Prediction of Death Counts Based on Short-term Mortality Fluctuations Data Series using Multi-output Regression Models

Md Imtiaz Ahmed*, Nurjahan†, Md. Mahbub-Or-Rashid‡, and Farhana Islam§
*Department of Computer Science and Engineering, Prime University
†Department of Internet of Things and Robotics Engineering, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh
‡Department of Computer Science and Engineering, Bangladesh University of Business and Technology
§Department of Educational Technology, Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

***Abstract*—Effective public health responses to unexpected epidemiological hazards or disasters need rapid and reliable monitoring. But, monitoring fast-changing situations and acquiring timely, accurate, and cross-national statistics to address short-term mortality fluctuations due to these hazards is very challenging. Estimating weekly excess deaths is the most solid and accurate way to measure the mortality burden caused by short-term risk factors. The Short-term Mortality Fluctuations (STMF) data series is one of the significant collections of the Human Mortality Database (HMD) that provides the weekly death counts and rates by age and sex of a country. Sometimes, the data collected from the sources are not always represented in specific age groups rather represented by the the total number of individual death records per week. However, the researchers reclassified their dataset based on the ranges of age and sex distributions of every country so that one can easily find out how many people died in per week of each country based on an equation and earlier distribution data. The paper focuses on the implementation of multi-output regression models such as logistic regression, decision tree, random forest, k nearest neighbors, lasso, support vector regressor, artificial neural network, and recurrent neural network to correctly predict death counts for specific age groups. According to the results, random forest delivered the highest performance with an R squared coefficient value of 0.9975, root mean square error of 43.2263, and mean absolute error of 16.4069.**

Keywords*—Multi-output regression model; short-term mortality fluctuations; machine learning; deep learning

I. Introduction

In the past few years, there have been many outbreaks of natural or man-made hazards which eventually turned into a pandemic situation. For instance, influenza outbreaks in 2014–15, 2016–17, and 2017–18, as well as the recent COVID-19 pandemic. These hazards induced significant increases in short-term mortality in several countries [1]. Accurate and statistical data is important to analyze the mortality rates and to provide an immediate response to short-term health concerns for reducing life loss. However, the recent COVID-19 pandemic pointed out the scarcity of reliable, accurate, and comparable international data required to track the spread of epidemics [2]. In May 2020, the Human Mortality Database (HMD, [3]) team released the Short-term Mortality Fluctuations (STMF) data series to meet the increasing need for such data. The information on how many people died in a calendar year has been kept in this dataset on a weekly basis. However, the researchers built their dataset on age-specific deaths in each country so that one can find out how many children, youth, or adults die in each country per week. In many cases, researchers have already stated that they cannot get accurate data into different ranges of ages but they can get the total death number of a city, a country, or a state. To mitigate these problems, researchers normally use the below Eq. 1 for distributing the total number of deaths to age-specific numbers. They use the equation 1 and use the earlier distribution that they already deposited into the database. However, the observed or forecasted death counts from annual age-specific groups are then converted to standard age groups using the following formula:

$$\hat{M}_b^s(x, x+m) = M_b^s(x, x+n) * \frac{M_b(x, x+m)}{M_b(x, x+n)} \tag{1}$$

In the above equation, $M_b^s(x, x+n)$ indicates the number of death according to the original data in the interval of age $[x, x+n)$ in s week of year b and n is the age interval length of original data. On the other hand, $\hat{M}_b^s(x, x+m)$ indicates the predicted number of death in the interval of age $[x, x+m)$ in s week of b year and m is the age interval length of estimated data. $M_b(x, x+m)$ and $M_b(x, x+n)$ represents the number of death in the whole year b.

Similar to the age specific distributions, they have calculated the specific groups based on sex by using the annual data stated in the following Eq. 2 when the age-specific sex group data is unavailable.

$$\hat{M}_b^{s,males}(x, x+m) = M_b^{s,total}(x, x+m) * \frac{M_b^{males}(x, x+m)}{M_b^{total}(x, x+m)} \tag{2}$$

The death rate according to the age groups has been estimated using total number of death in s week of b year and total population $P_b(x, x+m)$ of specific age groups using the following Eq. 3:

$$R_b^s(x, x+m) = \frac{M_b^s(x, x+m)}{P_b(x, x+m)/52} \quad (3)$$

However, The accuracy of the prediction of weekly age or sex group specific data from the combined data using the above distribution equation is not calculated or proved. As an efficient and easy alternative to the equation to solve the problem, we have proposed a system that uses several multi-output regression models. Compared to developing separate five single-output models for predicting five output features, multi-output regression has multiple advantages. Multi-output regression provides reduced training time, a unified prediction rule, and improved predictive generalization. As a result, much more complicated decision-making problems can be solved easily [4]. In this work, the prime objective of this research is to propose a model using multioutput regression to correctly perform the prediction of mortality data based on combined weekly mortality data. After collecting the dataset, we applied six ML models as well as two DL models named Linear Regression (LR), Decision Tree (DT), Random Forest (RF), K-nearest Neighbour (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Regressor (SVR), Artificial Neural Network (ANN) and Recurrent Neural Network (RNN). After then, we have compared the output of each model. Finally, we have explored the best-performing model for the problem.

The remainder part of the paper is organized as follows: the recent relevant works is shown on Section II, the materials and methods is described on Section III, the result of the experiment is shown on Section IV, discussion is demonstrated on Section V and finally the conclusion and future works on Section VI.

II. Related Works

Some researchers already exploited the benefit of multi-output regression in their work. For example, in [5], Cui et al. jointly predicted two healthcare resource utilization measures such as length of stay and cost using multi-output regression models. They used four regression models such as NN, DT, RF, and multi-task Lasso for the prediction. They have achieved best performance with RF model when features generated through skip-gram feature vectors according to the R^2 coefficient, RMSE, Mean-Absolute error (MAE), Median Absolute Error (Median-AE) among the uninterpretable methods. Boumezoued et al. [6] utilized linear regression and neural network model to the correction of the mortality data while birth by month data is not available. They worked on the database of human mortality. In [7], Shahid et al. used seven regression models including decision tree, random forest, linear regression, support vector regression, ridge regression, gradient boosting, and multi-layer perceptron to efficiently forecast road traffic flow. Before implementing the models, they have utilized five dimensionality reduction methods. Han et al. [8] applied multi-output least square support vector regressor (M-LSSVM) to predict the levels of gas in a multi-tank LDG system in real time. It encompasses both the individual fitting errors as well as the combined ones for each output. Tuia et al. [9] employed an multioutput support vector regressor (MSVR) model to estimate biophysical parameters such as fractional vegetation cover, chlorophyll content, and leaf area index from remote sensing images in a simultaneous manner. The study demonstrated that M-SVR is a viable substitute for nonparametric estimation of biophysical parameters and model inversion, compared to the single-output regression method. Li et al. [10] developed a system that utilizes multi-target regression models to predict the time series value of blood-drug efficacy in traditional chinese medicine datasets. The proposed system utilized the correlation between targets to enhance the performance of four learning techniques such as LR, Partial Least Squares, SVR, and ANN. SVR exhibits the best performance among the applied models. Meyer et al. [11] investigated the use of multi-target machine learning models for wind turbine normal behavior monitoring. The authors assessed 6 multi-target models such as DT, RF, KNN, MLP, CNN, and LSTM in a wind turbine case study and found that these models offer benefits over single-target modeling. Specifically, multi-target models can significantly decrease the effort required for the lifecycle management of normal behavior models while maintaining model accuracy. Kucuk et al. [12] predicted soil moisture through applying nine multi-output regression models such as LR, ridge regression, Lasso, RF, adaptive boosting, extreme gradient boosting, gradient boosting, histogram-based gradient boosting and extra tree regressor (ETR). They have shown that ETR delivers best performance with 0.81 r-squared coefficient value.

III. Methodology

The whole procedure has been subdivided into several parts such as dataset description, data preprocessing, implementation of multi-output regression models and finally the comparison of the performance of the algorithms. The conceptual flow of the procedure has been demonstrated on Fig. 1. At first, we have gathered the dataset in csv format. The data values has been scrubbed with the necessary features, and the final dataset has the features named Country Code, Year, Week, Sex, D_Total, D0_14, D15_64, D65_74, D75_84, and D85p. To handle multiple target features, we have utilized several regression models that perform better in multi-output regression problems such as LR, DT, RF, KNN, Lasso, SVR, ANN, and RNN. All the implementation were performed on python. Finally, the the performance of the model has been evaluated based on performance metrics.

A. Dataset Description

The STMF data series, a part of HMD, contains the records of human mortality rate according to every week of a year. The data has been stored both in the csv and excel file formats. There are total 19 features in the dataset such as CountryCode, Year, Week, Sex, next five features (5-9) includes death counts by age group (0-14, 15-64, 65-74, 75-84, 85+), total death counts through combining all age groups, next five features (11-15) such as death rates by age group (0-14, 15-64, 65-74, 75-84, 85+), total death rates through combining all sex groups, finally 17-19 attributes are explanatory indicators such as split, splitsex and forecast (see Table I). The four columns of the dataset are country in ISO-3 code format, year, week, sex of the the people who has died. It maintains the guidelines of ISO 8601-2004 to arrange the week. Generally, a year is

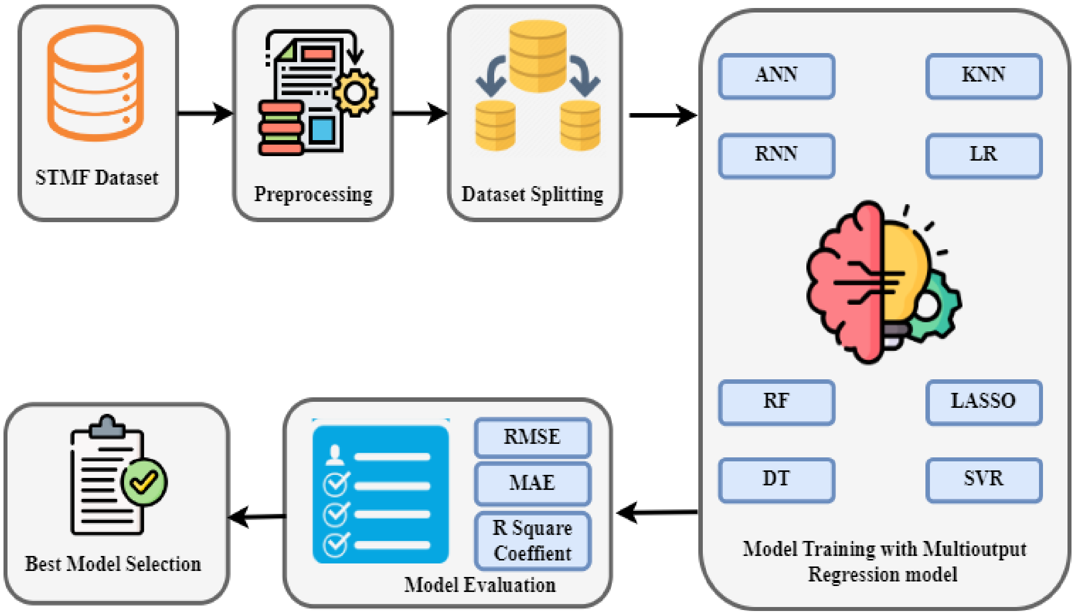

Fig. 1. Conceptual flow of the proposed model.

divided to 52 weeks except for some years of 53 weeks like 1992, 1998, 2004, 2009, 2015, and 2020. However, this paper focuses on the dataset with ten features where the five features such as country code, year, week, sex, D_Total were used as independent features and the five features such as D0_14, D15_64, D65_74, D75_84, and D85p were used as dependent features. On the other hand, there are 107211 records in the table.

B. Data Preprocessing

Data preprocessing is performed to remove any abnormalities in the dataset, identifying milling values as well as to prepare the data for further analysis. In this work, the dataset was checked whether it has any null values or not and it was replaced with zero using the fillna() option of python. After that, the character values changed using the encoder of python as the character or word data cannot be used for the application.

C. Multi-output Regression Models

Multi-output regression involves concurrently predicting multivariate output feature space from a given multivariate input feature space [13], [14]. Suppose $a \in R^u$ is a u-dimensional input feature space and $b \in R^v$ is a v-dimensional output feature space. So, multi-output regression can be stated as mapping from R^u to R^v [15]. In this work, we simultaneously predicted five output features using multi-output regression models. We have implemented eight different regression models namely ANN, RNN, LR, DT, RF, KNN, Lasso, and SVR to the data.

1) Artificial Neural Network (ANN): ANN [16]is a computational network, which is motivated by the structure and function of biological neural networks in the brain [17]. The neural networks have several's neurons that are interconnected to each layer named as nodes similar to biological neural networks. The basic objective is to simulate the neural network that makes up the human brain so that computers can be capable of comprehending information and making decisions in the same way humans do. There are major three layers of ANN.

- Input Layer: The input layers receive input in various formats from multiple sources provided by researchers. Inputs are provided in the form of a pattern and vector from those external sources.
- Hidden Layer: The hidden layer lies in the middle of the input and output layers. This layer extracts all hidden features and patterns based on a given weight.
- Output Layer: Each input is multiplied by its associated weight. If the summed-up weighted input is zero then a bias is added to make a non-zero or different output. After that an activation function is applied to the summed-up weighted inputs to get the desired output. The Eq. 4 displays the standard format of a transfer function.

$$y = \sum_{i=1}^{n} W_i * X_i + c \tag{4}$$

Here, the variable y represents the weighted sum, where X_i represents the input values, W_i represents their respective

TABLE I. DIFFERENT ATTRIBUTES OF THE DATASETS

S.N.	Attributes	Type of Attribute	Attribute Value
1	CountryCode	Nominal	Austrailia, Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, England and Wales, Estonia, Finland, France, Germany,Greece, Hungary, Iceland, Israel, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Norway, Poland, Portugal, Russia, Scotland, Slovenia, Slovakia, Spain, Switzerland, Sweden, USA
2	Year	Numerical	1990-2022
3	Week	Numerical	1-53
4	Sex	Nominal	Male(m), Female(f), Both(b)
5	Death counts by age group (0-14), D0_14	Numerical	Mean: 24.48840206
6	Death counts by age group (15-64), D15_64	Numerical	Mean:741.2852896
7	Death counts by age group (65-74), D65_74	Numerical	Mean: 574.2379793
8	Death counts by age group (75-84), D75_84	Numerical	Mean: 860.7234414
9	Death counts by age group (84+), D85p	Numerical	Mean: 874.3757614
10	The total death counts of all ages combined, DTotal	Numerical	Mean: 3075.110874
11	Death rates by age group (0-14), R0_14	Numerical	Mean: 0.000410215
12	Death rates by age group (15-64), R15_64	Numerical	Mean: 0.003046857
13	Death rates by age group (65-74), R65_74	Numerical	Mean: 0.020856924
14	Death rates by age group (75-84), R75_84	Numerical	Mean: 0.055437482
15	Death rates by age group (84+), R85p	Numerical	Mean: 0.166329407
16	The total death rates of all ages combined, RTotal	Numerical	Mean: 0.009862773
17	Split	Nominal	0,1
18	SplitSex	Nominal	0,1
19	Forecast	Nominal	0,1

weights, and c represents the bias term. The output is then produced by passing the weighted total through an activation function. The training of the model has been performed in 100 epochs with relu activation function as well as Adam optimizer.

2) Recurrent Neural Network (RNN): RNN is a type of artificial neural network in which the output from one phase is fed back as input for the subsequent phase. It possesses hidden layers that utilize RNN memory to preserve information from prior computations, thereby facilitating the extraction of significant information for the purpose of sequential data processing. Thus, many applications with sequential data such as speech recognition [18], language translation [19], and human activity recognition can be benefited from RNNs. RNN converts independent activations into dependent ones by giving each layer the same amount of weights and biases. This reduces the complexity of increasing parameters and helps to memorize each previous output, which will be used as input for the subsequent hidden layer. After then, each set of three layers can be connected to form a single recurrent layer. The Eq. 5 represents the formula for determining the current state:

$$S_t = f(S_{t-1}, X_t) \tag{5}$$

Here, S_t denotes the present state, S_{t-1} denotes the preceding state, and X_t denotes the input state. The Eq. 6 represents the formula for using the activation function:

$$S_t = f(W_{ss}S_{t-1} + W_{sx}X_t) \tag{6}$$

Here f is the activation function, W_{ss} represents the weight assigned to the recurrent neuron, and W_{sx} represents the weight assigned to the input neuron. The Eq. 7 represents the formula to determine output:

$$Y_t = W_{sy}S_t \tag{7}$$

Here Y_t represents the output and W_{sy} represents the weight assigned to the output layer.

3) Linear Regression (LR): LR is one of the widely used machine learning methods which estimates the linear relationship between dependent and independent variables. It demonstrates how the value of the dependent variable changes based on the value of the independent variable. Basically, it is employed in predictive analysis. It forecasts factors that are real or numerical, such as birthday, sales, salary, age, and product price. The main goal of linear regression is to determine the best-fitting linear equation that reduces the disparity between the anticipated and actual values of the dependent variable. The Eq. 8 represents the formula of the model.

$$y = b_0 + b_1x_1 + b_2x_2...b_nx_n \tag{8}$$

Here y denotes the dependent variable, also termed as target variable, x_1, x_2... x_n denotes the independent variables, which are known as predictor variables. b_0, b_1, b_2...b_n denotes the coefficients associated with each independent variable. b_0 is the line's intercept, and a_1 is the linear regression coefficient.

4) Decision Tree (DT): DT is a supervised learning algorithm applied to both classification and regression problems. It is a tree-like structured approach where the internal nodes indicate input features or attributes, branches indicate the decision-making process that is based on those features, and leaf nodes indicate the output or prediction of the model. The algorithm starts at the root node of the tree and at every decision node, the algorithm selects the branch to pursue by evaluating the present record's values with associated decision node values. Based on this comparison, the algorithm follows the corresponding branch to the next node. One of the main issues in DT algorithms is to determine the best attribute for the root node and subsequent sub-nodes. An attribute selection measure (ASM) can be used to find solutions to these issues. There are two widely used ASM techniques that are described in the following sections.

- Information Gain: After dividing the data depending on an attribute, it calculates the reduction in entropy or uncertainty in the target variable. The splitting attribute is selected based on the attribute that has the highest information gain. The Eq. 9 represents the formula to calculate Information Gain (IG).

$$IG = Entropy(s) - \sum \frac{|S_v|}{|S|} * Entropy(S_v) \tag{9}$$

Here, $Entropy(s)$ represents the entropy of the original dataset S. $|S_v|$ represents the instance number in S that have the value v for attribute, $|S|$ represents the total instance number in S, and Entropy(S_v) is the entropy of the subset S_v after splitting the data based on the attribute value v.

- Gini Index: Gini index quantifies the impurity or dissimilarity of a dataset's value while creating a decision tree. The objective of the Gini index is to reduce impurities from the root nodes to the leaf nodes. The attribute with the lowest Gini index should be chosen as the splitting attribute. Gini index can be calculated using the below formula stated in Eq. 10.

$$GI(S) = 1 - \sum p_i^2 \tag{10}$$

Here, p_i is the proportion of instances in S that belong to class i.

5) Random Forest (RF): RF is another popular supervised algorithm that integrates the power of decision trees and ensemble learning for solving classification and regression problems [20]. It functions by randomly choosing subsets of the training data and features known as bootstrap samples from the original dataset. Each decision tree is then created independently using these subsets through a recursive process. During prediction, each tree generates an independent prediction and the final prediction is then determined by taking the average prediction of all the trees. There should be a chance that certain decision trees may generate incorrect predictions, but when all the trees are combined, it provides an accurate prediction. RF also provides several other benefits, including the ability to handle nonlinear relationships, capture complex interactions among features, and improved accuracy, and robustness against outliers and noise compared to individual decision trees [21].

6) K Nearest Neighbour (KNN): KNN is one of the simplest yet versatile algorithms applied to both classification and regression tasks [22]. This algorithm is non-parametric, instance-based, and makes no assumptions on the distribution of the underlying data. KNN algorithm assigns labels to previously unlabeled data based on the features and labels of its K nearest neighbors in the training data. The process involves computing the distance between the new unseen input data and each training sample using a certain distance metric such as Euclidean distance, Minkowski distance, Manhattan distance, hamming distance. In the classification process, KNNs are used to assigning labels to a new data point based on the dominant class label among the neighbors. In regression, the predicted value is calculated by averaging the target values of K's nearest neighbors. The choice of K may have impact on the algorithm's performance. If the value of K is smaller, it may lead to a potentially more flexible and noisier prediction and if the value of K is larger, it may lead to potentially smoother but biased predictions. In order to identify the K nearest neighbors, Euclidean distance is used most of the time as a distance metric. The Eq. 11 represents the formula to determine the nearest neighbors between two data sets, p and q.

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{11}$$

Here, p and q denote the coordinates of data points in each dimension, and n represents the total number of dimensions or features.

7) Least Absolute Shrinkage and Selection Operator (LASSO): LASSO is a type of linear regression model that employs shrinkage to select the variables. It is beneficial in analyzing datasets with high dimensions, specifically those with many features and fewer observations [23]. During prediction, the linear regression model provides equal importance to all features. However, when there are many features, including irrelevant or redundant ones, the model may become complicated and overfit the training data, which results in poor generalization of new data. Lasso Regression addresses this problem by adding an L1 penalty term to the cost function. The L1 penalty promotes sparsity and facilitates efficient feature selection by shrinking the coefficients of irrelevant features toward zero, thereby eliminating the corresponding features from the model. The generic form of the cost function in LASSO regression is presented on Eq. 12.

$$J = \frac{1}{m} \sum_{i=1}^{m} \left(y^{(i)} - h\left(x^{(i)} \right) \right)^2 + \lambda \sum_{j=1}^{n} |w_j| \tag{12}$$

Here, the variable m denotes the count of training examples. The variable $y^(i)$ represents the target variable's value for the i-th training example. The expression $h(x^{(i)})$ denotes the hypothesis function's for prediction, while n denotes the total number of features. The weight assigned to the jth feature is represented by w_j.

8) Support Vector Regression (SVR): SVR is a supervised learning algorithm used for classification and regression problems [24], [25], [26]. It is an expansion of the Support Vector Machine (SVM) algorithm. The aim of SVR is to select a hyperplane with a maximum margin while allowing a certain level of error (epsilon) for data points that lie within that margin. The SVR algorithm tries to identify the optimal hyperplane by solving an optimization problem that minimizes the training data error and maximizes the margin. In the prediction phase, SVR applies the learned hyperplane to predict the values for new data points. The predicted values are defined by the position of the data points with respect to the hyperplane. SVR can handle non-linear relationships and high-dimensional data effectively with the help of kernel functions. Kernel functions are applied to convert the input data into a higher-dimensional space, where it can find a linear regression function. The selection of the kernel function depends on the type of data and the problem at hand. The three kernels that SVM most frequently uses are.

- Linear kernel: It deals with large sparse data and is used in text categorization. It measures the linearity between the input data and the target variable.
- Polynomial kernel: This kernel, known as a polynomial kernel, introduces polynomial features to capture nonlinear relationships.

- Radial Basis Function (RBF): It maps the input data into an infinite-dimensional feature space applying Gaussian functions.

D. Performance Metric

The purpose of accuracy metrics is to determine the performance of any model. In this section, we used three evaluation metrics such as R Square coefficient, RMSE, and MAE to measure the prediction accuracy of the regression models.

1) R Square Coefficient: R square coefficient is an evaluation metric that measures the fitness of a regression model [7]. It can be expressed as using Eq. 13.

$$R^2 = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \bar{x})^2} \tag{13}$$

Where, x_i and $\hat{x}_i$ are the actual and predicted output of i-th sample respectively, $\bar{x}$ is the average output. The highest value of R^2 is 1, indicating that the closer the value to 1 the better fitted the model is.

2) Root Mean Square Error(RMSE): RMSE is a general-purpose error metric used to measure the performance of a model according to prediction accuracy [27]. The smaller the RMSE value the higher the prediction accuracy. It can be expressed as the square root of the mean squared error. The equation to calculate MSE and RMSE for multi-output regression model is provided in Eq. 14 and 15, respectively.

$$MSE = \frac{1}{m} \frac{\sum_i (x_i - \hat{x}_i)^2}{N} \tag{14}$$

$$RMSE = \sqrt{MSE} \tag{15}$$

In the above equation stated in 14, N denotes the number of samples.

3) Mean Absolute Error(MAE): MAE calculates the difference between actual output and predicted output. MAE for multi-output regression model expressed in the Eq. 16.

$$MAE = \frac{1}{m} \frac{1}{n} \sum_{i=1}^{n} |(x_i - \hat{x}_i)^2| \tag{16}$$

IV. Experimental Results

The main objective of the proposed model is to predict the weekly death count based on age-specific user group. To perform the task, we have implemented several regression models. The dataset was divided into 80% to 20% where 80% data is used for training and 20% data for testing purposes. All the experiments were performed in Python. The performance of these models is evaluated based on three different metrics such as MSE, MAE and R squared coefficient. The higher value of the R squared coefficient is found with RF algorithm (0.9975), which is followed by DT (0.9958), RNN (0.9529), KNN (0.9430), ANN (0.9427), LR (0.8937), Lasso (0.8937), SVR (0.8438) which is shown on Table II. The higher value of the R squared coefficient, the lower value of RMSE and MAE indicates good fitted model for the task. It is evident that the value of MAE is lower with RF (16.4069) that is followed by DT (21.7217), KNN (106.2851), ANN (114.8771), RNN (124.3763), Lasso (211.7925), LR (212.0527) and SVR (245.8620). The lowest RMSE value is found on RF with 43.2263 which is followed by DT, RNN, KNN, ANN, LR, Lasso, and SVR. Therefore, it can be concluded that RF is the best-performing model for the task and after then DT showed almost similar types of prediction.

TABLE II. Comparison among the Regression Models based on RMSE, MAE and R Squared Coefficient

	RMSE	MAE	R Squared
RF	43.2263	16.4069	0.9975
DT	56.4134	21.7217	0.9958
RNN	333.8810	124.3763	0.9529
KNN	386.2374	106.2851	0.9430
ANN	389.7136	114.8771	0.9427
LR	525.6425	212.0527	0.8937
Lasso	525.6477	211.7925	0.8937
SVR	656.4845	245.8620	0.8438

V. Discussion

In this research, we figured out that, instead of using the distribution equation, the construction of a model with random forest and decision tree algorithms to perform the count of mortality in absence of age specific data from the total count of all ages is much easier and better solution. It is evident that the use of the multi-output regression model has proved its efficiency to perform the prediction. To the best of our knowledge, this work is the first attempt to propose a multi-output regression model as a solution to the distribution problem on the mentioned dataset. It can be summarised that ML techniques provide better output for multi-output regression than DL methods. Among the classifiers, RF showed the best performance based on RMSE, MAE, and R squared coefficient.

However, the performance of several algorithms can further be improved through tuning the hyper-parameters of the model. In addition, the utilization of these regression models can be applied to the similar domains through extending their potentiality.

VI. Conclusion

STMF data series is one of the most valuable data series of HMD. Various types of analysis can be performed using the weekly records according to the information of their age group and gender. However, the data that are collected from various countries sometimes lack the weekly death information. Currently, researchers used the distribution equation to calculate the age-specific weekly data. Multi output regression models are getting popularity in prediction related problems over the last few years. In this work, we have implemented such regression models because of the benefits over single output model. In this work, RF is selected as the best performing model based on R-square coefficient, RMSE and MAE.

References

[1] D. A. Jdanov, A. A. Galarza, V. M. Shkolnikov, D. Jasilionis, L. Németh, D. A. Leon, C. Boe, and M. Barbieri, "The short-term mortality fluctuation data series, monitoring mortality shocks across time and space," *Scientific Data*, vol. 8, no. 1, p. 235, Dec. 2021. [Online]. Available: https://www.nature.com/articles/s41597-021-01019-1

[2] L. Németh, D. A. Jdanov, and V. M. Shkolnikov, "An open-sourced, web-based application to analyze weekly excess mortality based on the short-term mortality fluctuations data series," *Plos One*, vol. 16(2), no. e0246663, 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0246663

[3] "Hmd," *The Human Mortality Database*, 2020. [Online]. Available: http://www.mortality.org/.

[4] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on Multi-Output Learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8892612/

[5] L. Cui, X. Xie, Z. Shen, R. Lu, and H. Wang, "Prediction of the healthcare resource utilization using multi-output regression models," *IISE Transactions on Healthcare Systems Engineering*, vol. 8, no. 4, pp. 291–302, Oct. 2018. [Online]. Available: https://doi.org/10.1080/24725579.2018.1512537

[6] A. Boumezoued and A. Elfassihi, "Mortality data correction in the absence of monthly fertility records," *Insurance: Mathematics and Economics*, vol. 99, pp. 486–508, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167668721000536

[7] N. Shahid, M. A. Shah, A. Khan, C. Maple, and G. Jeon, "Towards greener smart cities and road traffic forecasting using air pollution data," *Sustainable Cities and Society*, vol. 72, p. 103062, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670721003462

[8] Z. Han, Y. Liu, J. Zhao, and W. Wang, "Real time prediction for converter gas tank levels based on multi-output least square support vector regressor," *Control Engineering Practice*, vol. 20, no. 12, pp. 1400–1409, Dec. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0967066112001670

[9] D. Tuia, J. Verrelst, L. Alonso, F. Perez-Cruz, and G. Camps-Valls, "Multioutput Support Vector Regression for Remote Sensing Biophysical Parameter Estimation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 804–808, Jul. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5735189/

[10] H. Li, W. Zhang, Y. Chen, Y. Guo, G.-Z. Li, and X. Zhu, "A novel multi-target regression framework for time-series prediction of drug efficacy," *Scientific Reports*, vol. 7, no. 1, p. 40652, Jan. 2017. [Online]. Available: https://www.nature.com/articles/srep40652

[11] A. Meyer, "Multi-target normal behaviour models for wind farm condition monitoring," *Applied Energy*, vol. 300, p. 117342, Oct. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306261921007509

[12] C. Kucuk, D. Birant, and P. Yildirim Taser, "An intelligent multi-output regression model for soil moisture prediction," in *Intelligent and Fuzzy Techniques for Emerging Conditions and Digital Transformation*, C. Kahraman, S. Cebi, S. Cevik Onar, B. Oztaysi, A. C. Tolga, and I. U. Sari, Eds. Cham: Springer International Publishing, 2022, pp. 474–481.

[13] S. Xu, X. An, X. Qiao, L. Zhu, and L. Li, "Multi-output least-squares support vector regression machines," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1078–1084, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865513000196

[14] G. Liu, Z. Lin, and Y. Yu, "Multi-output regression on the output manifold," *Pattern Recognition*, vol. 42, no. 11, pp. 2737–2743, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320309001691

[15] D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen, "Survey on multi-output learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409–2429, 2020.

[16] S.-C. Wang, *Artificial Neural Network*. Boston, MA: Springer US, 2003, pp. 81–100. [Online]. Available: https://doi.org/10.1007/978-1-4615-0377-4_5

[17] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943. [Online]. Available: http://link.springer.com/10.1007/BF02478259

[18] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[20] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/10.1023/A:1010933404324

[21] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[22] V. Iswarya Kumari, M. Bhavya, N. Kasi Mounika, and M. Yoshitha, "Sales Prediction using Linear Regression," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 139–141, Nov. 2022. [Online]. Available: http://ijarsct.co.in/Paper7611.pdf

[23] S. Kwon, S. Han, and S. Lee, "A small review and further studies on the lasso," *Journal of the Korean Data and Information Science Society*, vol. 24, no. 5, pp. 1077–1088, 2013.

[24] M. Sajjad, S. U. Khan, N. Khan, I. U. Haq, A. Ullah, M. Y. Lee, and S. W. Baik, "Towards Efficient Building Designing: Heating and Cooling Load Prediction via Multi-Output Model," *Sensors*, vol. 20, no. 22, p. 6419, Nov. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/22/6419

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/10.1007/BF00994018

[26] P. Lu, L. Ye, W. Zhong, Y. Qu, B. Zhai, Y. Tang, and Y. Zhao, "A novel spatio-temporal wind power forecasting framework based on multi-output support vector machine and optimization strategy," *Journal of Cleaner Production*, vol. 254, p. 119993, May 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652620300408

[27] T. O. Hodson, "Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not," *Geoscientific Model Development*, vol. 15, no. 14, pp. 5481–5487, 2022. [Online]. Available: https://gmd.copernicus.org/articles/15/5481/2022/