposure to etiologies, and pathogens, which can lead to a higher incidence of infections, including pneumonia [115]. Research has demonstrated that twins are at a greater risk of respiratory illnesses due to shared exposure [112]. Additionally, twins are more likely to experience nutritional deficiencies, particularly if born preterm or with low birth weight. These factors can compromise their immune systems, making them more susceptible to infections like pneumonia [116]. Moreover, the demands of caring for twins can lead to divided parental attention, which may result in less vigilant monitoring of a child's health. This lack of attention can delay the recognition and treatment of pneumonia symptoms and increase the risk of severe illness, like pneumonia among children [117].

Conclusions and recommendations

In this study, six machine learning algorithms were developed and compared for pneumonia prediction, determinant factors identification and stratification. Accordingly, the random forest algorithm was the most accurate algorithm for pneumonia prediction with an 91.8% of AUC value. The magnitude of pneumonia among children aged 6–23 months was 31.3% and 28.1% before and after data normalization. According to the prediction model, pneumonia cases among children is become increasing, which indicate pneumonia is significant public health problem for young children. Based on an important feature selection for pneumonia among children, region, women's age,, husbands' education, birth weight, birth interval, initiation of breastfeeding, child sex, and health facility visits were stratified as the top importance predictors of pneumonia among children aged 6–23 months.

Strengths and limitation

This study is important for generating insights for stakeholders to take clinical and public health problems specifically in pneumonia prevention and interventions. The data-driven predictive model prioritized key determinant factors of pneumonia among children aged 6–23 months, which can support the timely lifestyle modification, and behavioral intervention, facilitate health service delivery and access for pneumonia case prevention. Moreover, the study is crucial for the advancement of research methodology, and add theoretical knowledge through learning in machine learning algorithms for healthcare practices and health data science. As a limitation, important variables of pneumonia might not be included in this study, and the strength of the relationship between variables might be limited based on the cross-sectional nature of the data. Moreover, the verbal reports of mothers of fever, cough, and rapid breaths among children might not confirm the presence of pneumonia. Methodologically, there is also high false negative of pneumonia case among children in the predictive model. Therefore, researchers are strongly recommended to advance the limitation the current study using more advanced deep learning algorithms to enhance the prediction accuracy and minimize false negative prediction probability of the algorithm.

Abbreviations

ANC	Antenatal care
AUC	Area under ROC Curve
DHS	Demographic and Health Survey
EDHS	Ethiopian Demographic and Health Survey
EDHS	Ethiopia Demography and Health Survey
LMICs	Low- and middle-income countries

SNNPR	South Nations and Nationality and Peoples of the region
TT	Tetanus Toxoid
ROC	Receiver operative curve
STATA	Statistical software for data science

Acknowledgements
We would like to express our deepest appreciation to the Measure DHS program for permitting data access and use for this study.

Clinical trial number
Not applicable.

Authors' contributions
AWD was involved in study design, data management, data analysis, interpretation, and discussion of the findings. RA, WG, GWK, MAT, AL and MHK had significant role in drafting and editing the manuscript. FB, SSA, MHJ, GND, LFW, and SP had significantly contribution in the revision of the manuscript. All authors approved the final submission of the revised manuscript for publication.

Funding
No funding was received for this study.

Data availability
The dataset used for analysis is available on the Measure DHS program (http://dhsprogram.com) website. All the data generated, and analyzed are included in this article.

Declarations

Ethics approval and consent to participate
Ethical approval and consent from study participants were not necessary for this study. This is because this study was based on a secondary data source that is publicly available from the Measure DHS program website (https://dhsprogram.com/Date/terms-of-use.cfm).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Debre Berhan University, Asrat Woldeyes Health Science Campus, Public Health Department, Debre Berihan, Ethiopia. [2]Mattu University, Health Science College, Mettu, Ethiopia. [3]Wallaga University, Health Science College, Nekemte, Ethiopia. [4]Madda Walabu University, Health Science College, Shashemene Campus, Shashemene, Ethiopia. [5]Department of Environmental/Ecological Studies and Sustainability, Akamai University, Kamuela, USA. [6]Women Researchers Council, Azerbaijan State University of Economics, Baku, Azerbaijan. [7]Health Science Campus, Health Informatics Department, Wollo University, Wollo, Ethiopia.

Received: 7 September 2024 Accepted: 3 April 2025
Published: 2 May 2025

References

1. Manohar P, et al. Secondary bacterial infections in patients with viral pneumonia. Front Med. 2020;7: 420.
2. Pneumonia, symptom, and treatment: Accessed from https://www.who.int/health-topics/pneumonia/#tab=tab_1.
3. Lema K, et al. Prevalence and associated factors of pneumonia among under-five children at public hospitals in Jimma zone, South West of Ethiopia, 2018. J Pulmonol Clin Res. 2018;2(1):25–31.
4. Nasrin S, et al. Factors associated with community acquired severe pneumonia among under five children in Dhaka, Bangladesh: A case control analysis. PLoS ONE. 2022;17(3): e0265871.
5. Beletew B, et al. Prevalence of pneumonia and its associated factors among under-five children in East Africa: a systematic review and meta-analysis. BMC Pediatr. 2020;20:1–13.
6. Walker CLF, et al. Global burden of childhood pneumonia and diarrhoea. Lancet. 2013;381(9875):1405–16.
7. Solomon Y, et al. Prevalence of pneumonia and its determinant factors among under-five children in Gamo Zone, southern Ethiopia, 2021. Front Pediatr. 2022;10: 1017386.
8. Odeyemi A, et al. Complications of pneumonia and its associated factors in a pediatric population in Osogbo, Nigeria. Nigerian J Paediatr. 2020;47(4):318–23.
9. Wanyana MW, et al. Factors associated with severe pneumonia among children< 5 years, Kasese District, Uganda: a case-control study, January–April 2023. Pneumonia. 2024;16(1):13.
10. Traore Y, et al. Incidence, seasonality, age distribution, and mortality of pneumococcal meningitis in Burkina Faso and Togo. Clin Infect Dis. 2009;48(Supplement_2):S181–9.
11. Deribew A, Tessema F, Girma B. Determinants of under-five mortality in Gilgel gibe field research center, Southwest Ethiopia. Ethiopian J Health Dev. 2007;21(2):117–24.
12. Isturiz RE, Luna CM, Ramirez J. Clinical and economic burden of pneumonia among adults in Latin America. Int J Infect Dis. 2010;14(10):e852–6.
13. Acemoglu D, Johnson S. Disease and development: the effect of life expectancy on economic growth. J Polit Econ. 2007;115(6):925–85.
14. Welte T, Torres A, Nathwani D. Clinical and economic burden of community-acquired pneumonia among adults in Europe. Thorax. 2012;67(1):71–9.
15. Chekole DM, et al. Prevalence and associated risk factors of pneumonia in under five years children using the data of the University of Gondar Referral Hospital. Cogent Public Health. 2022;9(1): 2029245.
16. Fasil A, et al. Epidemiological study of contagious caprine pleuropneumonia (CCPP) in selected districts of Gambella Region, Western Ethiopia. Afr J Agric Res. 2015;10(24):2470–9.
17. Amare RA, et al. Incidence of recovery from severe pneumonia and its predictors among children 2–59 months admitted to pediatric ward of Ayder Comprehensive Specialized Hospital, Tigray, Ethiopia: a retrospective cohort study. J Family Med Primary Care. 2022;11(9):5285–92.
18. Abuka T. Prevalence of pneumonia and factors associated among children 2–59 months old in Wondo Genet district, Sidama zone, SNNPR, Ethiopia. Curr Pediatr Res. 2017;21(1):19–25.
19. Demsash AW, et al. Spatial and multilevel analysis of sanitation service access and related factors among households in Ethiopia: using 2019 Ethiopian national dataset. PLOS Global Public Health. 2023;3(4): e0001752.
20. Demsash AW, Emanu MD, Walle AD. Exploring spatial patterns, and identifying factors associated with insufficient cash or food received from a productive safety net program among eligible households in Ethiopia: a spatial and multilevel analysis as an input for international food aid programmers. BMC Public Health. 2023;23(1):1141.
21. Roomaney RA, et al. Epidemiology of lower respiratory infection and pneumonia in South Africa (1997–2015): a systematic review protocol. BMJ Open. 2016;6(9): e012154.
22. Ramezani M, Aemmi SZ, Emami Moghadam Z. Factors affecting the rate of pediatric pneumonia in developing countries: a review and literature study. Int J Pediatr. 2015;3(6.2):1173–81.
23. Budge S, Ambelu A, Bartram J, Brown J, Hutchings P. Environmental sanitation and the evolution of water, sanitation and hygiene. Bull World Health Organ. 2022;100(4):286–8. https://doi.org/10.2471/BLT.21.287137. Epub 2022 Mar 3.
24. Chen C, et al. Prenatal and postnatal risk factors for infantile pneumonia in a representative birth cohort. Epidemiol Infect. 2012;140(7):1277–85.
25. Aslam A, et al. The state of the world's children 2014 in numbers: every child counts. revealing disparities, advancing children's rights. ERIC; 2014. Accessed from: https://www.unicef.org/media/89221/file/SOWC%202014.pdf.
26. Rawat A, et al. Health system considerations for community-based implementation of automated respiratory counters to identify childhood pneumonia in 5 regions of Ethiopia: a qualitative study. Int J Health Policy Manag. 2023;12:7385.

27. Madhi SA, et al. Vaccines to prevent pneumonia and improve child survival. Bull World Health Organ. 2008;86:365–72.
28. Kiguli S, et al. Nutritional supplementation in children with severe pneumonia in Uganda and Kenya (COAST-Nutrition): a phase 2 randomised controlled trial. Eclinicalmedicine. 2024;72:102640.
29. Enebeli U, Amadi A, Iro O. The Association between water, sanitation and hygiene practices and the occurrence of childhood pneumonia in Abia State, Nigeria. Int J Res Sci Innov. 2019;6:55–60.
30. Scott JAG, et al. The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. Clin Infect Dis. 2012;54(suppl_2):S109–16.
31. Sun X, Douiri A, Gulliford M. Applying machine learning algorithms to electronic health records to predict pneumonia after respiratory tract infection. J Clin Epidemiol. 2022;145:154–63.
32. Swetha K, et al. Prediction of pneumonia using big data, deep learning and machine learning techniques. In: 2021 6th International Conference on Communication and Electronics Systems (ICCES). IEEE; 2021. https://doi.org/10.1109/ICCES51350.2021.9489188.
33. Alowais SA, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689.
34. Demsash AW. Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday. Inform Med Unlocked. 2023;40:101298. https://doi.org/10.1016/j.imu.2023.101298.
35. Demsash AW, et al. Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset. PLoS ONE. 2023;18(10): e0288867.
36. Petersson L, et al. Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. BMC Health Serv Res. 2022;22(1):850.
37. Baghdadi NA, et al. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data. 2023;10(1):144.
38. The 2016 Ethiopian demography and health survey. 2016. Accessed from https://www.dhsprogram.com/pubs/pdf/FR328/FR328.pdf.
39. Atoloye K, et al. A spatio-temporal mapping and bayesian modelling of risk factors of pneumonia symptoms in under-five children in Nigeria. medRxiv, 2022: p. 2022.12. 19.22283675.
40. Wakeyo MM, et al. Short birth interval and its associated factors among multiparous women in Mieso agro-pastoralist district, Eastern Ethiopia: A community-based cross-sectional study. Front Glob Womens Health. 2022;3: 801394.
41. Kassie SY, et al. Spatial distribution of short birth interval and associated factors among reproductive age women in Ethiopia: spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey. BMC Pregnancy Childbirth. 2023;23(1):1–14.
42. Demsash AW, et al. Spatial distribution of vitamin A rich foods intake and associated factors among children aged 6–23 months in Ethiopia: spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey. BMC Nutr. 2022;8(1):1–14.
43. Muhwava LS, Morojele N, London L. Psychosocial factors associated with early initiation and frequency of antenatal care (ANC) visits in a rural and urban setting in South Africa: a cross-sectional survey. BMC Pregnancy Childbirth. 2016;16(1):1–9.
44. Demsash AW, et al. Spatial distribution of vitamin A rich foods intake and associated factors among children aged 6–23 months in Ethiopia: spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey. BMC nutrition. 2022;8(1):77.
45. Kolluri J, et al. Reducing overfitting problem in machine learning using novel L1/4 regularization method. In: 2020 4th international conference on trends in electronics and informatics (ICOEI)(48184). IEEE; 2020. https://doi.org/10.1109/ICOEI48184.2020.9142992.
46. Khan NM, et al. Analysis on improving the performance of machine learning models using feature selection technique. In: Intelligent systems design and applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6–8, 2018, vol. 2. Springer; 2020. https://doi.org/10.1007/978-3-030-16660-1_7.
47. Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient. Appl Intell. 2022;52(5):5457–74.
48. Chumachenko T, Bazilevych K. Dimensionality reduction of chronic kidney disease data using principal component analysis. 2023.
49. Demsash AW. Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday. Inform Med Unlocked. 2023:101298. https://doi.org/10.1016/j.imu.2023.101298.
50. Demsash AW, et al. Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset. PLoS ONE. 2023;18(10): e0288867.
51. Kalayou MH, Kassaw AAK, Shiferaw KB. Empowering child health: Harnessing machine learning to predict acute respiratory infections in Ethiopian under-fives using demographic and health survey insights. BMC Infect Dis. 2024;24(1):338.
52. Géron A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.; 2022. Accessed from: https://books.google.com.et/books?id=HHetDwAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.
53. Yu H-F, Huang F-L, Lin C-J. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach Learn. 2011;85:41–75.
54. James G. An introduction to statistical learning. Springer; 2013. https://doi.org/10.1007/978-1-0716-1418-1.
55. Fenta HM, et al. Factors of acute respiratory infection among under-five children across sub-Saharan African countries using machine learning approaches. Sci Rep. 2024;14(1):15801.
56. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann transl Med. 2016;4(11):218.
57. Narkhede S. Understanding auc-roc curve. Towards Data Sci. 2018;26(1):220–7.
58. El Khouli RH, et al. Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. J Magnetic Resonance Imaging. 2009;30(5):999–1004.
59. Team A. Accuracy vs. precision vs. recall in machine learning: what's the difference? 2024. Accessed from https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall.
60. Vakili M, Ghamsari M, Rezaei M. Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv preprint arXiv:2001.09636, 2020. https://doi.org/10.48550/arXiv.2001.09636.
61. Rahman MA, Kohli T. Mental health analysis of international students using machine learning techniques. PLoS ONE. 2024;19(6):e0304132.
62. Bitew FH, et al. Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. Genus. 2020;76:1–16.
63. Nirmolia N, et al. Prevalence and risk factors of pneumonia in under five children living in slums of Dibrugarh town. Clin Epidemiol Global Health. 2018;6(1):1–4.
64. Mengstie LA. Prevalence of pneumonia and associated factors among children aged 6–59 months in Angolela Tera district, North Shoa, Ethiopia, 2021, community-based cross-sectional study. Bull National Res Centre. 2022;46(1):231.
65. Igweonu-Nwakile C, et al. Prevalence of Pneumonia and Its Determinants among Under-five Children attending a Primary Health Care Clinic in Amuwo Odofin Local Government Area, Lagos, Nigeria. J Community Med Primary Health Care. 2023;35(1):40–9.
66. Chanie MG, et al. Predictors of community acquired childhood pneumonia among 2–59 months old children in the Amhara Region, Ethiopia. BMC Pulm Med. 2021;21(1):179.
67. Fox MP, et al. Low rates of treatment failure in children aged 2–59 months treated for severe pneumonia: a multisite pooled analysis. Clin Infect Dis. 2013;56(7):978–87.
68. Demsash AW, Walle AD. Women's health service access and associated factors in Ethiopia: application of geographical information system and multilevel analysis. BMJ Health Care Inform. 2023;30(1):e100720.
69. Hossain MZ, et al. Weather Variability, Socioeconomic Factors, and Pneumonia in Children Under Five-Years Old—Bangladesh, 2012–2016. China CDC Weekly. 2021;3(29):620.

70. Lee E, et al. Annual and seasonal patterns in etiologies of pediatric community-acquired pneumonia due to respiratory viruses and Mycoplasma pneumoniae requiring hospitalization in South Korea. BMC Infect Dis. 2020;20:1–10.
71. Umar KF, et al. Risk Factor of Paediatric Community-Acquired Pneumonia in Wajo Regency, Indonesia. National J Commun Med. 2024;15(02):98–104.
72. Bazie GW, Seid N, Admassu B. Determinants of community acquired pneumonia among 2 to 59 months of age children in Northeast Ethiopia: a case-control study. Pneumonia. 2020;12:1–10.
73. Endale A, et al. Determinants of community-acquired pneumonia among 2-59 months old children attending health facility in Hossaena Town, Ethiopia. 2022.
74. Aftab W, et al. Exploring health care seeking knowledge, perceptions and practices for childhood diarrhea and pneumonia and their context in a rural Pakistani community. BMC Health Serv Res. 2018;18:1–10.
75. Willson DF, et al. Complications in infants hospitalized for bronchiolitis or respiratory syncytial virus pneumonia. J Pediatr. 2003;143(5):142–9.
76. Apisarnthanarak A, et al. Ventilator-associated pneumonia in extremely preterm neonates in a neonatal intensive care unit: characteristics, risk factors, and outcomes. Pediatrics. 2003;112(6):1283–9.
77. de Buhr E, Tannen A. Parental health literacy and health knowledge, behaviours and outcomes in children: a cross-sectional survey. BMC Public Health. 2020;20:1–9.
78. Kajungu D, et al. Factors associated with caretakers' knowledge, attitude, and practices in the management of pneumonia for children aged five years and below in rural Uganda. BMC Health Serv Res. 2023;23(1):700.
79. Bakare AA, et al. Community and caregivers' perceptions of pneumonia and care-seeking experiences in Nigeria: A qualitative study. Pediatr Pulmonol. 2020;55:S104–12.
80. Hadisuwarno W, Setyoningrum RA, Umiastuti P. Host factors related to pneumonia in children under 5 years of age. Paediatr Indones. 2015;55(5):248–51.
81. Karmany PA, Rahardjo SS, Murti B. Effect of Low Birth Weight on the Risk of Pneumonia in Children Under Five: Meta-Analysis. Int Confer Public Health Proceed. 2020;5(1):106.
82. Sutriana VN, Sitaresmi MN, Wahab A. Risk factors for childhood pneumonia: a case-control study in a high prevalence area in Indonesia. Clinical and experimental pediatrics. 2021;64(11):588.
83. Fadl N, Ashour A, Yousry Muhammad Y. Pneumonia among under-five children in Alexandria, Egypt: a case-control study. J Egypt Public Health Assoc. 2020;95:1–7.
84. Kiconco G, et al. Prevalence and associated factors of pneumonia among under-fives with acute respiratory symptoms: a cross sectional study at a Teaching Hospital in Bushenyi District, Western Uganda. African Health Sciences. 2021;21(4):1701–10.
85. Kassie SY, et al. Spatial distribution of short birth interval and associated factors among reproductive age women in Ethiopia: spatial and multilevel analysis of 2019 Ethiopian mini demographic and health survey. BMC Pregnancy Childbirth. 2023;23(1):275.
86. Björkegren E, Svaleryd H. Birth order and health disparities throughout the life course. Soc Sci Med. 2023;318:115605.
87. Debere HR, Adjiwanou V. The effects of reproductive variables on child mortality in Ethiopia: evidence from demographic and health surveys from 2000 to 2016. Reproductive Health. 2024;21(1):4.
88. Nurmala I, Kurniawan F. Relationship between gravidity and low birth weight in Kendari City hospital. Indonesian J Contemp Multidiscip Res. 2023;2(3):445–64.
89. Geleta D, Tessema F, Ewnetu H. Determinants of community acquired pneumonia among children in Kersa District, Southwest Ethiopia: facility based case control study. J Pediatr Neonatal Care. 2016;5(2):00179.
90. Dina RA, Djuwita R. The role of exclusive breastfeeding in reducing pneumonia prevalence in children under five. J Gizi Pangan. 2021;16(28):89–98.
91. Karmany PAW, Rahardjo SS, Murti B. The effects of non-exclusive breastfeeding on the risk of pneumonia in children under five: Meta-analysis. Journal of Epidemiology and Public Health. 2020;5(4):393–401.
92. Sakib MS, Ripon Rouf ASM, Tanny TF. Determinants of early initiation of breastfeeding practices of newborns in bangladesh: evidence from bangladesh demographic and health survey. Nutrition and Metabolic Insights. 2021;14:11786388211054676.
93. Gedefaw M, Berhe R. Determinates of childhood pneumonia and diarrhea with special emphasis to exclusive breastfeeding in north Achefer district, northwest Ethiopia: a case control study. Open Journal of Epidemiology. 2015;5(02):107–12.
94. Yadate O, et al. Determinants of pneumonia among under-five children in Oromia region, Ethiopia: unmatched case-control study. Archives of Public Health. 2023;81(1):87.
95. Haile RN, Abate BB, Kitaw TA. Spatial variation and determinants of delayed breastfeeding initiation in Ethiopia: spatial and multilevel analysis of recent evidence from EDHS 2019. Int Breastfeed J. 2024;19(1):10.
96. Alebel A, et al. Timely initiation of breastfeeding and its association with birth place in Ethiopia: a systematic review and meta-analysis. Int Breastfeed J. 2017;12:1–9.
97. Lembang ND, P.R. The Relationship between nutritional status and shortness of breath in children with pneumonia at Mappi General Hospital, Indonesia. Open Access Indonesian J Med Rev. 2023;3(4):437–40. https://doi.org/10.1080/24694193.2019.1578435.
98. Rajaraman S, Guo P, Xue Z, Antani SK. A deep modality-specific ensemble for improving pneumonia detection in chest x-rays. Diagnostics. 2022;12(6):1442.
99. Kyu HH, Vongpradith A, Sirota SB, Novotney A, Troeger CE, Doxey MC, Bender RG, Ledesma JR, Biehl MH, Albertson SB, Frostad JJ. Age–sex differences in the global burden of lower respiratory infections and risk factors, 1990–2019: results from the Global Burden of Disease Study 2019. Lancet Infect Dis. 2022;22(11):1626–47.
100. Asmamaw DB, et al. Early Postnatal Home Visit Coverage by Health Extension Workers and Associated Factors Among Postpartum Women in Gidan District, Northeast Ethiopia. Int J Public Health. 2023;68: 1605203.
101. Demsash AW, et al. Birth preparedness and pregnancy complication readiness and associated factors among pregnant women in Ethiopia: A multilevel analysis. PLOS Global Public Health. 2024;4(5): e0003127.
102. Nguyen TK, Tran TH, Roberts CL, Fox GJ, Graham SM, Marais BJ. Risk factors for child pneumonia-focus on the Western Pacific Region. Paediatric respiratory reviews. 2017;21:95–101.
103. Ooko SA. Media framing of infectious diseases in Kenya: a case study of pneumonia (Doctoral dissertation, University of Nairobi). 2014.
104. Kumar S, Mohanraj R, Dhingra B, Agarwal M, Suresh S. Optimizing care-seeking for childhood pneumonia: a public health perspective. Indian Pediatrics. 2021;58:1030–5.
105. Rodrigues F, Ziade N, Jatuworapruk K, Caballero-Uribe CV, Khursheed T, Gupta L. The Impact of Social Media on Vaccination: A Narrative Review. Journal of Korean Medical Science. 2023;38(40):e326.
106. Naugle DA, Hornik RC. Systematic review of the effectiveness of mass media interventions for child survival in low-and middle-income countries. J Health Commun. 2014;19(sup1):190–215.
107. Shiferaw K, Mengiste B, Gobena T, Dheresa M. The effect of antenatal care on perinatal outcomes in Ethiopia: A systematic review and meta-analysis. PloS one. 2021;16(1):e0245003.
108. Lassi ZS, Padhani ZA, Rabbani A, Rind F, Salam RA, Das JK, Bhutta ZA. Impact of dietary interventions during pregnancy on maternal, neonatal, and child outcomes in low-and middle-income countries. Nutrients. 2020;12(2):531.
109. Newnham JP, Dickinson JE, Hart RJ, Pennell CE, Arrese CA, Keelan JA. Strategies to prevent preterm birth. Frontiers in immunology. 2014;5:584.
110. Workineh Y, Hailu D, Gultie T. Determinants of pneumonia among under two children in southern Ethiopia: A case control study 2016. Curr Pediatr Res. 2017;21(4):604–12.
111. Le Roux DM, et al. Factors associated with serious outcomes of pneumonia among children in a birth cohort in South Africa. PLoS ONE. 2021;16(8):e0255790.
112. Baseer KAA, Mohamed M, Abd-Elmawgood EA. Risk factors of respiratory diseases among neonates in neonatal intensive care unit of Qena University Hospital, Egypt. Annals of global health. 2020;86(1):22.
113. Vieira SE, et al. Urban air pollutants are significant risk factors for asthma and pneumonia in children: the influence of location on the measurement of pollutants. Archivos de Bronconeumologia (English Edition). 2012;48(11):389–95.
114. Källander K, Young M, Qazi S. Universal access to pneumonia prevention and care: a call for action. Lancet Respir Med. 2014;2(12):950–2.

115. Wonodi CB, et al. Evaluation of risk factors for severe pneumonia in children: the Pneumonia Etiology Research for Child Health study. Clinical infectious diseases. 2012;54(suppl_2):S124–31.
116. Demsash AW, Asefa EY, Bekana T. Mothers' experience of losing infants by death and its predictors in Ethiopia. PLoS ONE. 2024;19(6):e0303358. https://doi.org/10.1371/journal.pone.0303358.
117. Sutcliffe AG, Derom C. Follow-up of twins: health, behaviour, speech, language outcomes and implications for parents. Early Human Dev. 2006;82(6):379–86.

Publisher's Note