OPEN ACCESS Freely available online

International Journal of Advancements in Technology

Research Article

Dynamic Ensemble Modelling for Prediction of Influenza Like Illnesses: A Framework

Samy Ghoneimy[1], Hossam M. Faheem[2], Noha Gamal[3*]

[1]British University in Egypt, Egypt; [2]Ain Shams University, Egypt; [3]Ahram Canadian University, Egypt

ABSTRACT

One of the advantages we have today in the fight against coronavirus (COVID-19) that wasn't as advanced in the SARS outbreak of 2003 is big data analytics and the major advancements in machine intelligence and artificial intelligence technologies. The United States of America's statistical surveillances have listed pneumonia/influenza as the seventh leading cause of death. Severe influenza seasons can result in more than 60,000 excess deaths and more than 200,000 hospitalizations. US witnessed fifty-five-thousand deaths (55,000 people) caused by pneumonia/influenza among total number of nine-hundred-thousand deaths (900,000 people) (%6.0)-during Influenza outbreak in 2018. Patients aged 65 years or older are at particular risk for death from viral pneumonia as well as from influenza not complicated by pneumonia. Deaths in these patients account for 89% of all pneumonia and/or influenza deaths. The healthcare industry needs researchers who are interested in applying machine learning for surveillance, prediction and diagnosis of diseases. Many healthcare-related researches, states that machine learning (ML) is the life-saving technology that will renovate healthcare services. This technology challenges the traditional reactive approach to healthcare. It is the predictive, proactive, and preventive life-saving qualities that make it a critically essential capability in every health system. In order to help in the prediction of pneumonia/influenza outbreaks, regression and classification techniques such as Ridge, Decision Tree Regression/Classification, Multiple Linear Regression, Logistic Regression Classification, K-Nearest Neighbor and Support Vector Machine Regression can be applied to predict forthcoming instances based on a trustworthy training and validation datasets. Accurate predictions will help healthcare stakeholders and governments to address the medical and physical needs during outbreak season. In this paper we exploit a methodology for predicting the number of deaths due to Influenza and Pneumonia in USA Cities using different machine supervised learning algorithms. Each algorithm is implemented, fitted to training dataset, validated by the validation dataset, and evaluated by means of Root Mean Square Error (RMSE) and R2 metric. KNN is the most fitted to the dataset by giving 92.6% accuracy. The least fitted algorithm is Logistic Regression by giving 51% accuracy. The remaining tested algorithms give accuracy levels from 80% to 92%. Evaluation Metrics, R2, and RMSE are obtained both analytically and programmatically using Python-based simulation. Results from both methods are well-matched. The promising results encourage the idea of enhancing the performance of the predictor. A new predictor (KMR-Stack) is implemented by integration of the best three fitted algorithms (KNN, Multiple Linear Regression, Ridge) in one stack. KMR-Stack exceeded KNN accuracy ratio by giving 94.9% accuracy. In KMR-

Correspondence to: N oha Gamal, Lectur er of Electronics and Communications Engineering, Ahram Canadian University , Egypt, E-mail: Noha.gamal@acu.edu.eg

Received: May 20, 2020; **Accepted:** May 27, 2020; **Published:** June 8, 2020

Citation: Ghoneimy S, Faheem HM, Gamal N (2020) Dynamic Ensemble Modelling for Prediction of Influenza Like Illnesses: A Framework. Int J Adv Technol 11:235. doi: 10.35248/0976-4860.20.11.235

Stack, another improvement was made in comparison with other stacking models introduced in the literature. The improvement included in the dynamicity of choosing the base-model regressors Hence, the stacked-integrated use of different machine learning algorithms showed increased prediction accuracies compared to the use of each individual algorithm, therefore improves influenza surveillance and potentially contributes in developing a robust defence strategy, which will collectively enhance human health.

Keywords: Healthcare; Machine learning; Dynamic; Stacking; Data analysis; Regresión; Classification; R2; RMSE; Ana lytical; Modelling; Decision tree; SVR; Logistic regression; Linear regression; KNN; Ridge

INTRODUCTION

The abrupt outbreak of coronavirus 2 (severe acute respiratory syndrome "SARS-CoV-2") has been leading universal population into a prominent crisis [1]. At present, healthcare organizations are in an urgent need for decision-making techniques to handle this virus and many other pandemics, that will help healthcare stakeholders in getting proper suggestions in real-time to avoid communicable diseases spread. Aspired by the huge advancement in computing reached by the twentieth century, the artificial Intelligence came into existence to imitate human brains in some information sciences domains [2]. That's whereby systems are developed to behave intellectually, reason rationally and have the flexibility to effectively interpret the surroundings in real time. A machine acting like a human has made it possible to simulate and solve many complex problems that need professional expertise. Hence, one of the widespread subfields of Artificial Intelligence became Machine Intelligence or Machine Learning (ML). Learning can be merely defined as the acquisition of knowledge or skills through a process of teaching, study, or experience. Although learning is an easy task for human beings, to attain new knowledge or skills from surrounding data, it is too hard and complicated process for machines. Furthermore, the intelligence level of a machine is directly proportional to its learning competence. The learning of algorithm helped the machine to understand a task from its experience. So, whenever the machine is able to make predictions from instances of desired behavior or past annotations and information, we can say that the machine learns. A more formal definition of machine learning was given by Tom Mitchell [3] " A computer program is said to learn from experience E regarding the class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Many ML algorithms were developed to run on different datasets and to solve specific problems. Most of these problems have some sort of history that helps assembling datasets. Dataset is the collection of experiences, obtained by performing tasks in the preceding experiences. The automatic ML algorithms learn from the preceding incidence to produce a new outcome or make predictions for forthcoming benefits [4].

Machine learning brought revolutionary solutions in prediction and classification problems by empowering the algorithms with data mining and big data analytics techniques. Machine learning integrate with various interdisciplinary fields like healthcare, games industry, home assistant, self- driving cars etc. In our research, we focus on the necessity of machine learning in healthcare which will aid in feature selection and in the processing of bulky and complex datasets. An analysis of the clinical disease datasets will help healthcare industry stakeholders to plan and provide sustainability in the process, leading to better outcomes, accurate inspection and diagnosis, lower costs of care, and increased patient satisfaction.

At present Big Data analysis is heading the research criteria of many disciplines that are now not directly associated to computer science, information or Mathematics [5]. Nevertheless, the real benefit of Big Data is no longer on the data itself, but in the ability to discover (unexpected) patterns and assemble information from it with excellent Data Science techniques [6]. There are several applications for Machine Learning (ML), the most significant of which is data mining/analysis. People are often susceptible to making mistakes during data analysis or, probably, when trying to discover associations between multiple features. This can make it considerably difficult for them to find solutions to certain problems. Machine learning can often be effectively applied to these problems, refining the efficiency of systems and the designs of machines [7].

In this paper we exploit a methodology for predicting the number of deaths due to Influenza and Pneumonia in USA Cities using different machine supervised learning algorithms (Ridge, Decision Tree Regression, Multiple Linear Regression, Logistic Regression, Support Vector Machine Regression). Each algorithm is implemented, fitted to training dataset, validated by the validation dataset, and evaluated by means of Root Mean Square Error (RMSE) and R2 metric. The promising learning results encourage the idea of enhancing the performance of the predictor. A new predictor (KMR-Stack) is implemented by integration of the best three fitted algorithms (KNN, Multiple Linear Regression, Ridge) in one stack. In KMR-Stack an improvement was made in comparison with other stacking models introduced in the literature. The improvement included in the dynamicity of choosing the base-model regressors. Our main goal in this research is to improve influenza surveillance and potentially contributes in developing a. defense strategy, which will collectively enhance human health. The remaining of the paper is organized as follows: Section II, illustrates the related work and literature review. The proposed framework is introduced in section III. Lastly, section IV, presents conclusion and future studies

RELATED WORK

Many researchers have worked on different machine learning algorithms for disease diagnosis, detection or prediction.

Researchers have been accepted that machine-learning algorithms work well in diagnosis of different diseases. In line with the research area of our paper, some of related proposed methodologies are explored as follows:

Shamshirband S et al. used Support Vector Machine and Firefly Algorithm (SVM-FFA) to predict malaria transmission to show which of the two has a better performance in prediction [8]. The work relates to malaria epidemy which is widespread in the state of Rajasthan leading to death and illness; lack of primary healthcare makes the situation worse. The four model systems designed were SVM-FFA, Auto-Regressive Moving Average (ARMA), Artificial Neural Networks (ANN) and SVM using LibSVM library in MATLAB. The R2 statistic and NMSE parameter were used for evaluating the performance of proposed algorithms, such that use of R2 gave accurate result for SVM-FFA in predicting malaria incidences. In conclusion, it was established that the novel approach of SVM-FFA is best among all the models.

David H Wolpert et al. surveyed many researches related to stacking in machine learning [9]. Stacking is an integrated technique that has been becoming very popular among the research community. Stacking is an efficient technique in which the predictions, generated from various different machine learning model, are used as inputs in a meta-learner, a second layer machine learning model. Unlike other ensemble combination rules [10], which are used by Sajid N et al. just to combine predictions of different models [11], the meta-learner learns how to combine the predictions at training level. Therefore, it provides a specific and unique way to combine predictions across multiple datasets and produce an efficient results without the need of tuning different ensemble combination rules. Practically, with some adjustments, stacking has shown to exceed other ensemble techniques from performance point of view. Saso D ˇzeroski et al. improves the performance by using multi-response model trees at the meta-learner level [12].

Kesorn K et al. introduced a surveillance system to monitor the effect of Dengue Hemorrhagic Fever (DHF) and Aedes aegypti mosquito infection rate, based on climate and geographical area using the Support Vector Machine (SVM) [13]. The nine major areas considered for a dengue epidemic rate were selected within the year 2007-2013. These areas are: temperature, rainfall, humidity, wind speed, Aedes aegypti larvae infection rate, a male mosquito infection rate, a female mosquito infection rate, population density, and morbidity rate. The method takes place in three stages. For the model construction, classification algorithms (like K-Nearest Neighbor (KNN), Decision Tree (DT), Neural Networks (NN), Support Vector Machine (SVM)) were used with different kernels. The 10-fold cross-validation technique was employed to validate the result for SVM effectiveness using the accuracy, sensitivity, and specificity as overall performance metrics. SVM-RBF kernel shows better performance with 96.296% accuracy, which is better among techniques such as SVM-L, SVM-P, KNN, DT, and NN.

Rane AL developed a survivability kit for prediction of some common epidemic diseases like ColdsFlu Gripe, Dengue, Malaria, Cholera, Leptospirosis, Chikungunya, Chickenpox, and Diarrhea [14]. To perform the study, data were collected from the hospital of Nasik, Maharashtra (India) from 316 patients. Algorithms like Decision Tree J48 (DT J48), Multi-layer Perceptron Neural Network (MLPNN), Support Vector Machine (SMO), K-Nearest Neighbor (LWL), and Naïve Bayes (NB) were assessed by 10-fold cross-validation and were implemented in WEKA software.

PROPOSED FRAMEWORK

In this paper we exploit a methodology for predicting the number of deaths due to Influenza and Pneumonia in USA Cities using different machine supervised learning algorithms (KNN, Ridge, Decision Tree Regression, Multiple Linear Regression, Logistic Regression, Support Vector Machine Regression). Each algorithm is implemented, fitted to training dataset, validated by the validation dataset, and evaluated by means of Root Mean Square Error (RMSE) and R2 metric. The promising learning results encourage the idea of enhancing the performance of the predictor. A new predictor (KMR-Stack) is implemented by integration of the best three fitted algorithms (KNN, Multiple Linear Regression, Ridge) in one stack. Our main goal in this research is to improve influenza surveillance and potentially contributes in developing a robust defense strategy, which will collectively enhance human health.

Data collection and preprocessing

Dataset (Raw data) of Deaths in 122 U.S. cities-2018 is referenced from data.cdc.gov/dataset. Each week, the vital statistics offices of 122 cities across the United States report the total number of death certificates processed and the number of those for which pneumonia or influenza was listed as the underlying or contributing cause of death by age group (Under 28 days, 28 days-1 year, 1-14 years, 15-24 years, 25-44 years, 45-64 years, 65-74 years, 75-84 years, and ≥ 85 years), with total number of incidences ~ 4300.

Table 1, declares the parameters included in the dataset, the data type of each parameter, and short description of each parameter. Figure 1 shows dataset visualization. Raw data mentioned above need preprocessing that includes various operations. Each operation aims to help machine learning build better predictive models as shown in Figure 2.

Table 1: Columns in this dataset.

Column name	DT	Description
Reporting area	Text	U.S. City Name
MMWR year	Num	Year number of the reported season
MMWR week	Num	Week number of the reported season
All causes, by age (years), All Ages	Num	Count of Deaths (all causes, all ages)

All causes, by age (years), ≥ 65	Num	Count of Deaths (all causes, age>=65)
All causes, by age (years), 45-64	Num	Count of Deaths (all causes, 64=>age>=45)
All causes, by age (years), 25-44	Num	Count of Deaths (all causes, 44=>age>=25)
All causes, by age (years), 1-24	Num	Count of Deaths (all causes, 24=>age>=1)
All causes, by age (years), LT 1	Num	Count of Deaths (all causes, age<1)
P and I Total	Num	Count of Deaths (Cause=influenza and Pneumonia)

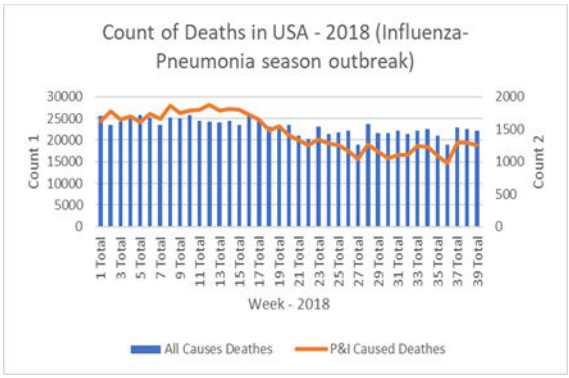

Figure 1: Dataset visualization.

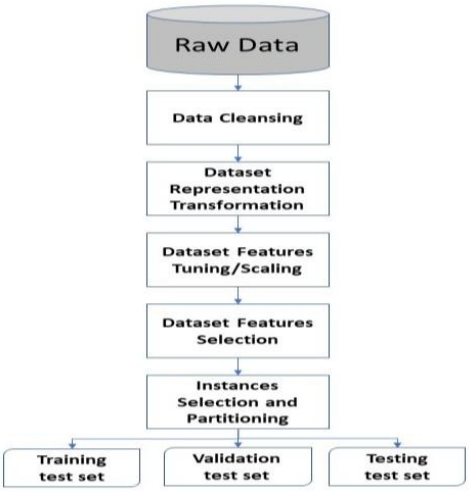

Figure 2: Data preprocessing operations (structured data).

Proposed model description

Proposed model consists of three main stages. First stage is the data preprocessing resulting two data sets (training and testing). Second stage is the implementation of six individual regression algorithms, KNN, Ridge, Decision Tree Regression, Multiple Linear Regression, Logistic Regression, Support Vector Machine Regression.

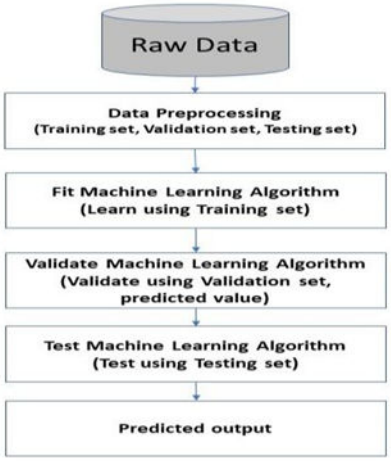

Figure 3: Proposed model's two stages processes.

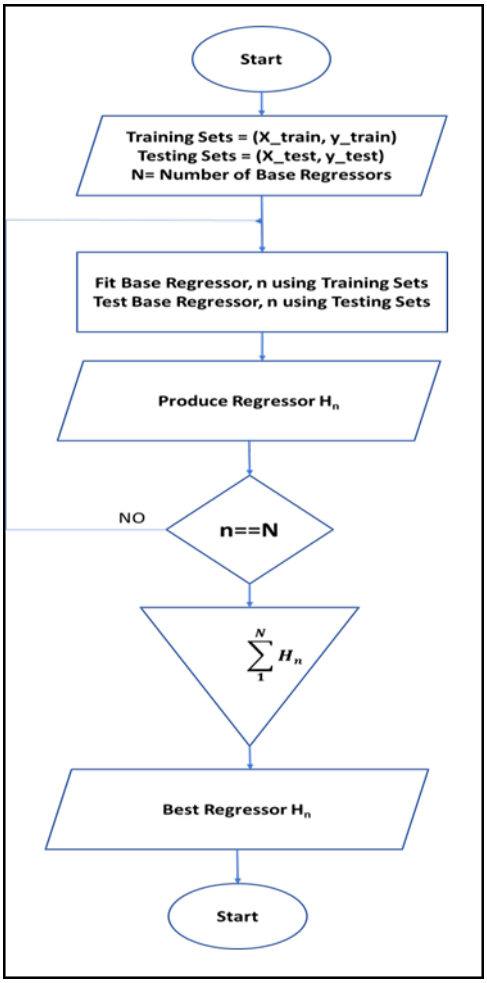

Figure 4: Stacking of N progressors for a combined one base model.

The output of second stage is the evaluation metrics of each algorithm individually. First and second stages processes are summarized in Figure 3, taking into consideration that second stage contains six individual regressors needs to be fitted to training dataset to generate the dynamic selection of base model regressors, will be used in final stage. Last stage considers a Stacking process of regression algorithms give the best

evaluation metrics in second stage to create stacking ensemble function.

Stacking is an integrated learning technique that combines multiple regression models via a meta-regressor. The base level models are trained based on a complete training set, then the meta-model is trained on the outputs of the base level model as features. The base level often consists of different learning algorithms and therefore stacking ensembles are often heterogeneous. The flowchart shown in Figure 4, summarizes stacking of N progressors to create a combined one base model for first level of regression. A second level progressor should be applied to give a final prediction of the integrated stack.

Machine learning techniques

The pioneer of Artificial Intelligence, Arthur Samuel, who devised the term machine learning, quoted that, "machine learning, as a way of programming, gives the computer the ability to learn" [15]. Machine learning is categorized into three types, namely supervised learning, unsupervised learning, and reinforcement learning.

- Build a mathematical model of a set of data that contains both the inputs (parameters or features) and the desired outputs (dependent outcome) [16]. The data is known as training data, and consists of a set of training examples. Each training instance has one or more inputs and a desired output. In the mathematical model of supervised learning, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Using iterative implementation, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs [17]. optimum learning will allow the algorithm to accurately determine the output for inputs that were not a part of the training data (test data). An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task [18,19]. There are two groups of algorithms in supervised learning, Classification and Regression. The main difference between them is that the output variable in regression is numerical (or continuous) while that for classification is categorical (or discrete). In our research, we will get use of many supervised learning algorithms to predict the number of deaths due to Influenza and Pneumonia in USA Cities during seasonal outbreaks
- Receive a set of data that contains only inputs (independent parameters or features), and find structure in the data, like grouping or clustering of data points. The algorithms therefore learn from test data that has not been labeled, classified or categorized. Uunsupervised learning algorithms recognize commonalities in the data and respond in consideration of the presence or absence of such commonalities in each new piece of data [20]
- Is a part of human interactive psychology, which uses an agent to act according to the circumstances towards maximizing the rewards. The reinforcement learning goals work by setting explicit goals; it works by sensing the environment. Applications of reinforcement learning are vast, and it is used mostly in game development, manufacturing, inventory management, delivery management, power system, finance sector [20,21]

Supervised learning algorithms-regression

Regression Analysis is a statistical process for assessing the relationships between the dependent variables or criterion variables and one or more independent variables or predictors. Regression analysis explains the changes in criterions in relation to changes in select predictors. The conditional expectation of the criterions based on predictors where the average value of the dependent variables is given when the independent variables are changed. Three major uses for regression analysis are determining the strength of predictors, forecasting an effect, and trend forecasting. Types of regression we will assess in our research are: KNN, Ridge, Decision Tree Regression, Multiple Linear Regression, Logistic Regression, Support Vector Machine Regression are explained below [22]:

- It is a technique for analyzing multiple regression data. When multicollinearity occurs, least squares estimates are unbiased. A degree of bias is added to the regression estimates, and a result, ridge regression reduces the standard error
- It uses the relationship between two sets of continuous numerical measures. The first set is called the predictor or independent variable. The other is the response or dependent variable. The goal of linear regression is to identify the relationship in the form of a formula that defines the dependent variable in terms of the independent variable. Once this relationship is quantified, the dependent variable can be predicted for any instance of an independent variable
- It sounds similar to linear regression but is actually focused on problems involving categorization instead of quantitative forecasting. Here the output variable values are discrete and finite rather than continuous and with infinite values as with linear regression. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable=response or outcome variable) and a set of independent (predictor or explanatory) variables. The output of logistic regression is a value between 0 and 1. Results closer to 1 indicate that the input variable more clearly fits within the category. Results closer to 0 indicate that the input variable likely does not fit within the category
- Trees use a decision to categorize data. Each decision is based on a question related to one of the input variables. Wit h each question and corresponding response, the instance of data gets moved closer to being categorized in a specific way. This set of questions and responses and subsequent divisions of data create a tree-like structure. At the end of each line of questions is a category
- is also a classification and regression algorithm. The learning process is composed of the training set of data being stored. A s new instances are evaluated, the distance to each data point in the training set is evaluated and there is a consensus decision as to which category the new instance of data falls into based on its proximity to the training instances. This categorization algorithm allows for multivalued categorizations of the data (the "k" is the number of neighbors it checks)

- Support vector machine (SVM): is a discriminative classifier which can be used for both classification and regression problems. The goal of SVM is to identify an optimal separating hyperplane which maximizes the margin between different classes of the training data. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples to create the largest possible distance to reduce an upper bound

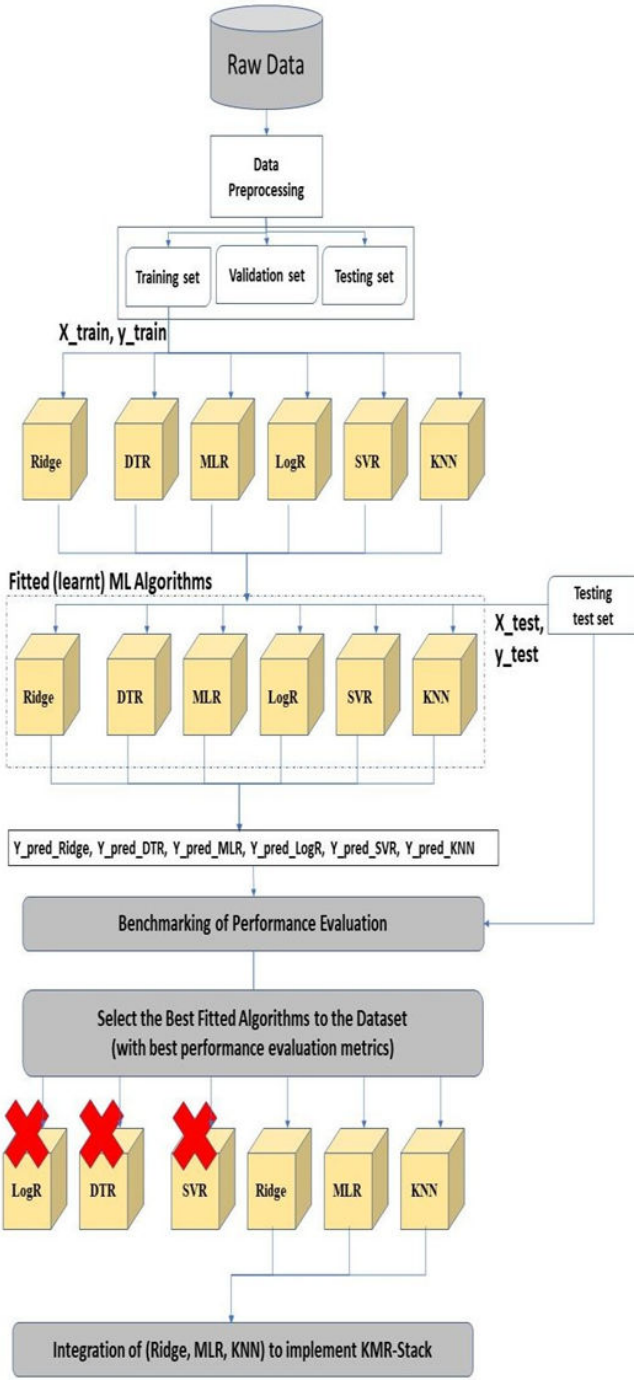

Figure 5: Detailed proposed franework architecture (first, second stages).

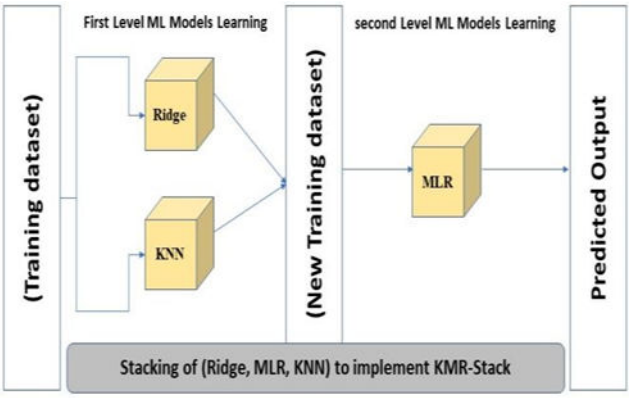

Figure 6: Stacking ensemble function architecture (Third stage).

Basic and enhanced models architectures

For easier model implementation, we divided the system into two separate functions, Base Model Function and Stacking Ensemble Function. Base Model function includes the implementation of first and second stages as shown in Figure 3 and detailed in Figure 5.

Stacking Ensemble Function includes the implementation of third stage as shown in Figure 6.

Regression and ensemble regression analysis

The model specification in linear regression is that the dependent variable, y_i is a linear combination of the parameters. There is one independent variable xi and two parameters β_0 and β_1 in the simple linear regression. For modeling n data points in more generalized linear regression (Multiple Linear Regression) the following analysis can be used:

$$yi = 1xi1 + \hat{\beta 1}xi1 - \hat{\beta pxip}, i = 1, \ldots, n \quad \ldots\ldots\ldots\ldots (1)$$

$$\varepsilon i = yi - \hat{\beta 1}xi1 - \hat{\beta pxip}, i - 1, \ldots, n \quad \ldots\ldots\ldots\ldots (2)$$

$$\sum_{i=1}^{n}\sum_{k=1}^{p} x_{ij}x_{ik}\beta_k = \sum_{i=1}^{n} x_{ij}y_{i,\,j=1,\ldots,p} \quad \ldots\ldots\ldots\ldots (3)$$

$$\text{Or, } (XTS)\hat{\beta} = (XTY) \quad \ldots\ldots\ldots\ldots (4)$$

$$\hat{\beta} = (XTY)(XTX) - 1 \quad \ldots\ldots\ldots\ldots (5)$$

$$\hat{Y} = X\hat{\beta} = (XTY)(XTX) - 1 = HY \quad \ldots\ldots\ldots\ldots (6)$$

$$ei = yi - \hat{y}_i \quad \ldots\ldots\ldots\ldots (7)$$

$$\sigma 2 = MSE = \frac{\sum e_i^2}{n-p-1} \quad \ldots\ldots\ldots\ldots (8)$$

$$R2 = \frac{\sigma\hat{y}^2}{\sigma y^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad \ldots\ldots\ldots\ldots (9)$$

Where y_i is the ith observation in the dependent variable vector Y, x_{ip} is the $(i,p)^{th}$ element in the independent variable matrix X, β_p is the p^{th} parameter in coefficient matrix β. The observation forecast is calculated as in (6), The error mean square MSE is an estimate of the variance σ, of the random error terms, e_i. R^2 is a quantity that measures how much of the variance in y is explained by the model, $\hat{Y}$. Under "general conditions", R2 is also the square of the correlation.

Instead of a single prediction, an ensemble regression is made up of a group of interrelated predictions all associated with one observation. This constrains the statistics of the ensemble as set out in the following system of relationships.

$$((\hat{y} - Y)^2) = (E2) + ((\hat{y} - Y)^2) \text{..... (10)}$$

The mean squared error of the individual ensemble members, $\hat{Y}_\tau$ corresponds to the distribution of the ensemble and the squared error in the mean of the ensemble $\hat{Y}_m$. Where E^2 is the ensemble spread mean.

$$(E2) = (\frac{1}{N}\sum_{\tau=1}^{N}(\hat{y}_\tau - \hat{y}_m)^2) \text{..... (11)}$$

The sample variance of the individual ensemble forecasts, σ_τ^2, can be related to the mean ensemble spread and variance of the ensemble mean, σ_m^2, by:

$$\sigma_\tau^2 = \sigma_m^2 + (E^2) \quad \text{.....(12)}$$

$$R^2 = R\tau^2 \frac{\sigma_\tau^2}{\sigma_m^2} \quad \text{.....(13)}$$

A prediction of an ensemble is often seen as a set of potential states from a given initial state. One of the various solutions will be "The Best Solution". Normally, each ensemble member is assumed to have an equal chance to be "the best". Without actually identifying a best member, Yb, we can speculate that it is directly linked to the real observation Y by:

$$Y = \beta_0 + \beta_1 Y_b + \varepsilon_b \text{ ..... (14)}$$

Following the assumption that, any ensemble member Y_τ of N ensemble members is likely to be "the best" regressor equally with all other ensemble members, then for any given observation j on M-sample size the expected value of Yb can be computed using $\hat{Y}_\tau$ as follows:

$$EXPVAL(Yb) = (\frac{1}{N}\sum_{\tau}^{N}\hat{y}_\tau = \hat{y}_m) \text{ ..... (15)}$$

The mean of Yb over M sample size can be obtained as follows:

$$EXPVAL(Yb) = (\frac{1}{M}\frac{1}{N}\sum_{j=1}^{M}\sum_{\tau=1}^{N}Y_{\tau,j} = (B) \text{..... (16)}$$

$$(EX(\sigma 2b)) = \frac{1}{M}\frac{1}{N}\sum_{j=1}^{M}\sum_{\tau=1}^{N}(Y_{\tau,j} - B)^2 = \sigma_\tau^2 \text{ .. (17)}$$

For simplicity,

$$(EXP(\sigma 2b)) = \sigma 2b \text{ ..... (18)}$$

Using the same transformations as in equations (9), (13):

$$Rb = \frac{R_m^2}{R_\tau} \text{ ..... (19)}$$

RESULTS AND DISCUSSION

In this paper a robust model for predicting the number of deaths due to Influenza and Pneumonia in USA Cities was implemented and evaluated by using different machine supervised learning algorithms. Raw data with the total number of Deaths in 122 U.S. cities during Influenza and Pneumonia outbreak season in 2018 is referenced from data.cdc.gov/dataset. Total number of death certificates processed is reported on weekly basis, and the number of those for which pneumonia or influenza was listed as the underlying or contributing cause of death by age group (Under 1 year, 1-24 years, 15-24 years, 25-44 years, 45-64 years, and ≥ 65 years). Total number of instances remained in the dataset after passing the data preprocessing stage is ~ 4300 records. The dataset was divided up in an 8:2 ratio and each part was used for constructing the regression model and prediction respectively.

Each algorithm is implemented, fitted to training dataset, tested by the testing dataset, and evaluated by means of Root Mean Square Error (RMSE) and R2 metric.

R2 score varies between 0 and 100%. It is the proportion of the variance in the dependent variable that is predictable from the independent variable(s) or, the total variance explained by model)/total variance. So, if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value of R2 illustrates a low level of correlation, meaning a regression model that is not valid or not fitting to our problem.

Mean Square Error (MSE) is the average of the square of the errors. The larger the number the larger the error. Error in this case means the difference between the observed values y_1, y_2, y_3, etc. and the predicted ones pred (y_1), pred (y_2), pred (y_3), etc. each difference is squared $(\text{pred}(y_n)\text{-}y_n))^2$ so that negative and positive values do not cancel each other out.

Evaluation Metrics, R2, and RMSE are obtained both analytically using the equation system listed eqn. (1) to eqn. (19) and programmatically using Python-based simulation. Results from both methods are well-matched as shown in Table 2. KNN is the most fitted to the dataset by giving 92.6% accuracy. The least fitted algorithm is Logistic Regression by giving 51% accuracy. Logistic Regression results were expected. It was only tested to proof the validity of dataset used, by being not suitable for prediction by means of logistic regression. The remaining tested algorithms give accuracy levels from 80% to 92% as shown in the Table 2.

After getting evaluation metrics of our individual regressors, an enhanced approach was introduced by implementing a stacking. Stacking is to learn several different regressors (base-model) and combine them by training a metamodel to output predictions based on the multiple predictions returned by these base models. A new predictor (KMRStack) is implemented by integration of the best three fitted algorithms (KNN, Multiple Linear Regression, Ridge) in one stack enhancing the performance of the predictor.

KMR-Stack exceeded KNN accuracy ratio by giving 94.9%. Hence, the stacked-integrated use of different machine learning algorithms showed increased prediction accuracies compared to the use of each individual algorithm, therefore improves influenza surveillance and potentially contributes in developing a robust defense strategy, which will collectively enhance human health. Results show that, regarding the subjected problem with

the given dataset, the most appropriate Machine Learning Algorithms are KNN, Ridge and MLR respectively.

Table 2: Evaluation metrics (analytical vs. simulated)

Regression Algorithm	R2 Value (Simulated)	RMSE Value (Simulated)	R2 Value (Analytically calculated)	RMSE Value (Analytically calculated)
K-nearest neigbour (K=11)	0.926966	6.07093	0.932	5.950
Ridge	0.920198	6.34592	0.914	6.282
Multiple linear regression	0.917707	6.44429	0.921	6.277
Decision tree regression	0.881933	7.77189	0.750	6.762
Support vector regression	0.817007	9.60973	0.613	8.457
Logistic regression	0.512966	15.6773	0.446	12.072
KMR-stack (proposed)	0.94952	5.63292	0.959	5.577

Hereafter, a graphical visualization of all tested regression algorithms is presented in Figures 7 to 13 illustrates a visualization of a scatter plot for the total number of Deaths due to Influenza and Pneumonia in relevance to the total number of Deaths due to all causes. Each figure has two scatters the red scatter is for the real values of training/testing sets and the blue/yellow scatter is for the learned (in case of training set)/predicted (in case of testing set):

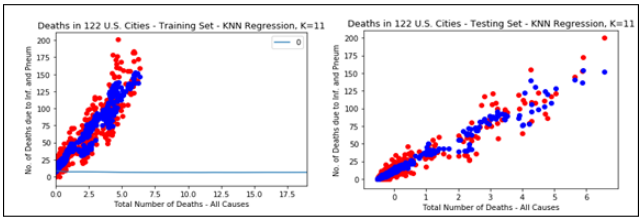

Figure 7: KNN Visualizations.

- Ridge training and testing results visualization

Figure 8: Ridge visualizations.

- MLR training and testing results visualization

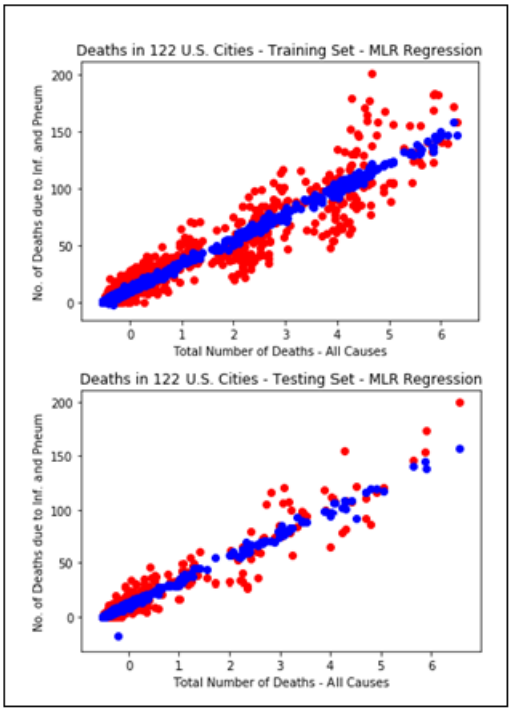

Figure 9: MLR visualizations.

- DTR training and testing results visualization

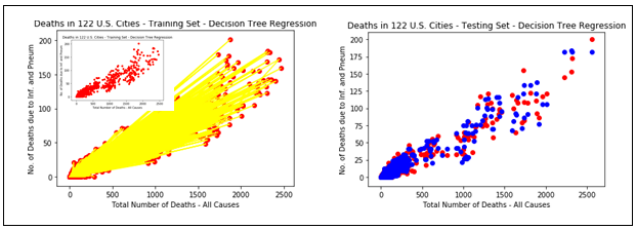

Figure 10: DTR visualizations (Training actual and learned values are identically matched).

• SVR training and testing results visualization

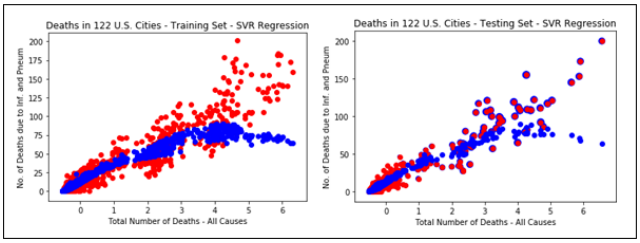

Figure 11: SVR visualizations.

• LogR training and testing results visualization

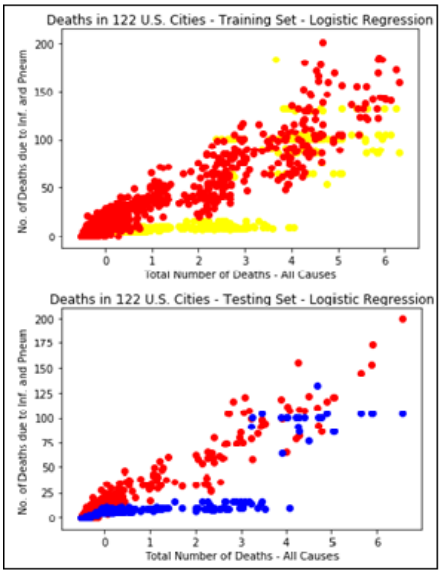

Figure 12: LogR visualizations.

• KMR-stack training and testing results visualization

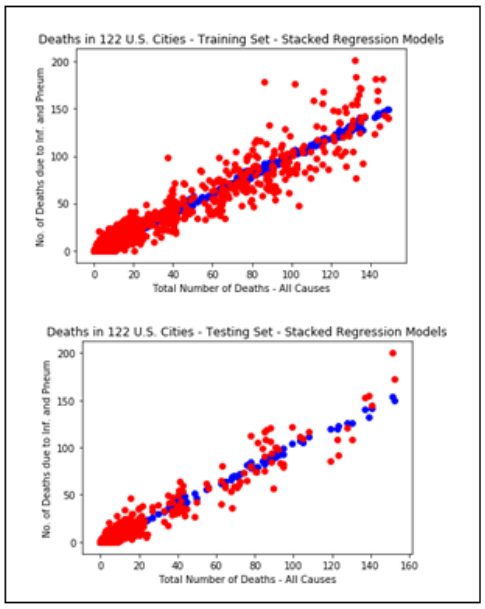

Figure 13: KMR-stack visualizations.

CONCLUSION AND FUTURE STUDIES

Annually, the US healthcare systems produces nearly one trillion GBytes of data. These remarkable quantities of data have been accompanied by an increase in cheap, largescale computing power. Together, they elevate the possibility that artificial intelligence and machine learning, in particular can generate insights that can recognizably improve the whole healthcare industry. We see, with machine learning applications, healthcare and medicine segment can advance into a new era and completely transform the healthcare operations. In this study, the machine learning techniques was used for selecting the most significant features to be utilized in predicting the number of deaths due to Influenza and Pneumonia. US witnessed fifty-five-thousand deaths (55,000 people) caused by pneumonia/influenza among total number of nine-hundred-thousand deaths (900,000 people) (%6.0)-during Influenza outbreak in 2018. According to our paper, machine learning algorithms namely MLR, KNN, SVM, Ridge, DTR, and LogR were applied to measure the performance evaluation while predicting the number of deaths due to Influenza and Pneumonia. The experiment results show that KNN, Ridge, MLR achieved the best accuracy rates by giving R2 scores of (0.926, 0.92, 0.917) respectively, that was also validated as evaluation Metrics, R2, and RMSE are obtained both analytically and programmatically using Python-based simulation. Results from both methods are well-matched. The enhanced KMR-Stack produced R2 score of 0.949 with more than 2% increase in prediction accuracy. In KMRStack an improvement was made in comparison with other stacking models introduced in the literature. The improvement included in the dynamicity of choosing the base-model regressors. An algorithm was implemented to choose which models to be trained and used in the stacking base model regressors by measuring the performance evaluation of six different regressors then select the best evaluated three progressors to implement stacking model. More complex datasets integration will be introduced and more advanced classification techniques will be evaluated and enhanced in our future research.

REFERENCE

1. Roy AN, Jose J, Sunil A, Gautam N, Nathalia D, Suresh A. Prediction and spread visualization of COVID-19 pandemic using machine learning. Preprints. 2020;2020050147.
2. Vaishya R., Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr: Clinical Research and Reviews. 2020;14:337-339
3. Mitchell Tom M. Machine learning and data mining. Communications of the ACM. 1999;42.
4. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nature Methods. 2018;15:233-234.
5. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Quarterly. 2012;36:1165-1188.

6. Ramírez-Gallego S. A distributed evolutionary multivariate discretizer for big data processing on apache spark. Swarm Evol Comput. 2018;38:240-250.
7. Char DS, Shah NH, Magnus D. Implementing machine learning in health care addressing ethical challenges. N Engl J Med. 2018;378:981-983.
8. Shamshirband S, Mohammadi K, Tong CW. A hybrid SVM-FFA method for prediction of monthly mean global solar radiation. Theor Appl Climatol. 2016;125:153.
9. David HW. Stacked generalization. Neural Networks. 1992;5:241-259.
10. Pierre G, Damien E, Louis W. Extremely randomized trees. Machine Learning. 2006;63:3-42.
11. Sajid N, Dhruba KB. Classification of microarray cancer data using ensemble approach. Netw Model Anal Health Inform Bioinforma. 2013;2:159-173.
12. Saso DZ, Bernard Z. Is combining classifiers with stacking better than selecting the best one? Machine Learning. 2004;54:255-273.
13. Kesorn K, Ongruk P, Chompoosri J, Phumee A, Thavara U, Usavadee T, et al. Morbidity rate prediction of dengue hemorrhagic fever (DHF) using the support vector machine and the Aedes aegypti infection rate in similar climates and geographical areas. PloS One. 2015;10:e0125049.
14. Rane AL. Clinical decision support model for prevailing diseases to improve human life survivability. International Conference on Persasive Computing, IEEE. 2015:1-5.
15. Samuel AL. Some studies in machine learning to use the game of checkers. IBM J Res Dev. 1959;3:210-229.
16. Russell SJ, Norvig P. Artificial intelligence: a modern approach (Third ed.). Prentice Hall, 2010.
17. Mohri M, Rostamizadeh A, Talwalkar A. Foundations of machine learning. The MIT Press, 2012.
18. Alpaydin E. Introduction to machine learning. MIT Press, 2010.
19. Alex R, Stephen B, Paroma V, Chris Ré, other members of Hazy Research. Weak supervision: the new programming paradigm for machine learning, 2019.
20. Jordan MI, Bishop CM. "Neural Networks". In Allen B. Tucker (ed.). Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, Florida: Chapman and Hall/CRC Press LLC, 2004.
21. Van Otterlo M, Wiering M. Reinforcement learning and markov decision processes. Reinforcement Learning. 2012;12:3-42.
22. Amin M, Ali A. Performance evaluation of supervised machine learning classifiers for predicting healthcare operational decision. c-section classification database report, uci machine learning repository, University of California, Irvine, USA. 2018.