Machine Learning for Epidemiological Analysis in The Industrial Area for a Sustainable Life

J. Susymary, P. Deepalakshmi

Abstract*: Pollution exposure and human health in the industry contaminated area are always a concern. The need for industrialization urges to concentrate on sustainable life of residents in the vicinity of the industrial area rather than opposing the industrialists. Literature in epidemiological studies reveal that air pollution is one of the major problems for health risks faced by residents in the industrial area. Main pollutants in industry related air pollution are particulate matter ($PM_{2.5}$, PM_{10}), SO_2, NO_2, and other pollutants upon the industry. Data for epidemiological studies obtained from different sources which are limited to public access include residents' sociodemographic characters, health problems, and air quality index for personal exposure to pollutants. This combined data and limited resources make the analysis more complex so that statistical methods cannot compensate. Our review finds that there is an increase in literature that evaluates the connection between ambient air pollution exposure and associated health events of residents in the industrially polluted area using statistical methods, mainly regression models. A very few applies machine learning techniques to figure out the impact of common air pollution exposure on human health. Most of the machine learning approach to epidemiological studies end up in air pollution exposure monitoring, not to correlate its association with diseases. A machine learning approach to epidemiological studies can automatically characterize the residents' exposure to pollutants and its associated health effects. Uniqueness of the model depends on the appropriate exhaustive data that characterizes the features, and machine learning algorithm used to build the model. In this contribution, we discuss various existing approaches that evaluate residents' health effects and the source of irritation in association with air pollution exposure, focuses machine learning techniques and mathematical background for epidemiological studies for residents' sustainable life.

***Keywords* : Epidemiological studies, sustainable life, air pollutants, air pollution exposure, sociodemographic characters, health problems, statistical methods, machine learning, mathematical background.**

I. INTRODUCTION

Epidemiological studies have been often made to assess the association of the health consequences of individuals and ambient air pollution in European as well as Asian countries [1]-[4]. A few literatures have been there to find the health impacts of residents due to industrial pollution. The main concern for these kinds of studies are limited to capacity and resources for comprehensive environmental data caused by industrial pollution. A statistical approach necessitates known prior assumptions about the data, its functional form and probability distributions. Since data is collected from variable resources, which makes them high dimensional, a statistical method alone cannot handle. Then how to find those environmental risks within the limited public access? Automated empirical model requires no prior knowledge of data. Machine learning (ML) is one if this kind.

Machine learning is learning by example. It focuses on the mathematics of computing to build a model automatically from the data that has no known prior assumptions, its functional form and probability distributions. Machine learning model makes the machine to learn with a comprehensive set of example data known as a training set that spans as much of parameter space. Then ML tests its generalization by using an independent validation dataset to check how it performs when present with a data that the algorithm never seen. This validation dataset can be a randomly selected subset of training data that was held back for validation. So, for a successful machine learning model, a comprehensive training dataset and an algorithm is necessary. Machine learning algorithms are of four distinct groups, namely supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning is of two types namely classification and regression. Supervised learning has been extensively used for prediction and time series forecasting. There are different supervised learning approaches to make the machine learn, from basic decision trees to clustering layers of neural networks. It typically performs the following tasks:

- Identify data, integrate, and preprocess the dataset
- Select an appropriate machine learning algorithm
- Empirical model building
- Train, the model on test data
- Generate scores and findings

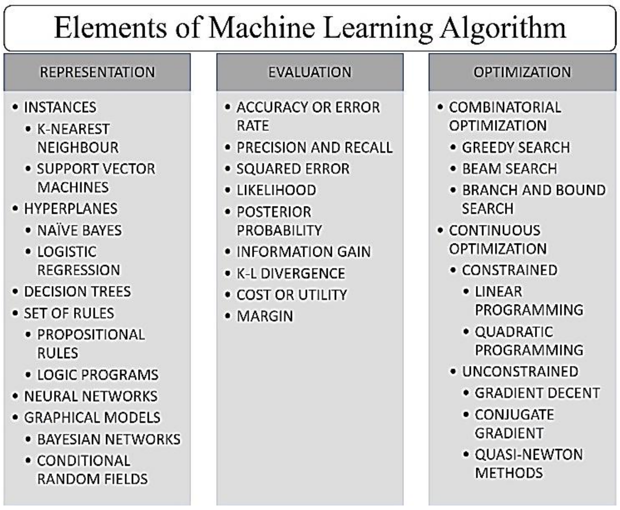
A supervised learning would be feasible for epidemiological analysis since it focuses mainly on prediction or forecasting [5], [6]. The supervised machine learning algorithm comprises of three elements namely representation, evaluation, and optimization. Depending upon the model, representation can do with a set of classifiers such as, instances, hyperplanes, decision trees, set of rules, neural networks or through graphical models. Evaluation by using accuracy or error rate, precision and recall, squared error, the likelihood, posterior probability, information gain, K-L divergence, cost or utility, or margin. Optimization techniques include either combinatorial or continuous. The combinatorial optimizations are the greedy search, the beam search, and the branch and the bound search.

Revised Manuscript Received on December 15, 2019.
* Correspondence Author

J. Susymary*, Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Virdhunagar, Tamil Nadu, India. Email: susymaryj@gmail.com.

P. Deepalakshmi, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Virdhunagar, Tamil Nadu, India. Email: deepa.kumar@klu.ac.in.

The continuous optimization is either constrained based or unconstrained based. Fig. 1 outline the elements of machine learning algorithm.



Elements of Machine Learning Algorithm

REPRESENTATION
- INSTANCES
  - K-NEAREST NEIGHBOUR
  - SUPPORT VECTOR MACHINES
- HYPERPLANES
  - NAÏVE BAYES
  - LOGISTIC REGRESSION
- DECISION TREES
- SET OF RULES
  - PROPOSITIONAL RULES
  - LOGIC PROGRAMS
- NEURAL NETWORKS
- GRAPHICAL MODELS
  - BAYESIAN NETWORKS
  - CONDITIONAL RANDOM FIELDS

EVALUATION
- ACCURACY OR ERROR RATE
- PRECISION AND RECALL
- SQUARED ERROR
- LIKELIHOOD
- POSTERIOR PROBABILITY
- INFORMATION GAIN
- K-L DIVERGENCE
- COST OR UTILITY
- MARGIN

OPTIMIZATION
- COMBINATORIAL OPTIMIZATION
  - GREEDY SEARCH
  - BEAM SEARCH
  - BRANCH AND BOUND SEARCH
- CONTINUOUS OPTIMIZATION
  - CONSTRAINED
    - LINEAR PROGRAMMING
    - QUADRATIC PROGRAMMING
  - UNCONSTRAINED
    - GRADIENT DECENT
    - CONJUGATE GRADIENT
    - QUASI-NEWTON METHODS

Fig. 1.Elements of machine learning algorithm

The rest of this article is organized as follows. In section 2, we mainly described the materials and methods used to collect different forms of data, and challenges in statistical approach on epidemiological studies in the industrial area. The different air pollution exposure monitoring measures that has been applied are also mentioned in this section. Section 3 discuss various machine learning techniques and their performance evaluation in each article. Section 4 describes the mathematical background for the study. Then follows the findings of the review, current system and future directions, and conclusion.

II. LITERATURE ON EPIDEMIOLOGICAL STUDIES IN INDUSTRIAL AREA

Recent published papers related to epidemiological studies in the industrial area has been taken and reviewed its background and objectives, data collection methods, and measures taken to analyse the collected data. Articles that have been reviewed in this section performs statistical regression analysis with evaluation measures, an odds ratio and a confidence interval (95%) to identify the association between risk factors and related diseases [7]-[13]. A combination of questionnaire study and air pollution exposure assessment using a dispersion model were used as materials for epidemiological studies that associates air pollution caused by industries and related health issues of individuals living in the proximity of it in [7]-[9],[11],[13]. In addition to that a spirometer has been used to test the lung function in [8] and [13]. Municipal records and interviews were used as part of the study in [10]. An additional clinical examination was conducted to measure nitric oxide inhalation in [11]. It has been inferred that the possible-combined data will be comprised of sociodemographic characters, health problems, clinical data, meteorological data, and air pollution exposure data.

Fig. 2 shows an example of integrated dataset for epidemiological studies in the industrial area. Identified risk factors include $PM_{2.5}$, PM_{10}, SO_2, NO_X, benzene, phenol and FeNO with associated diseases such as myocardial infraction, chest weakness, longstanding cough, a shortfall of breath, asthma attack, decreased lung function, dry cough, wheezing, respiratory symptoms, high blood pressure, lung cancer, cardiac diseases, haematological malignancies, etc. Table 1 shows an overview of the epidemiological studies in the industrial area.

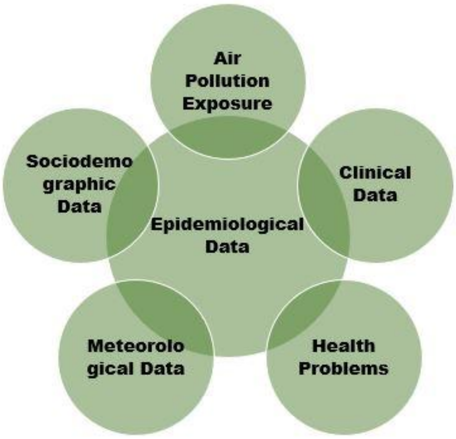


Fig. 2.A combined dataset

A. Air Pollution Exposure Monitoring

All the above reviewed papers used air pollution dispersion modelling with geocode of the residents' address as the distance metric to the measure proximity index of individuals for the exposure assessment. Most of the time, the relationship between predictors and the pollutant concentrations were non-linear. So linear regression models will not work well. But the progression of air pollution exposure assessment can be range from ground-based, single site air pollution exposure model to an advanced spatial-temporal model which uses advanced machine learning algorithms [14]-[20]. Recent advances in sensor technology facilitates personal exposure monitoring in a sustainable manner [21], [22]. Mobile applications in android also do the personal air pollution exposure in a stratified manner [23], [24].

B. Challenges

Challenges of an epidemiological model is related to the nature of data. Data is complex due to its stratified nature and heterogeneous definition which needs cross domain data integration. Lack of uniformity of data, including quantitative and qualitative variables measured with different scales to characterize the unique exposure makes integration difficult. Integrating variables with zero values known as data transformation, and normalization are needed before analysis in order to avoid biased parameter estimates. But in some machine learning techniques, normality assumptions are different. Since the epidemiological analysis is either predictive or forecasting in nature, accuracy is a major concern than time complexity. The machine learning techniques that are appropriate in sustainable epidemiological analysis which has a potential to incorporate analytical procedures that characterize the relationship between health risks and risk factors is necessary for the sustainable life of residents living in the industrial area.

Table- I: Overview of epidemiological studies in the industrial area

Reference	Purpose of Study	Location	Materials	Approach/Model	Identified Risk Factors and Diseases
[7]	Figure out the health problems and irritation related to exposure to the air pollutants in the industrially polluted region	East Estonia	• Questionnaire survey ○ Sociodemographic characters ○ Health problems • Air pollution exposure using a dispersion model using meteorological data	Statistical, Multiple Regression analysis	• Benzene - myocardial infraction • Phenol - chest tightness, long-term cough, myocardial infraction • $PM_{2.5}$ – chest tightness, shortness of breath, an asthma attack
[8]	Evaluation of lung function and breathing problems among school students due to industrial air pollution	Southwest of Netherlands	• Questionnaire survey to parents ○ Sociodemographic characters ○ Health problems • Lung function measurement using spirometer • Air pollution exposure using a dispersion model using meteorological data	Statistical, Linear and Logistic Regression analysis	• $PM_{2.5}$ and NO_X – decreased lung function • $PM_{2.5}$ – dry cough
[9]	The role of risk factors in industrial air pollution and human health	Southwest of Netherlands	• Questionnaire study ○ Sociodemographic characters ○ Health problems • Air pollution exposure using a dispersion model using meteorological data	Statistical, Logistic Regression analysis,	• $PM_{2.5}$ and NO_X – wheezing, dry cough • $PM_{2.5}$ – high blood pressure • Parental worry is a mediator
[10]	Finding the role of air pollution from an oil refinery for hematological malignancies' death	Italy	• Municipal files • Interviews with relatives of departed	Statistical, Conditional Logistic Regression with time weighted average residential proximity	• Benzene – increased risk of the hematological malignancies' death
[11]	Association between the concentration of nitric oxide due to industrial pollution and respiratory symptoms, asthma in school children	Estonia	• Questionnaire survey ○ Sociodemographic characters ○ Health problems • Clinical examination to measure nitric oxide (FeNO) • Air pollution exposure using a dispersion model using meteorological data	Statistical, Logistic Regression analysis, Chi-Squared Test	• FeNO – respiratory symptoms • Benzene and formaldehyde – rhinitis without cold, attacks of asthma
[12]	Longstanding air pollution exposure from multiple sources and fatality in the industrial area	Italy	• Municipal registers • Long term air pollution exposure using a dispersion model using meteorological data	Statistical, Cox Survival analysis using a linear term and dichotomous variable	• PM_{10} from industry - cancers, cardiac diseases • NO_X from traffic - cancers, neurological diseases • PM_{10} and NO_X - Neurological diseases, the lung cancer
[13]	Exposure to pollutants from an oil refinery and decreased lung function among school children	Italy	• Questionnaire study ○ Sociodemographic characters ○ Health problems • Spirometry tests for lung function • Air pollution exposure monitoring	Statistical, Multiple Regression model	• SO_2 – reduction of lung function, airway inflammation

III. LITERATURE ON MACHINE LEARNING FOR EPIDEMIOLOGICAL STUDIES

It has been learned that the temporal texture of epidemiological data, and the demand for real time prediction creates an awareness whether to stick on at time series forecast, or predictive model. Bayesian models incorporate qualitative information for training whereas Non-Bayesian models utilize quantitative information without qualitative inputs. In simple regression models, normal distributions will not hold environmental exposures. Even complex regression models require specification of both response and explanatory variables. Neural network models can learn from complex stratified groups. Literature learning in this section focuses different machine learning algorithms and its performance in epidemiological studies even though most of the pollution exposure data in the reviewed papers are not related to industrial proximity.

A neural network representation for a machine learning predictive model in a time series perspective could be found in [25]. They endured recurrent neural network (RNN), and convolutional neural network (CNN) models respectively for a longstanding correlation between data and mine information from data from different sources as a combined model with residual links for optimization, CNNRNN-Res and compared them against autoregressive models (AR, GAR, VAR) and a non-linear model Gaussian Process Regression (GPR) in the US and Japan datasets. The evaluation metrics are Root Mean Square Error (RMSE) and Pearson's Correlation Coefficient. CNNRNN-Res have a domination performance (Lower RMSE value) even with a smaller number of model parameters, and a higher standard deviation than the other compared models. Fig. 3 shows the results summary of ML algorithms.

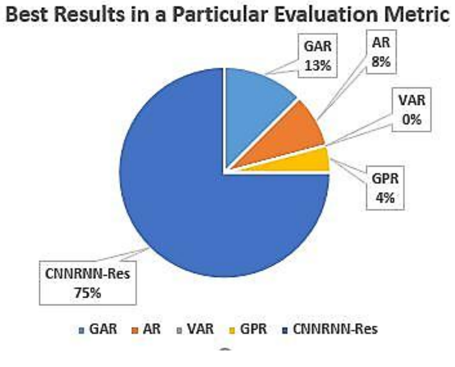


Fig. 3. Percentage of the best results for each algorithms

The effect of multiple air pollutants on health issues using a two-stage approach were implemented in [26]. Ambient air pollution exposure was identified using a CaRT model in stage 1. The correlation of magnitude of the pollutants, and math scores were evaluated using a multivariate linear regression model (MLR) in stage 2 and resulted in an odds ratio, -1.19 points and 95% confidence interval. Fig. 4 shows the performance of CaRT and multivariate regression models.

Three ANN models, namely multilayer perception (MLP), extreme learning machines (ELM), and echo state networks (ESN) are used to evaluate the number of hospital admissions due to breathing illness and the impact of $PM_{2.5}$, temperature, relative humidity in [27]. The statistical performance of each model is compared with the result obtained from Anova Friedman's test. They have used absolute principal component analysis (APCA) to extract four principal elements from ten, with group data variance, eigenvalues, and rotation matrix. Out of three ANN models, MLP gave the best results (lower Mean Square Error, and p-value < 0.001). ANN seems to be more sensitive than any other statistical regression models. Fig. 5 shows the efficiency of ANN algorithms.

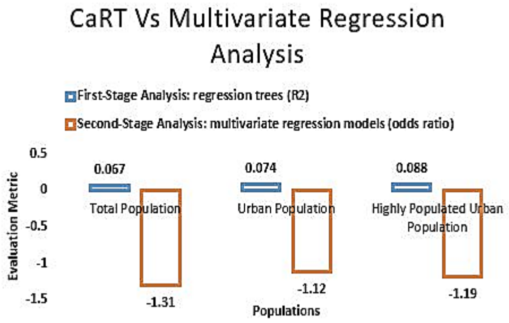


Fig. 4.Performance of CaRT and multivariate regression models

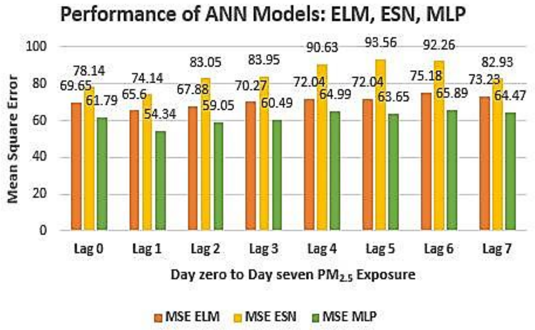


Fig. 5.Performance of ANN models

LASSO regression was used to spot the correlation amidst pollutant concentrations, and disease by a case control study with multiple statistical testing penalty in [28]. Variables with non-zero estimates are selected as narrow down variables using conditional logistic regression with a ten-fold cross validation for the tuning parameter, λ in stage 1, and then performed LASSO regression to find the disease association with an odds ratio (pooled = 3.11) in stage 2. Fig. 6 shows the performance of the regression model.

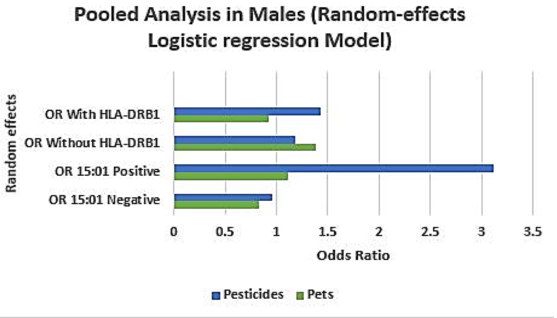


Fig. 6.Performance of regression model

A country level data analysis to associate $PM_{2.5}$ and PM_{10} pollutants for hospital admissions due to upper respiratory infections in Taiwan have been done with Multilayer Perception (MLP) on 33% training set and 67% testing set in [29]. Normalization in the pre-processing stage on both particulate matter concentrations and combined particulate matters has been done before given as input to the MLP. A higher accuracy rate in overall population, 81.75% for $PM_{2.5}$ and 83.21% for PM_{10} concentrations. Fig. 7 shows the performance of MLP model.

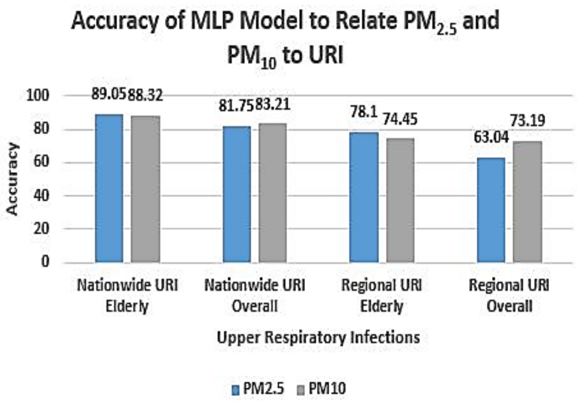


Fig. 7.Performance of MLP model

An information-based machine learning model with an artificial neural network (ANN), using a back-propagation (BP) algorithm was developed for estimating indoor airborne culturable fungi by monitoring indoor air quality data in [30]. The model with one hidden layer shows higher accuracy than the compared SVM model. The ANN model with ten hidden nodes shows an accuracy rate of 83.11%. Already a general regression neural network (GRNN) model was developed by the same authors for an abrupt measurement of indoor bacterial concentrations with simple indoor air quality inputs [31]. Fig. 8 shows the accuracies in the hidden layers of ANN model.

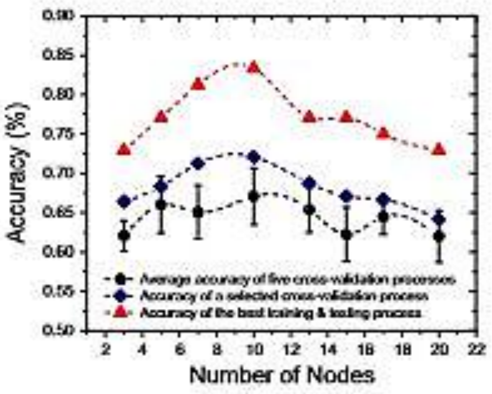


Fig. 8.[30] Trends in accuracy in the layers of ANN model

A deep neural network (DNN) model was developed to pinpoint environmental risk aspects linked with acute breathing problems based on integrated respiratory disease data, air pollution data, and meteorological environment data in [32]. The data was converted into HDFS format in the pre-processing phase and train deep learning models in six stratified groups. Risk factors are compared with the normalized weight feature. The developed model's precision was greater and training time was brisker than the compared C4.5 classification algorithm [33]. The risk factors are same for both the algorithm. Fig. 9 shows the performance of the DNN model for each stratified population.

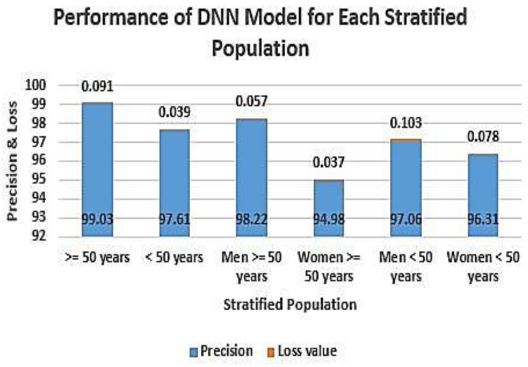


Fig. 9.Performance of the DNN model

From the reviewed articles, it has been learned that neural network models have more significance in accuracy, and efficiency than any statistical and machine learning models. Also, with some simple input parameters, the machine learning techniques such as ANN and SVM accomplish explicit predictions of secondary variables that are crucial to detect from traditional dimensions [34]-[39]. The deep neural network outperforms in bigger and deeper networks. A feed forward neural network training process can learn intricate characteristics from diverse stratified groups of data. The deep learning method spare artificial neural network models. The Table 2 shows an overview of machine learning approaches to epidemiological studies.

IV. MATHEMATICAL BACKGROUND FOR DATA ANALYSIS

The performance of the model depends on the algorithm's accuracy, training time, complexity, the number of parameters and features used, choosing parameter settings and validation strategies, identifying under lifting and over lifting, estimating confidence interval, and uncertainty. A thorough understanding of mathematics behind these factors such as linear algebra for optimization methods, probability and statistics for inference, multi-variate calculus, and complex optimizations for computational efficiency and scalability is required to build a sustainable model.

A. Setting up of the Data

A dataset of m objects can be obtained as vectors of features, a_j outcomes or observations or labels, y_j for each feature; where $j = 1,2, \dots, m$ after cleaning and formatting. The outcome y_j preserve a real number for regression, a label for classification where a_j lies in any of M classes $(M > 2)$, multiple labels for different paradigm, no labels (y_j is null) for subspace identification and clustering (partition a_j into few clusters).

B. Mapping Function in Data Analysis

The first step in the data analysis task of machine learning is learning phase or training phase. This is the process of finding a function φ known as the mapping function. It is the identification of a data fitting

Table- II: Overview of machine learning for the epidemiological studies

Reference	Purpose of Study	Dataset	Models	Evaluation Metrics	Remarks
[25]	Deep Learning for Epidemiological Studies	• Japan-Prefectures • US-Regions • US-HSS	• GAR • AR • VAR • GPR • CNNRNN-Res	• Root Mean Square Error • Pearson's Correlation Coefficient	CNNRNN-Res is robust and outperforms other models. (Lower RMSE value)
[26]	Identification of air pollutants associated with initial cognitive problems among children in U.S using machine learning techniques	• ECLS-B, NCES 2010-009 [40] • Outcome assessment for mathematics • EPA [41] • USDA [42]	• Two Stage: • Stag1- CaRT, • Stage2-Multivariate Linear Regression	• CaRT- Squared Error • Regression- Odds Ratio	High isophorone exposure levels (OR= -1.19) were found by regression model
[27]	Artificial neural network models for assessing the impact of $PM_{2.5}$ on breathing problems	• Daily PM2.5 concentrations • Sampling campaigns • Hospital admission register • Temperature • Relative humidity	• ANN Models – • MLP • ELM • ESN	• Mean Square Error (MSE) • Mean Absolute Percentage Error (MAPE) • Number of artificial Neurons (NN) • Anova Friedman's test	MLP gave the best results, lower MSE with p-value < 0.001and robust for small datasets
[28]	Assessment of environmental risk factors for multiple sclerosis using machine learning techniques	• Population based Case control study	• Logistic Regression, • Least Absolute Shrinkage and Selection Operator (LASSO)	• Odds Ratio	Environmental risk factors using logistic regression and its association of disease using LASSO model. (OR pooled=3.11)
[29]	A nation-wide analysis of outpatient visits in hospitals for upper respiratory infections in relation to $PM_{2.5}$ and PM_{10} using multilayer perception algorithm	• Daily nationwide and regional outdoor $PM_{2.5}$ and PM_{10} concentrations from Taiwan Environment Protection Administration • Upper Respiratory Infections (URI) data from Centers for Disease Control Taiwan • Insurance records	• Multilayer Perception (MLP)	• Accuracy/Error rate	MLP detect the correlation of URI and $PM_{2.5}$ (81.75% accuracy), PM_{10} (83.21% accuracy) concentrations in overall population better than regional level.
[30]	A neural network approach to explore the rapport amidst indoor air quality and the level of airborne culturable fungi	• 249 data sets of Indoor air quality indicators from 85 residential buildings in China	• ANN (Back-propagation) • SVM	• Accuracy/Error rate	ANN using back-propagation with one hidden layer has positive prediction accuracy than SVM. ANN with ten hidden nodes has the prediction accuracy of 83.33% with +(-) 30%
[32]	Identification of environmental risk aspects of intense respiratory diseases in Chinese population with distinct age and gender using deep learning	• Respiratory disease data • Air pollution data • The meteorological environment data	• Stratified Deep Neural Network (DNN) • C4.5 classification algorithm	• The feature weight of the risk factor • Normalized weight • Precision and recall	Higher precision for DNN, fast learning method, and efficient than C4.5 algorithm. Risk factors are same for both the model

problem (find the best parameter), when φ is defined in terms of some parameter vector x. It is a function that approximately maps a_j and y_j for each $j: \varphi(a_j) \approx y_j$ for $j = 1,2,\dots,m$. If there are no labels y_j or if some labels are missing, the function φ will perform something suitable with data $\{a_j\}$. It satisfies some additional properties such as simplicity and structure that makes the model robust and generalizable.

The mapping function φ is used both for analysis and prediction. In analysis, the parameter vector x that defines the mapping function φ reveals the structure of the data. For example: feature selection reveals the elements of vectors a_j that determines the output y_j and bulge the importance of these features. Also, it uncovers some hidden structure such as low dimensional subspaces that enclose the vectors a_j, clusters that hold the vectors a_j and a decision tree of vectors a_j and labels y_j. In prediction, for a new data vectors a_j, the predicted output is defined as $y_j \leftarrow \varphi(a_j)$.

The challenges of mapping function are noise or errors in a_j and y_j, over lifting to the particular sample, missing data (vectors a_j), missing labels (output y_j), streaming of data in learning phase. The applications of mapping function include Least Squares, Matrix Completion, Sparse Inverse Covariance, Sparse Principal Component Analysis (SPCA), SPCA with Low-Rank, Subspace Identification, Linear Support Vector Machines, Non-linear SVM, Logistic Regression, and Neural Networks [43].

As learned from literature studies of machine learning approach to epidemiological studies, neural network models outperform other supervised learning models. In the neural network model, input is the vectors a_j and output is the odds of a_j belonging to each class. At every layer, input is converted to output by a linear transformation possessed with feature-wise function using (1):

$$a^{l+1} = \sigma(W^l a^l + g^l) \qquad (1)$$

where a^l is node values at layer, l, and (W^l, g^l) are parameters in the network, σ is the feature-wise function that causes transformation to scalar input:

Logistic function: $t \rightarrow 1/(1 + e^{-t})$;
Hinge: $t \rightarrow max(t, 0)$;
Bernoulli: random! $t \rightarrow 1$ with probability $1/(1 + e^{-t})$ and $t \rightarrow 0$ otherwise.

C. Optimization in Data Analysis

Optimization methodology integrates with applications. The Equation (2) is the basic paradigm for most of the optimization formulations in data analysis is:

$$\min_{x} \frac{1}{m} \sum_{j=1}^{m} h_j(x) + \lambda\Omega(x) \qquad (2)$$

where

h_j count on parameters x of the mapping function φ, and data items (a_j, y_j);

Ω is the regularization term, not always but often non-smooth, convex, and separable in the elements of x. It can also be a symbol for a simple set $x \geq 0$.

$\lambda \geq 0$ is the regularization parameter.

Some of the optimization algorithms are full gradient algorithms such as a gradient with projection, and a gradient for shrinking to handle Ω, an accelerated gradient, a stochastic gradient which is widely used for neural networks [44], hybrids of the full gradient and a stochastic gradient, the coordinate descent, the conditional gradient, newton's method, and approximate newton [45].

V. Findings of The Survey

We performed an unbound search for recent articles organised by epidemiological studies in the industrial area, the articles for the air pollution exposure assessment as part of the data collection for epidemiological studies, and the articles that are organised by the machine learning approach in epidemiological studies. Fig. 10 shows the categorization of the articles organised by the study. The articles that fit the study setting in the industrial area follows a statistical regression analysis which will not fit well for non-linear and limited data. The finding suggests the importance of controlling the model complexity (the number of model parameters) for data insufficient problems. The need for an algorithm that successfully captures the non-linear features is a concern in the model building. The study setting for the epidemiological data analysis can be a hypothesis generation, a prediction, and a time series predication or forecasting in nature. Our literature study finds that the machine learning models such as the neural networks, that can depict the said issues will be suitable for epidemiological analysis to find the association of risk factors and diseases in the industrial area. Fig. 11 shows the categorization of study setting for machine learning algorithms. Table 3 shows the overall analysis of the machine learning algorithms in the present survey.

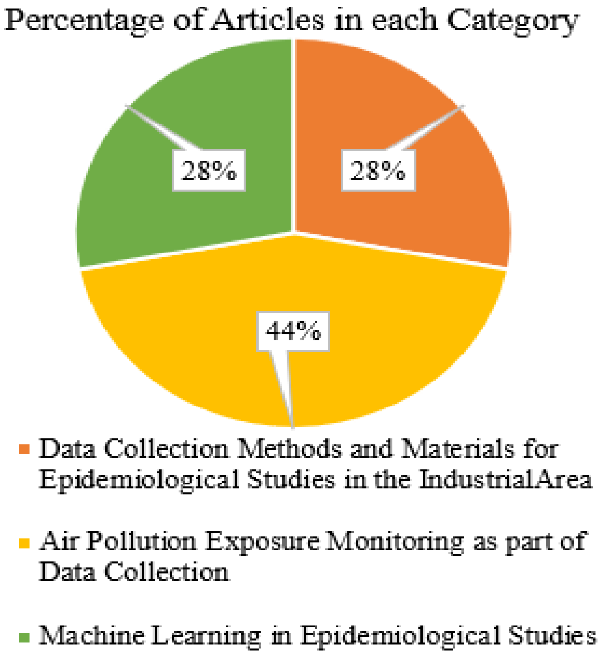


Fig. 10. Categorization of articles

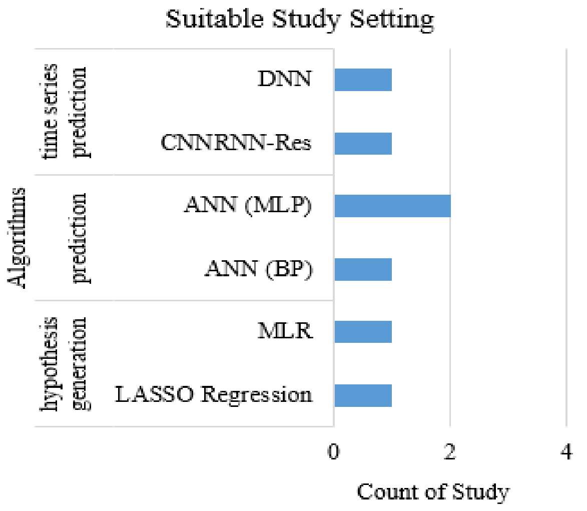


Fig. 11. Categorization of study setting in machine learning

Table- III: Characterization of ML models

ML Model	Performance					
	Heterogeneous Data	Parameters required	Missing Data	Limited Data	Suitable Study Setting	High Dimensional Data
CNNRNN-Res	High	Less	Yes	Yes	Time Series	Yes
MLR	Moderate	More	No	No	Hypothesis	No
ANN (BP)	High	Less	Yes	Yes	Prediction	Yes
LASSO regression	Moderate	More	No	No	Hypothesis	No
ANN (MLP)	High	Less	Yes	Yes	Prediction	Yes
DNN RMSE value	High	Less	Yes	Yes	Time Series	Yes

In Fig. 12, the machine leanring models are compared with their performance outcome, when the algorithm deals with the following characteristics of data:

- Heterogeneous data
- Amount of parameters required
- When there is missing data
- When there is limited data
- High demensional data

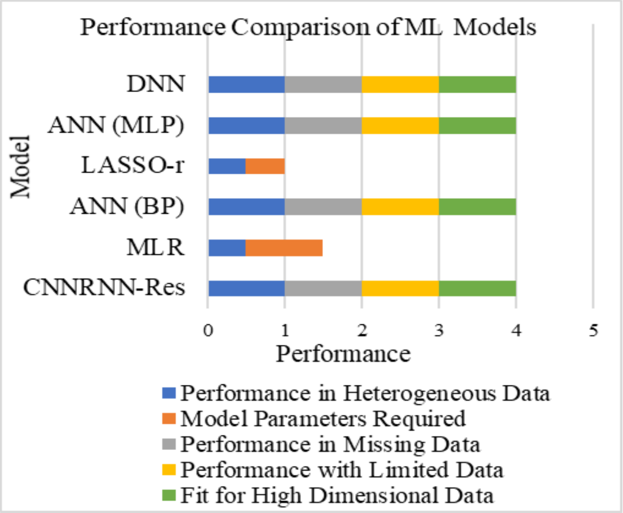


Fig. 12. Performance comparison of ML models

VI. CURRENT SYSTEM AND FUTURE DIRECTIONS

To date, the fundamental transformation of machine learning approach to epidemiological studies is critical. It must face the challenges such as rising burden of illness due to air pollution in the industrial area, and higher societal expectations to find the association between air pollution exposure and health effects. The advantages of machine learning have discussed extensively in the literature. The new era of machine learning technique like neural network framework can resolve those challenges.

VII. CONCLUSION

The review indicates that so many statistical models have been developed for the evaluation of air pollution exposure and related health effects of individuals in the industrial area. As data is stratified in nature, evaluation of the association between the variables is a concern in statistical models. A limited number of machine learning models have been developed to evaluate the health risks associated with pollutants in general, indoor, outdoor, or traffic, and it could perform fast estimation with a higher accuracy rate. As read, the neural network model outperforms all other machine learning models. This strategy could apply to evaluate the industrial air pollution exposure and its health effects in individuals living in the ethnicity of the industry.

REFERENCES

1. M. Seifi, S. Niazi, G. Johnson, V. Nodehi, and M. Yunesian, “Exposure to ambient air pollution and risk of childhood cancers: A population-based study in Tehran, Iran,” Science of the Total Environment, vol. 646, pp. 105–110, January 2019.
2. P. Torres, J. Ferreira, A. Monteiro, S. Costa, M. C. Pereira, J. Madureira, A. Mendes, and J. P. Teixeira, “Air pollution: A public approach for Portugal,” Science of the Total Environment, vol. 643, pp. 1041–1053, December 2018.
3. K. Orru, S. Nordin, H. Harzia, and H. Orru, “The role of perceived air pollution and health risk perception in health symptoms and disease: a population-based study combined with modelled levels of PM10,” International Archives of Occupational and Environmental Health, vol. 91, Issue 5, pp. 581-589, July 2018.
4. A. Kamimura, B. Armenta, M. Nourian, N. Assasnik, K. Nourian, and A. Chernenko, “Perceived Environmental Pollution and Its Health in China, Japan, and South Korea,” J Prev Med Public Health, vol. 50, Issue 3, pp. 188-194, April 2017.
5. C. Kreatsoulas, and S. V. Subramanian, “Machine learning in social epidemiology: Learning from experience,” SSM - Population Health, vol. 4, pp. 347-349, April 2018.
6. B. Seligman, S. Tuljapurkar, and D. Rehkopf, “Machine Leaning approaches to the social determinants of health in the health and retirement study,” SSM - Population Health, vol. 4, pp. 95-99, April 2018.
7. H. Orru, J. Idavain, M. Pindus, K. Orru, K. Kesanurm, A. Lang, and J. Tomasova, “Residents’ Self-Reported Health Effects and Annoyance in Relation to Air Pollution Exposure in an Industrial Area in Estern-Estonia,” International Journal of Environmental Research and Public Health, vol. 15, Issue 2, p. 252, February 2018.
8. A. D. Bergstra, B. Brunekreef, and A. Burdorf, “The Effect of industry-related air pollution on lung function and respiratory symptoms in School Children,” Environment Health, vol. 17, Issue 1, p. 30, March 2018.
9. A. D. Bergsra, B. Brunekreef, and A. Burdorf, “The Mediating role of risk perception in the association between industry-related air pollution and health,” Plos One, vol. 13, Issue 5, e0196783, May 2018.
10. A. Micheli, E. Meneghini, M. Mariottini, M. Baldini, P. Baili, F. Di Salvo, and M. Sant, “Risk

of death for hematological malignancies for residents close to an Italian petrochemical refinery: a population-based case-control study," Cancer Causes Control, vol. 25, pp. 1635-1644, October 2014.

11. J. Idavain, K. Julge, T. Rebane, A. Land, and H. Orru, "Respiratory symptoms, asthma and levels of fractional exhaled nitric oxide in schoolchildren in the industrial areas of Estonia," Science of the Total Environment, vol. 650, pt. 1, pp. 65-72, February 2018.
12. L. Bauleo, S. Bucci, C. Antonucci, R. Sozzi, M. Davoli, F. Forastiere, and C. Ancona, "Long-term exposure to air pollutants from multiple sources and mortality in an industrial area: a cohort study," Occup Environ Med, vol. 76, Issue 1, pp. 48-57, January 2019.
13. F. Barbone, D. Catelan, R. Pistelli, G. Accetta, D. Grechi, F. Rusconi, and A. Biggeri, "A Panel study on Lung Function and Bronchial Inflammation among Children Exposed to Ambient SO2 from an Oil Refinery," International Journal of Environmental Research and Public Health, vol. 16, p. 1057, March 2019.
14. S. Muttoo, L. Ramsay, B. Brunekreef, R. Beelen, K. Meliefste, and R. N. Naidoo, "Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa," Science of the Total Environment, vol. 610-611, pp. 1439-1447, January 2018.
15. G. Hoek, "Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants," Curr Envir Health Rpt, vol. 4, pp. 450-462, December 2017.
16. N. J. Aquilina, J. M. Delgado-Saborit, S. Bugelli, J. P. Ginies, and R. M. Harrison, "Comparison of Machine Learning Approaches with a General Linear Model to Predict Personal Exposure to Benzene," Environmental Science & Technology, vol. 52, Issue 19, pp. 11215-11222, August 2018.
17. L. D. Hill, A. Pillarisetti, S. Delapena, C. Garland, D. Pennise, A. Pelletreau, P. Koetting, T. Motmans, K. Vongnakhone, C. Khammavong, M. R. Boatman, J. Balmes, A. Hubbard, and K. R. Smith, "Machine-learned modelling of PM2.5 exposures in rural Lao PDR," Science of the Total Environment, vol. 676, pp. 811-822, August 2019.
18. T. Xue, Y. Zheng, D. Tong, B. Zheng, X. Li, T. Zhu, and Q. Zhang, "Spatiotemporal continuous estimates of PM2.5 concentrations in China, 2000-2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations," Environment International, vol. 123, pp. 345-357, February 2019.
19. J. Taylor, C. Shrubsole, P. Symonds, I. Mackenzie, and M. Davies, "Application of an indoor air pollution metamodel to a spatially-distributed housing stock," Science of the Total Environment, vol. 667, pp. 390-399, February 2019.
20. B. N. Vu, O. Sanchez, J. Bi, Q. Xiao, N. N. Hansel, W. Checkley, G. F. Gonzales, K. Steenland, and Y. Liu, "Developing an Advanced PM2.5 Exposure model in Lima, Peru," Remote Sensing, vol. 11, p. 641, March 2019.
21. G. Marques, and R. Pitarma, "A Cost-Effective Air Quality Supervision Solution for Enhanced Living Environments through the Internet of Things," Electronics, vol. 8, p. 170, February 2019.
22. M. D. Adams, and D. Corr, "A Mobile Air Pollution Monitoring Data Set," Data, vol. 4, p. 2, December 2018.
23. S. Dhingra, R. B. Madda, A. H. Gandomi, R. Patan, and M. Daneshmand, "Internet of Things Mobile – Air Pollution Monitoring System (IoT-Mobair)," IEEE Internet of Things Journal, March 2019.
24. A. S. Mihaita, L. Dupont, O. Chery, M. Camargo, and C. Cai, "Evaluating air quality by combining stationary, smart mobile pollution monitoring and data driven modelling," vol. 221, February 2019.
25. Y. Wu, Y. Yang, H. Nishiura, and M. Saitoh, "Deep Learning for Epidemiological Predictions," In SIGIR'18: The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, July 8-12, 2018, Ann Arbor, MI, USA, ACM, New York, NY, USA.
26. J. A. Stingone, O. P. Pandey, L. Claudio, and G. Pandey, "Using Machine learning to identify air pollution exposure profiles associated with early cognitive skills among US Children," Environmental Pollution, vol. 230, pp. 730-740, November 2017.
27. G. Polezer, Y. S. Tadano, H. V. Siqueira, A. F. L. Godoi, C. I. Yamamoto, P. A. de Andre, T. Pauliquevis, M. F. Andrade, A. Oliveira, P. H. N. Saldiva, P. E. Taylor, and R. H. M. Godoi, "Assessing the impact of PM2.5 on respiratory disease using artificial neural networks," Environmental Pollution, vol. 235, pp. 394-403, April 2018.
28. E. M. Mowry, A. K. Hedstrom, M. A. Gianfrancesco, X. Shao, C. A. Schaefer, L. Shen, K. H. Bellesis, F. B. S. Briggs, T. Olsson, L. Alfredsson, and L. F. Barcellos, "Incorporating machine learning approaches to assess to putative environmental risk factors for multiple sclerosis," Multiple Sclerosis and Related Disorders, vol. 24, pp. 135-141, August 2018.
29. M. J. Chen, P. H. Yang, M. T. Hsieh, C. H. Yeh, C. H. Huang, C. M. Yang, and G. M. Lin, "Machine Learning to relate PM2.5 and PM10 concentrations to outpatient visits for upper respiratory tract infections in Taiwan : A nationwide analysis," World Journal of Clinical cases, vol. 6, Issue 8, pp. 200-206, August 2018.
30. Z. Liu, K. Cheng, H. Li, G. Chao, D. Wu, and Y. Shi, "Exploring the potential relationship between indoor air quality and the concentration of airborne culturable fungi: a combined experimental and neural network modeling study," Environ Sci Pollut Res, vol. 25, Issue 4, pp. 3510-3517, February 2018.
31. Z. Liu, H. Li, and G. Cao, "Quik estimation model for the concentration of indoor airborne culturable bacteria: an application of machine learning," Int J Environ Res Public Health, vol. 14, Issue 8, p. 857, July 2017.
32. S. Chen, and S. Wu, "Deep learning for identifying environmental risk factors of acute respiratory diseases in Beijing, China: implications for population with different age and gender," International Journal of Environmental Health Research, March 2019.
33. S. Chen, L. Yang, S. Wu, and J. Li, "C4.5 classification-based quantitative analysis of risk factors for respiratory diseases,", Chin J Med Lib Inf Sci, vol. 25, Issue 8, pp. 35-41, 2016.
34. G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting, vol. 14, pp. 35-62, 1998.
35. H. Li, Z. Zhang, and Z. Liu, "Application of Artificial Neural Networks for Catalysis: A Review," Catalysts, vol. 7, Issue 10, p. 306, October 2017.
36. J. A. K. Suykens, and J. Vandewalle, "Least Squares Support Vector Machine Classifiers," Neural Processing Letters, vol. 9, Issue 3, pp. 293-300, June 1999.
37. Z. Liu, H. Li, X. Zhang, G. Jin, and K. Cheng, "Novel Method for Measuring the Heat Collection Rate and Heat Loss Coefficient of Water-in-Glass Evacuated Tube Solar Water Heaters Based on Artificial Neural Networks and Support Vector Machine," Energies, vol. 8, Issue 8, pp. 8814-8834, August 2015.
38. Z. Lu, K. Liu, H. Li, X. Zhang, G. Jin, and K. Cheng, "Artificial Neural Networks-Based Software for Measuring Heat Collection Rate and Heat Loss Coefficient of Wate-in-Glass Evacuated Tube Solar Water Heaters," PLos One, vol. 10, Issue 12, p. e0143624, December 2015.
39. D. B. Jani, M. Mishra, and P. K. Sahoo, "Application of artificial neural network for predicting performance of solid desiccant cooling systems - A review," Renewable and Sustainable Energy Reviews, vol. 80, pp. 352-366, December 2017.
40. M. Najarian, K. Snow, J. Lennon, and S. Kinsey, "Early childhood longitudinal study, birth cohort (ECLS-B)," Preschool-Kindergarten 2007 Psychometric Report, April 2010.
41. EPA Technology Transfer Network: Air Toxics Website. National Air Toxics Assessments, 2013.
42. USDA Rural-urban Continuum Codes, 2003.
43. S. Sra, S. Nowozin, and S. J. Wright, "Optimization for Machine Learning," MIT Press, 2011.
44. J. Zhang, "Gradient Descent based Optimization Algorithms for Deep Learning Models Training," arXiv:1903.03614v1 [cs.LG], March 2019.
45. J. Nocedal, and S. J. Wright, "Numerical Optimization," Springer-Verlag New York, Inc, 1999

AUTHORS PROFILE

J. Susymary, MCA, M.Phil., completed her Master's degree in Computer Applications and Master of Philosophy in Computer Science in Madurai Kamaraj University, Tamil Nadu, India. She is currently pursuing Ph.D. in Computer Applications in Kalasalingam Academy of Research and Education (KARE), Krishnankoil, Virudhunagar District, Tamil Nadu, India. Her research interest includes Graph Mining, Data Analytics, Machine Learning, Deep Learning She is also an IEEE student member.

Dr. P. Deepalakshmi is currently working as a Professor in Department of Computer Science and Engineering at Kalasalingam Academy of Research and Education (KARE), Krishnankoil, Virdhunagar District, Tamil Nadu, India. She is also serving as Dean, School of Computing. Her research interest includes Optimization Techniques, Network Routing, Distributed Computing, Network Security, Data Analytics, Machine Learning Techniques. She also takes care of KARE ACM student chapter as faculty mentor. Contact her at deepa.kumar@klu.ac.in.