# Analysis and Prediction of Childhood Pneumonia Deaths using Machine Learning Algorithms

Felipe A. L. Soares, Efrem E. O. Lousada, Tiago B. Silveira, Raquel A. F. Mini, Luis E. Zárate, Henrique C. Freitas

Graduate Program in Informatics, Pontifícia Universidade Católica de Minas Gerais, Belo Horizonte, Brazil
falsoares@sga.pucminas.br, elousada@gmail.com, tbsilveira@sga.pucminas.br, {raquelmini, zarate, cota}@pucminas.br

**Abstract.** Acute Respiratory Tract Infections are among the leading causes of child mortality worldwide. Specifically, community-acquired pneumonia has different causes, such as: passive smoking, air pollution, poor hygiene, cardiac insufficiency, oropharyngeal colonization, nutritional deficiency, immunosuppression, and environmental, economic and social factors. Due to the variation of these causes, knowledge discovery in this area of health has been a great challenge for researchers. Thus, this paper presents the steps for the construction of a database and evaluation results applied to the analysis and prediction of potential deaths caused by childhood pneumonia using the Pictorea method. For this, the Random Forest and Artificial Neural Network algorithms were used, and after comparison, the Neural Network algorithm showed higher accuracy by up to 87.57%. This algorithm was used to analyze and predict the number of deaths from pneumonia in children up to 5 years old, and the results were presented using Root Mean Square Error and scatter plots. A domain specialist validated the results and defined that the pattern found is relevant for future studies in the medical field, helping to analyze the behavior of countries and predict future scenarios.

CCS Concepts: • **Computing methodologies** → **Machine learning**.

Keywords: Artificial neural network, Pneumonia, Data analysis and prediction, Potential deaths, Random forest.

## 1. INTRODUCTION

Childhood pneumonia is considered the leading cause of child deaths worldwide [Laiakis et al. 2010]. According to the World Health Organization, these diseases account for approximately 14% of all deaths, ranging from 31 and 91 deaths per 100,000 inhabitants in high-income and low-income per capita countries, respectively [Organization 2014]. Acute Respiratory Infections (ARI) are more common in children and elderly, as these age groups are vulnerable to complications. In [Liu et al. 2017], the authors concluded that ARI is among the leading causes of child mortality in the year 2013. More than 3 million of children under 5 years old died from infectious diseases worldwide, of which 14.9% were primarily caused by pneumonia.

Community-acquired pneumonia in children has several cause factors, such as: passive smoking, air pollution, poor hygiene, cardiac insufficiency, oropharyngeal colonization, nutritional deficiency, immunosuppression, and environmental, economic and social factors [Scotta et al. 2019]. Understanding how these factors really influence the onset of ARI, specifically pneumonia, has caught the attention of researchers [Chaves et al. 2017; Afifi et al. 2017].

Several sources (e.g., World Health Organization[1], World Bank Open Data[2] and Institute for Health Metrics and Evaluation[3]) provide socioeconomic data on countries that are valuable sources for data analysis. However, this wide database can be a problem in terms of processing and knowledge discovery. Different variables from different sources must be correlated towards useful information for

[1]https://www.who.int/en. Accessed on July 10, 2021
[2]https://data.worldbank.org. Accessed on July 10, 2021
[3]http://www.healthdata.org. Accessed on July 10, 2021

public actions. How to provide this useful information is our main issue.

As stated by [Yang et al. 2015], the healthcare sector generates huge amounts of data that can be exploited to find hidden patterns and knowledge. These can be used in decision-making in diagnosis, treatment, disease prevention, and in creating public policies to define countries' behavior to combat diseases. According to [Jothi et al. 2015], data mining and machine learning algorithms are being used in healthcare, primarily for disease prevention and to assist diagnostics.

Given this context, this paper proposes the use of the knowledge discovery process, based on Pictorea methodology proposed by [Montevecchi and Zárate 2014], to analyze and predict childhood pneumonia deaths in different regions as well as the various risk factors associated with it, such as health expenses, total population, demographic density, vaccinated population, population using sanitation facilities, smokers and others. To that end, two machine learning techniques were used, a Random Forest (RF) algorithm and an Artificial Neural Network (ANN).

This paper has three main contributions: i) a relevant data dictionary built alongside a domain specialist based on different factors correlated to pneumonia; ii) an approach using machine learning algorithms for analyzing and predicting potential childhood pneumonia deaths with high accuracy; and iii) a consistent database as baseline for training and testing the algorithms. All developed database and scripts are publicly available in our research repository[4].

This paper is organized as follows: Section 2 presents the related work about data mining, machine learning and knowledge discovery in the health field. Section 3 presents the description and construction of the database used to apply the method. In Section 4, the experiments and results are presented and discussed. Finally, Section 5 presents the conclusions and future works.

## 2. RELATED WORK

The literature presents many research works with focus on data mining and machine learning techniques applied to different health fields. Specifically, regarding community-acquired pneumonia, methods have been applied to estimate the risk of hospitalization or death. Statistical techniques as logistic regression are presented in [Duan et al. 2016]. The work developed in [Chaves et al. 2017] uses fuzzy logic to build a computational model that estimates the influence of exposure to air pollutants on the number of hospitalizations for pneumonia. [Caruana et al. 2015] used GA$^2$M in a dataset to assess pneumonia risk and presented the results in a hospital readmission case study. [Afifi et al. 2017] presented a new prediction method to aid algorithms in early detection of the risk of pneumonia, having a prediction accuracy of 73%.

There are works that used machine learning and knowledge discovery aimed at childhood pneumonia. [Laiakis et al. 2010] explored metabolomic analysis to improve the diagnosis of pneumonia. For this, the authors used SIMCA-P$^+$ and Random Forests, and the unsupervised data were analyzed using Principal Component Analysis (PCA), while the supervised data were analyzed using Partial Least Squares-Discriminant Analysis (PLS-DA) and Orthogonal Projection to Latent Structures (OPLS). The results obtained showed that the metabolomic analysis satisfactorily distinguished patients with severe pneumonia.

[Naydenova et al. 2015] used machine learning techniques to process various clinical measurements in order to diagnose. Machine learning techniques, such as Support Vector Machines and Random Forests, were used to develop the predictive algorithm based on four characteristics: temperature, respiratory rate, heart rate and oxygen saturation. The results were evaluated in a data set with 1093 children, with a sensitivity of 96.6%, specificity of 96.4% and an Area Under the Curve of 97.8%.

Research works [Alimadadi et al. 2020] about COVID-19 have used artificial intelligence and machine

[4]https://github.com/cart-pucminas/pneumonia. Accessed on July 10, 2021

learning to provide databases [COVID-19 2020], which can be used in the future to correlate, prevent and avoid childhood deaths by pneumonia. For instance, a technique presented by [Apostolopoulos and Mpesiana 2020] uses X-ray images utilizing transfer learning with convolutional neural networks to detect biomarkers related to COVID-19 disease. The authors achieved accuracy, sensitivity and specificity by up to 96.78%, 98.66% and 96.46%, respectively.

According to our review, even though there are works that use prediction in the context of childhood pneumonia, the literature does not present specific works using machine learning algorithms based on socioeconomic factors in a given region applied to the analysis and prediction of the total deaths of children from pneumonia.

## 3. RESEARCH METHODOLOGY

The database about pneumonia deaths from 186 countries for 2000, 2012 and 2013 can be accessed at *Kaggle*[5]. This database provides the following data: country; health expenditure (HE); percentage of vaccinated population (PPV); price of *PCV10* and *PCV13* pneumonia vaccines (PPCV10 and PPCV13); how the pneumonia vaccine is acquired (HVA); gross national income for the years 2012 and 2013 (GNI); percentage of population using sanitation facilities for 2012 and 2013 (PSF); birth rate per 1,000 inhabitants for the years 2012 and 2013 (BR); and death rate of children under 5 years old in 2013 (DRC). Thus, the database has a total of 186 instances and 9 attributes.

The Knowledge Discovery in Databases (KDD) process has a sequence of steps that comprises the entire data cycle until extracted knowledge becomes useful and not obvious. According to [Montevecchi and Zárate 2014], the Pictorea KDD process is a pedagogical and canonical method that has thirteen steps: 1 - Problem space exploration, 2 - Solution space definition, 3 - Understanding the problem domain, 4 - Characterization of the problem by attributes, 5 - Database creation, 6 - Data exploration, 7 - Data pre-processing, 8 - Dimensionality reduction and sample selection, 9 - Data transformation, 10 - Data mining, 11 - Standards discovery, 12 - Statistical validation, 13 - Viewing. This steps were developed following the generic proposed in [Fayyad et al. 1996], for the development, monitoring and documentation of stages and activities of database knowledge discovery projects, as it can be developed by less experienced professionals.

During these thirteen steps, a domain specialist only does not participate in five (5, 6, 8, 10 and 12). In all others, she/he participates in tune with KDD analyst. The participation of a domain specialist is important to obtain results that are in line with the objectives established by the KDD process, as well as, to helping to define data needed for analysis, assigning weights to potential problems and analyzing the results found [Montevecchi and Zárate 2014].

**Step 1** - Prior to the steps responsible for building the base and manipulating the data, Pictorea suggests steps related to the study of the problem and the suggestion of the techniques that will be applied. These steps are performed with a domain specialist, whose area of expertise (in this work, Pediatric Medicine) is directly related to the proposed problem. After problem definition, expectations about the outcome are defined. In this work, the expectation is to propose a prediction that from economic and social factors of a country can present the approximate number of deaths in children up to 5 years old caused by pneumonia where, the value is expressed in units of death per 1,000 children born in the year.

**Step 2** - The Pictorea method proposes the prior definition of the machine learning task that will be applied to the proposed problem. In this work, the machine learning algorithms chosen are Random Forest (RF) and Artificial Neural Network (ANN). The RF produces multiple decision trees using a randomly selected subset of samples and features training [Breiman 2001]. Thus, it combines these decision tree-based predictors to perform the prediction. ANN, on the other hand, uses models

[5]https://www.kaggle.com/c/pneumonia-child-mortality-data624-16a. Accessed on May 8, 2020

inspired by the processing and learning capacity of the human brain with generalization capacity based on the knowledge learned, presenting adequate outputs for inputs not observed during training process [Kraft et al. 2003].

**Step 3** - In the third step of Pictorea (regarding understanding the problem domain), a domain specialist analyzes the available data sources for better understanding of the problem. For this work, new variables, such as percentage of male smokers (PMS) and percentage of female smokers (PFS), percentage of people with Human Immunodeficiency Virus (PPHIV), demographic density (DD), total population (TP) and life expectancy (LE) are identified to enrich the database. According to a domain specialist, these are features that may influence the death rate due to pneumonia. Finally, a domain specialist suggested splitting the death rate of children under 5 years old (DRC) attribute (output) into 2 new attributes to be explored in the predictions: number of pneumonia deaths in newborns from 0 to 27 days old (NDN) and number of pneumonia deaths in children up to 5 years old (NDC). Therefore, more specific results can be obtained for different age groups.

**Step 4** - The fourth step corresponds to the characterization of the problem through attributes. The goal is to identify relevant features for the considered problem domain. For this, we consider the tacit knowledge of a domain specialist in order to find out which data were really relevant and which were disposable. Subsequently, a database is built for the project.

**Step 5** - First, it is necessary ensure integrity of the database, because in the same instance there are data from three different years (2000, 2012, 2013). The year became a new attribute of the database, reorganizing the data to the years for each country. Thus, it was necessary to search the country data in each of these three years to compose the database. Each of the 186 instances is fragmented into 3 new ones, tripling the value of records to 558. In the end, the dataset has data from 3 years, 186 countries, 558 instances and 16 attributes. After composing the database considered in this work, the absence of data for some countries was observed. Among these missing data we mention: sanitation facilities, gross national income and birth rate. In addition, the new attributes defined as output for the predictions need to be searched and inserted. External data sources [6,7] were searched to recover the required data. The attributes related to the annual percentage by country of childhood deaths by pneumonia, percentage by year and by country of the population using sanitation facilities, gross national income, and vaccinated population, were found and inserted into the database.

In order to obtain the final database, the new attributes defined in the third step were inserted into the data. Thus, the enriched final database has the attributes: country, year, Health expenditure (HE), Population vaccinated against pneumonia (%) (PPV), If the pneumonia vaccine is on the regular vaccination schedule (RVC), Gross National Income per capita (GNI), Population using basic sanitation facilities (%) (PBSF), Population using improved sanitation facilities (%) (PISF), Total population (TP), People born in the year (PB), Birth rate per 1,000 inhabitants (BR), Demographic density (DD), Male smokers (%) (PMS), Female smokers (%) (PFS), People with HIV (%) (PPHIV), Life expectancy (LE). The database has 2 attributes that are used as outputs for predicting and generating knowledge: the number of deaths from pneumonia in newborns from 0 to 27 days old (NDN) and the number of pneumonia deaths in children up to 5 years old (NDC).

**Step 6** - The sixth step is to perform the database exploration, identifying the data type, minimum value, maximum value, mean, median, mode, standard deviation, coefficient of variation, absolute deviation, amplitude and variance, aiming to analyze the representativeness of the dataset. Most of the data in the database is continuous quantitative type, and only country data such as whether pneumonia vaccine is on the regular vaccination schedule (RVC) is from categorical type.

Analyzing the data, we notice a great inequality between some countries. For example, health expenditure (HE) ranges from US$ 3.00 to US$ 9,715.00 per capita and the percentage of children

[6] http://apps.who.int/gho/data/node.home. Accessed on May 08, 2020
[7] https://data.worldbank.org. Accessed on May 08, 2020

vaccinated against pneumonia (PPV) ranges from 1% to 99%. Standard deviation, amplitude and variance were also high in some data such as health expenditure (HE) and gross national income (GNI). Given the need to better understand the outputs to be predicted, statistical description were also applied to this data.

**Step 7** - This step of pre-processing is to analyze and explore the data that will be used in the knowledge discovery process, analyzing missing data and outliers. With the reorganization of the database performed in the fifth step, many missing data emerged that needed treatment. For each attribute, the process for parsing outliers and missing data is performed, applying the analysis in the database attributes. The analysis of missing data was performed for all attributes. After ensuring the integrity of the database carried out in Step 5, few missing data was found. Only two instances (Afghanistan and South Sudan, both for the year 2000) had many missing attributes, and were therefore removed from the database. Removing outliers was not necessary.

**Step 8 and 9** - In this stage of Pictorea, all numeric attributes of the input set were standardized using the standard score formula (z-score). Data needs to be transformed so that it can be input to mining algorithms. Data related to the number of people with HIV (PPHIV), health expenditure (HE) and gross national income (GNI) are transformed. Initially, the HIV-related figure is the number of people who have HIV in a given country. However, since the population is very large and the expected results are on the scale of deaths per 1,000 born, countries with larger populations would be more likely to have more people with HIV than less populous countries. This attribute has been changed to rate of people who have HIV in a particular country. Thus, data related to the number of people with HIV, health expenditure and gross national income were changed to per capita (in the financial data) and percentage (in the remaining) to not negatively influence the prediction. In addition to these transformations, for the operation of the machine learning algorithms, it was necessary to transform the categorical data into discrete numeric variables.

**Steps 10, 11, 12 and 13** - After data preparation, the machine learning techniques for prediction are applied. For the three experiments a Random Forest algorithm configured with 1,000 trees and a Tensor Flow ANN (Multilayer Perceptron) were used with the first layer having 1,000 neurons and the second having 100, using a total of 2,000 epochs and generating an output of the quantitative type. These were implemented in *Python* and the settings were defined from empirical tests.

Experiment 1 was carried out with the objective of predicting the categorization from percentage ranges of the number of pneumonia deaths in newborns from 0 to 27 days old (NDN). Experiment 2 aims to predict the number of pneumonia deaths in newborns from 0 to 27 days old (NDN). Finally, experiment 3 aims to predict the number of children up to 5 years old deaths from pneumonia (NDC). In the three experiments, the set of input features was: Dataset = {HE, PPV, RVC, GNI, PBSF, PISF, TP, PB, BR, DD, PMS, PFS, PPHIV, LE}.

Patterns discovered by the previous procedure are statistically verified to estimate their representativeness with data not considered during the training phase of the models. For this, the hold-out method is used, which divides the data into a fixed percentage of examples, where a certain amount of data size $p$ is considered for training, while *1-p* tests. Thus, we used $p = 2/3$ and the base divided into 2 sets: 2/3 from the training database and 1/3 from the validation database. Given this, out of the total of 556 data, 371 data are used for training algorithms and 185 for testing.

## 4. EVALUATION RESULTS

The first experiment is performed to analyze and predict the number of newborns deaths from pneumonia. For this, the results related to the number of newborns dead by pneumonia (NDN) per 1,000 born are categorized (see Table I), in order to allow the percentage estimate of newborns mortality in a given location. Thus, the number of deaths from this disease is divided by the number of people born in the country, and it was given as a percentage. Subsequently, this percentage is categorized,

under a domain specialist's guidance, into 5 different classes, as shown in Table I.

Table I: Categorization of newborn pneumonia deaths

| **Value Range** | **Class** | **Total Data** | **Training Data** | **Testing Data** |
|---|---|---|---|---|
| 0% to 0.5% | 1 | 282 | 156 | 126 |
| 0.5% to 1.5% | 2 | 142 | 101 | 41 |
| 1.5% to 2.5% | 3 | 103 | 89 | 14 |
| 2.5% to 3.5% | 4 | 25 | 22 | 3 |
| 3.5% to 6% | 5 | 4 | 3 | 1 |

After applying the Random Forest (RF) and ANN techniques, Tables II and III show the confusion matrix of both results of the predictions in the test database. In this scenario, Random Forest (RF) correctly categorizes 151 data (representing accuracy of 81.62% of the test base), while ANN correctly categorizes 162 (representing 87.57%).

Table II: RF Confusion Matrix - NDN

| Actual/Prediction | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | **112** | 14 | 0 | 0 | 0 |
| 2 | 2 | **34** | 5 | 0 | 0 |
| 3 | 0 | 10 | **4** | 0 | 0 |
| 4 | 0 | 0 | 2 | **1** | 0 |
| 5 | 0 | 0 | 1 | 0 | **0** |

Table III: ANN Confusion Matrix - NDN

| Actual/Prediction | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | **119** | 6 | 1 | 0 | 0 |
| 2 | 6 | **33** | 2 | 0 | 0 |
| 3 | 0 | 6 | **8** | 0 | 0 |
| 4 | 0 | 0 | 2 | **1** | 0 |
| 5 | 0 | 0 | 0 | 0 | **1** |

In addition, the ANN algorithm is able to correctly predict the atypical case and, therefore, more difficult to predict, where the categorized actual value was 5, while the Random Forest could not get this prediction right. Analyzing both tables, it is also noticed that even the predictions made incorrectly are close to the actual values, both algorithms achieved good results with the training base. To confirm the better performance of the proposed ANN compared to RF, other 2 experiments (1.1 and 1.2) are performed by randomly varying the training and testing bases, maintaining hold-out and $p = 2/3$. Table IV illustrates comparisons of hit percentage for each experiment.

All experiments prove that the ANN algorithm has greater accuracy than the RF. In the first experiment, accuracy results found by the ANN are 5.95%, 8.64% and 6.49%, respectively, higher than the RF. In addition, the confusion matrix shows that the ANN model can better predict atypical cases, such as actual (real) values 4 and 5 in the second experiment and actual value 4 in third experiment. Given the superiority of the ANN model for predicting results in this database, it is used for other predictions (Experiments 2 and 3) within the same database and new knowledge findings.

The experiment 1 dealt with a multiclass classification problem, as their results are categories created from the percentage of children deaths from pneumonia in a given country. In this second experiment, the results are expressed by the number of deaths of newborns (up to 27 days old) due to pneumonia per 1,000 children born (NDN), thus the regression model.

To evaluate the prediction results in this second experiment, the Root Mean Square Error (RMSE) is calculated and the results are presented in a scatter plot. According to [Willmott 1982], RMSE is one of the best general measures of model performance and its error value is presented in the same

Table IV: Experiment results

| **Experiment** | **RF Hits** | **RF Accuracy** | **ANN Hits** | **ANN Accuracy** |
|---|---|---|---|---|
| 1 | 151 | 81.62% | 162 | 87.57% |
| 1.1 | 143 | 77.30% | 158 | 85.94% |
| 1.2 | 149 | 80.54% | 161 | 87.03% |

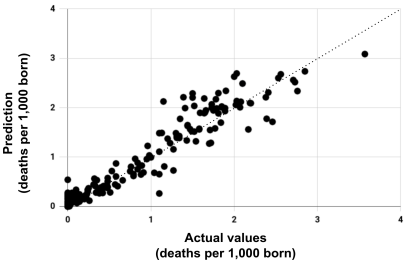


Fig. 1: Prediction results of NDN (Experiment 2)

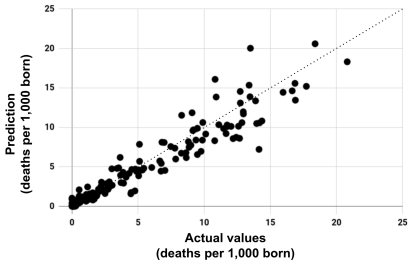


Fig. 2: Prediction results of NDC (Experiment 3)

dimensions as the analyzed variable. The set of features used to make the predictions can be found in Steps 10-13 of Subsection 3.

First, the database was divided into training and testing, using hold-out and p = 2/3. Subsequently, the ANN model was used to make the predictions, seeking in this experiment to predict the number of deaths of newborns (up to 27 days old) due to pneumonia per 1,000 children born (NDN). We analyzed 3 different results using the RMSE as a performance measure, dividing the results according to the countries financial development [Nations 2020]. The first application of RMSE was carried out on the predictive results obtained by analyzing 5 developed countries (Germany, Australia, USA, Portugal and Ireland), then it was applied in 4 developing countries (China, Chile, Egypt, Peru) and finally applied in 5 underdeveloped countries (Angola, Haiti, Afghanistan, Niger and Mali).

For developed country, RMSE is 0.05396, for developing countries, it is 0.12143 and for underdeveloped countries, it is 0.28196. In addition to the RMSE calculation, a scatter plot is created to compare actual and predicted values. Figure 1 illustrates this scatter plot. The dashed line represents a perfect model in which all results would have total accuracy. It is noticed that the results are close or in the dashed line, which reflects the accuracy of the algorithm used.

In the third experiment the results are in the continuous quantitative type. They represent the number of deaths of children up to 5 years old due to pneumonia per 1,000 children born (NDC). For this experiment, the same proposal presented in experiment 2 was used. The ANN model was used to make the predictions, seeking to predict the number of deaths of children up to 5 years old due to pneumonia per 1,000 children born (NDC). The same division of countries was carried out to apply the RMSE. Therefore, for the results of developed countries, the RMSE is 0.13032, for developing countries, it is 0.25948 and for underdeveloped countries, it is 2.38311. These values are larger than the second experiment because the data range for the results of this experiment is much larger than the previous experiment.

Figure 2 illustrates the scatter diagram as shown in the previous experiment. For this experiment, the results are also close or in the dashed line, which reflects the accuracy of the algorithm used. In the last step of the Pictorea process, the results found by the experiments performed in the ANN algorithm are analyzed. A domain specialist validated the data pattern as relevant for future applications in the health area, for instance, to predict future disease scenarios.

## 5. CONCLUSIONS

The purpose of this work is to develop an approach capable of analyzing and predicting childhood deaths by pneumonia based on different regions and attributes. This way, our research analyzed the number of newborns up to 27 days old and the number of children (up to 5 years old) who died from pneumonia in social groups for different countries. Consequence of these findings is to support new strategies concerning public policy towards childhood death prevention and reduction.

For this, it was necessary to build a database. All steps were performed using the Pictorea method-

ology, which was extremely important to support the work of a KDD specialist and a domain specialist. Thus, a domain specialist's knowledge played a valuable role, mainly in establishing what was really important to analyze and evaluate. In comparison to Random Forest algorithm, the Neural Network algorithm presented better results for predicting with accuracy of up to 87.57% in the experiments performed. It was used to predict the number of childhood deaths due to pneumonia. For these experiments, the results were satisfactory, with RMSE showing low errors in predictions.

A domain specialist validated the results and pattern relevance in order to perform future studies in the health field, such as the behavior of countries to combat childhood pneumonia. Actions to accelerate the reduction of this disease can be done based on the predicted numbers in any social group or period. For future work, we will collect new data from Covid-19 databases to extend our results towards pneumonia deaths.

## REFERENCES


Afifi, W. N. W. M., Warsito, I. F., Sayahkarajy, M., and Supriyanto, E. The development of an online pneumonia risk prediction system. In *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*. pp. 1–5, 2017.

Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., and Cheng, X. Artificial intelligence and machine learning to fight covid-19. *Physiological Genomics* 52 (4): 200–202, 2020.

Apostolopoulos, I. D. and Mpesiana, T. A. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med*, 2020.

Breiman, L. Random forests. *Machine Learning* 45 (1): 5–32, Oct, 2001.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '15. ACM, New York, NY, USA, pp. 1721–1730, 2015.

Chaves, L. E., Nascimento, L. F. C., and Rizol, P. M. S. R. Modelo fuzzy para estimar o número de internações por asma e pneumonia sob os efeitos da poluição do ar. *Revista de Saúde Pública* vol. 51, pp. 1–8, 2017.

COVID-19. Open research dataset (cord-19), 2020. Accessed: 2020-04-05.

Duan, Z., Han, X., Bai, Z., and Yuan, Y. Fine particulate air pollution and hospitalization for pneumonia: a case-crossover study in shijiazhuang, china. *Air Quality, Atmosphere & Health* 9 (7): 723–733, 2016.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI magazine* 17 (3): 37, 1996.

Jothi, N., Rashid, N. A., and Husain, W. Data mining in healthcare – a review. *Procedia Computer Science* vol. 72, pp. 306 – 313, 2015. The Third Information Systems International Conference 2015.

Kraft, M. R., Desouza, K. C., and Androwich, I. Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the.* pp. 9 pp.–, 2003.

Laiakis, E., Morris, G., Fornace, A., and Howie, S. Metabolomic analysis in severe childhood pneumonia in the gambia, west africa: findings from a pilot study. In *PLoS One*. Vol. 5, 2010.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the sustainable development goals. *The Lancet* 388 (10063): 3027–3035, 2017.

Montevecchi, A. and Zárate, L. *Pictorea: Um método para descoberta de conhecimento em bancos de dados convencionais*. Novas Edições Acadêmicas, United States, 2014.

Nations, U. World economic situation and prospects, 2020. Accessed: 2020-04-20.

Naydenova, E., Tsanas, A., Casals-Pascual, C., and De Vos, M. Smart diagnostic algorithms for automated detection of childhood pneumonia in resource-constrained settings. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*. pp. 377–384, 2015.

Organization, W. H. The top 10 causes of death, 2014. Accessed: 2018-01-04.

Scotta, M. C., Marostica, P. J., and Stein, R. T. 25 - pneumonia in children. In *Kendig's Disorders of the Respiratory Tract in Children (Ninth Edition)*, Ninth Edition ed., R. W. Wilmott, R. Deterding, A. Li, F. Ratjen, P. Sly, H. J. Zar, and A. Bush (Eds.). Content Repository Only, Philadelphia, pp. 427 – 438.e4, 2019.

Willmott, C. J. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* 63 (11): 1309–1313, 1982.

Yang, J.-J., Li, J., Mulder, J., Wang, Y., Chen, S., Wu, H., Wang, Q., and Pan, H. Emerging information technologies for enhanced healthcare. *Computers in Industry* vol. 69, pp. 3 – 11, 2015. Special Issue: Information Technologies for Enhanced Healthcare.