Tree-Based Regressor Ensemble for Viral Infectious Diseases Spread Prediction

Henry Mutisya Ngie
College of Pure and Applied Science (COPAS), Jomo Kenyatta University of Agriculture and Technology P.O Box 42603 - 00100 Nairobi - Kenya

Lawrence Nderu
College of Pure and Applied Science (COPAS), Jomo Kenyatta University of Agriculture and Technology P.O Box 42603 - 00100 Nairobi - Kenya

Dorcas Gicuku Mwigereri
College of Pure and Applied Science (COPAS), Jomo Kenyatta University of Agriculture and Technology P.O Box 42603 - 00100 Nairobi - Kenya

Abstract

Tree based regression models provide statistical bases for prediction of continuous response variable scores. They are non-linear models founded on simplicity and efficiency when deployed on multi variable data domains. Their fast prediction speed, ability to identify strong variables in prediction and reliance on statistical means to deal with missing values in datasets during prediction make these models common in modern machine learning. Some of these models such as CART, RETIS and M5 have been utilized in the past yielding reliable prediction outcomes which are yet to be achieved through the use of single classifier models due the growing dataset complexities as a result of recent trends in data science including big data and internet of things.

Combination of several classifiers through ensemble approach can boost feature selection and enhance classifier prediction capabilities. This research paper demonstrates ensemble of Decision Tree (DT) and Logistic Regression (LR) models to develop a tree-based regressor model christened *Simultaneous Tree-based Regressor Interactive Model (STRIM)*, with improved interaction effect especially on continuous response variable predictions. The model involves particle swarm optimization (PSO) for parameter tuning in an effort to ensure a balanced and reliable prediction achievement in the spread of infectious diseases, incorporating time series modeling. The model aimed at providing a solution to the prediction of infectious diseases spread using publicly available Covid-19 global data for evaluation through prediction of Covid-19 spread patterns. Existing models used in the domain are largely black-box and therefore the need for a glass-box model capable of disclosing the impact of prediction features to the final prediction results. STRIM proved to be a robust interpretable classifier model compared to single classifiers considered for the ensemble providing 0.99 accuracy levels of prediction.

Keywords: Covid-19, Ensemble learning, Tree-based regressor, feature selection, parameter tuning, Particle swarm optimization

1. INTRODUCTION

Towards the end of 2019, hospital in Wuhan China experienced patients with severe pneumonia like symptoms yet responsive to no drugs or vaccines [1]. Human to human transmission of the problem was noted among patients portraying the disease as highly infectious with continuous human to human transmissions reported. Health experts later linked the disease with the symptoms of Severe Acute Respiratory Syndrome (SARS) disclosing that Corona virus was behind its spread and cause. What started as a local epidemic in China later spread to the entire world with alarming high infection rates and deaths recorded especially in Europe and America leading to the World Health Organization (WHO) naming the disease as novel Corona Virus Disease 2019 (nCov-19) or (COVID-19) in addition to declaring it a global pandemic. WHO directions on containing the pandemic was to engage three important steps: (identify, isolate and contact-trace) however for this to be successful, governments across the world prefer to model their infection patterns so as to have an overview of expectations including infection rates, spread, high risk regions, peak and curve flattening point.

Several machine learning approaches have since been involved in modeling these patterns globally with significant results so far achieved. This paper proposes a tree based regressor model - Simultaneous Tree-based Regressor Interactive Model (STRIM) for application in the same space. Tree based regressor algorithms were initially proposed in a publication by Morgan and Sonquist done in 1963 [1]. They have since become an interesting research area for data scientists. This paper borrows sentiments by Braiman (1984) demonstrating regressors and regression as a simple common analytical technique for identification and modelling of relationships between explanatory input and output variables. Tree based ensemble models just like typical ensembles share common strengths which include high accuracies, robustness and interpretability in prediction than the ensembled classifiers [2]. They are known to have empirically demonstrated positive results as demonstrated in [3] where one such an algorithm (Intelligent Mining for TB infection prediction) used in predicting regional tuberculosis infection patterns reported an R^2 value of 94% [2]. This research seeks to improve these outcomes through incorporation of optimization based feature selection using the particle swarm optimization (PSO). The approach has been previously applied in single classifier models such as neural networks where it has enhanced reliability and accuracy of prediction

of cancer diagnosis among patients as demonstrated in the related work section.

2. RELATED WORK

There has been little research on infectious diseases (especially Covid-19) modelling published as at the time of this paper, with much information about the virus still not established including its treatment vaccine. This section demonstrates previous work by several researchers on infectious diseases spread prediction especially Covid-19, using several epidemiological machine learning algorithms.

2.1 ARIMA Model

In [4] Autoregressive Integrated Moving Average (ARIMA) model developed in R was described. The model is founded on three sequential parameters namely; p, d, q which can be written in linear form. The model was further enhanced with AUTOARIMA packages in R to facilitate simple time series analysis so as to express the pandemic spread [4]. The model was successfully applied on a 45-day (15th Feb to 31st Mar 2020) COVID-19 patients dataset in Italy gathered from the country's health ministry. It accurately predicted outcomes, at 93.75% level calculated using the Mean Absolute Prediction Error (MAPE) [4] showing an upward trend in the virus spread across Italy with the disease peak point predicted to be the last two weeks of March 2020. Although the country later imposed a nationwide lockdown to combat the disease, the prediction was true with the peak orchestrated by an increase in travelers into the country. The model was further used to predict the next 60 days registered and recovered cases in the country predicting an increase in both to between 106,000 - 183,000 and 17,000 – 82,000 respectively [4].

2.2 SIR Models

Empirical research has also demonstrated the use of SIR epidemiological models which classify cases in three categories of susceptible, infected and recovered [5] [6]. The three classes are simply referred to as vulnerable to the infection (Class S), infected population (Class I) and (Class R) removed representing recovered, developed immune, isolated or even dead population (Class R). When developing epidemiological models, one domicile challenge is allocating the value of R since class R members may get re-infected. Common SIR Models include SEIR, SEIS, MSEIR and MSEIRS among others. Several countries have so far used SEIR for COVID-19 predictions which has a consideration of an incubation period for patients. The model has so far expressed accuracies in prediction despite the model's instability in dynamic contact networks [5] [7]. Despite the little success of the models, there have been associate drawbacks such as short lead time as the model accuracies decrease with time for instance in Italy it decreased from 100% in day one to 85% in day six [6]. In addition there is a possibility of significant data loss due to the disease patterns leading to possible prediction errors. Research demonstrates the need for advancing SIR to achieve scalable models with reliable performance and accuracy.

2.3 Machine Learning Models

Machine learning (ML) models have also been utilized in this space outweighing the SIR models due to the reliable performance levels observed in ML models for long lead times [5] [6] [8]. The use of Machine learning for infectious diseases modelling and prediction has been practiced for some time. For instance, Random forest was used for Swine fever prediction, artificial neural networks (H1N1 flu, and Oyster norovirus), CART in dengue fever which was also modelled using neural networks and Bayesian network among others [9-11].

There however exists a gap in peer reviewed literature on use of machine learning algorithms application in infectious diseases spread prediction especially COVID-19 modelling despite their use in several pandemics. Nevertheless many researchers have expressed hope in the use of the models for COVID-19 due to their promising results in previous infectious diseases modelling. Ideally, despite the high performance levels observed previously in the use of machine learning algorithms potential areas of weakness have been identified in research. [6], Proposes improvement of both ML and SIR models due to their underlying weaknesses and improvement of time series advancements in machine learning. This paper aims at advancing time series modelling in machine learning through ensemble approach borrowing concepts of machine learning application in time series modeling projects such as weather forecasting and natural disaster modeling. With ensemble machine learning it was possible to model daily spread patterns in terms of cases and peak times by borrowing experiences of countries that have gone through an almost full cycle of infectious diseases. However notable concerns still remain on the dynamics surrounding the mitigation measures taken by different governments on combating spread of infectious diseases. It is obvious that they impact prediction but it is also notable that some of the combating strategies are a tall order for some states and governments. Much about data used in developing the model, architecture and performance is discussed in next parts of this paper, organized into; methodology which describes the model development approach, results section elaborating the model performance and evaluation of its performance, the model's prediction outcomes from its application on data from valid sources, discussion section which describes the model performance, contribution to the world of research and finally the conclusions section.

3. METHODOLOGY

This paper is founded on Cross Industry Standard Process for Data Mining (CRISP-DM) [9] scientific methodology. Proposed and developed in 1990s by a consortium of European companies [10], the methodology follows a six step process beginning with industrial understanding. At this point an understanding of infectious diseases spread predictor modelling was carried out so as to inform formulation of study goals as well as the need for data mining. This was followed by identification and understanding of data sources relevant to the study, in which the researcher identified public data sources relevant to the study including data.humdata.org and Kaggle as potential for the research. The third step was data pre-processing during which the sourced data for use in the model development was pre-processed and cleaned to eliminate any potential imbalances. Transformation of the data to match the later demonstrated variables for use in training and testing the model was also carried out at this stage. The end result was a relevant dataset for use in the research. The fourth step in CRISP-DM is the actual model formation using identified machine learning algorithms. At this point STRIM classifier was developed through ensemble of decision tree and regression trees. The model was developed via the use of tools such as python with various scientific analytical operations carried to enhance its performance as per its proposal. In the fifth step, the classifier was evaluated and assessed for validity and success against the single classifiers used in the model with or without PSO in line with the objectives of the study. The sixth step in the process was deployment of the developed model for use in machine learning based prediction of infectious diseases spread. According to [10] CRISP-DM allows backtracking since its driven by experience and experimentation of the research problem therefore offering an iterative model development process which is key in ensuring that a model follows the correct path at all times. It is for this reason that this research paper considered its use.

To ensure objective assessment of the model's predictive powers and robustness, cross-validation was used [10] with 10 folds. It gives comparative accuracies of several models all together and therefore it was found prudent for both the resultant regressor model and involved single classifier evaluation. Cross-validation typically splits data into mutually exclusive subsets for use as training and testing data. In extra ordinary circumstances such as neural network application an additional third subset called validation set is created [10 – 12]. Figure 1 demonstrates application of CRISP-DM in this paper.

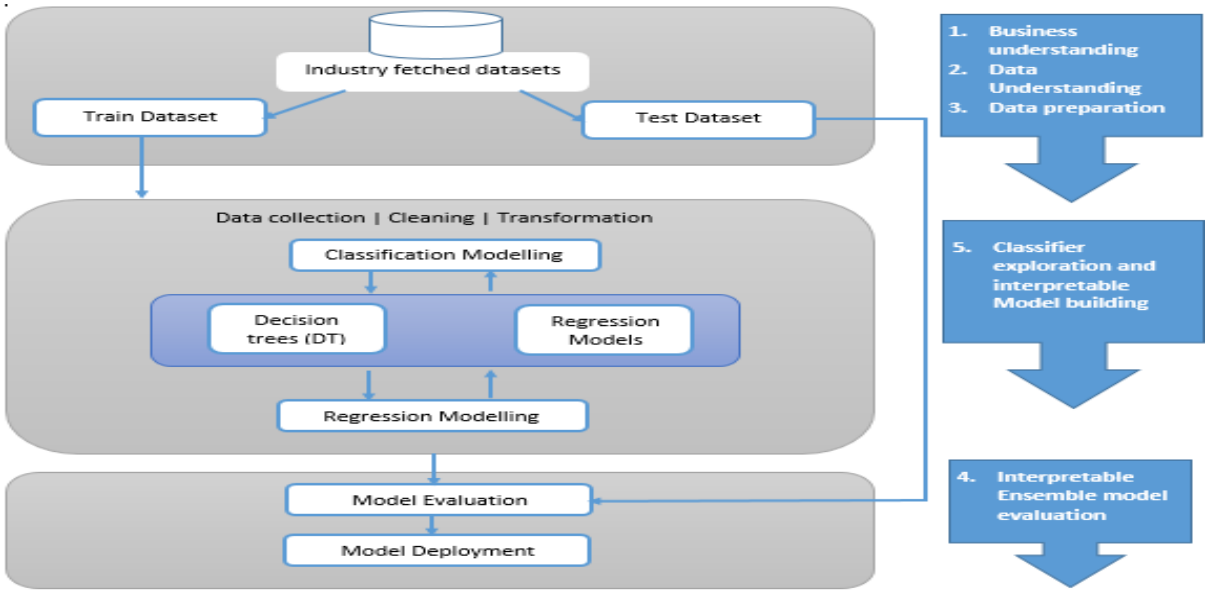

Figure 1 CRISP-DM approach framework for developing STRIM

3.1 Research Data

The sample data for use in this model was collected from public data sources such as The World Health Organization (WHO) website accessed from data.humdata.org and Kaggle accessed through kaggle.com/c/covid19-global-forecasting. The dataset contained global data regarding the virus regional spread in over a hundred countries indicating the daily infected cases and fatalities sampled between January and June 2020 relative to when respective countries/states reported the pandemic. The dataset has 20,000 entries as of this research date on which data mining heuristics were instrumental in ensuring that valid and relevant data was considered for the model formation. Therefore variables were deeply identified as identification variables (country short name, country/state), input variable in the dataset included (population, weight, date, cumulative cases and cumulative deaths). Finally the two output variables in the dataset were identified as (new cases and new deaths).

Table 1 an explanation of the dataset variables used in developing the model

No	Abbreviation	Variable name	Category	Description
1	OID	OBJECTID	Identification variable	Record ID in the dataset
2	ISC2	ISO_2_CODE	Identification variable	Country's ISO 2 character name
3	ISC3	ISO_3_CODE	Identification variable	Country's ISO 3character name
4	ADMNO	ADM0_NAME	Identification variable	The country's name
5	DATE	DATE_EPICRV	Input variable	The date that results were received and recorded in the WHO dataset
6	NCASE	NEWCASE	Output variable	New cases recorded daily into the WHO dataset
7	CCASE	CUMCASE	Input variable	The daily cumulative cases recorded daily into the WHO dataset
8	NDEATH	NEWDEATH	Output variable	New fatalities reported and recorded in the WHO dataset
9	CDEATH	CUMDEATH	Input variable	Cumulative number of fatalities reported and recorded in the WHO dataset

In the dataset framing, the rows represented sampled daily cases of the virus identification and reporting based on WHO records. Each of the columns in the dataset is a representation of either an identifier variable, input typically representing the attributes of the two output variables in the study.

3.2 Feature selection and Model Optimization with PSO

Particle Swarm Optimization (PSO) is an evolutionary computation technique for intelligence gathering (pattern recognition) in datasets proposed and developed by Dr. Kennedy and Dr. Eberhart in 1995 [16]. The idea of PSO can be traced in intelligent swarming behavior of bird flocks, bees and the human being's social behavior [17]. It offers a meta-heuristic approach to the problem under optimization. This means no assumptions are made on the problem hence wide solution search for optimized performance. PSO can be further enhanced through combinatorial approaches as several researchers have previously done [18]. The equations below are respectively used in switching the particle positions repetitively until an optimal criterion is met. Step two is essentially repeated until this criterion is achieved.

$$Vid = W * Vid + c1 * rand1 * (Pbestid - Xid) + c2 * rand2 * (Gbestid - Xid) \quad (1)$$

$$Xid = Xid + Vid \quad (2)$$

Where Vid and Xid represent the velocity and position of the i^{th} particle with d dimensions, rand1 and rand2 are the uniform random functions.

3.3 Applied Classifiers

3.3.1 Decision Trees

These are classification models composed of a handful of input variables (attributes) contributing to their prediction capabilities [10]. They are majorly build on tree like formations of branches and nodes. A branch in the tree represents the output conditions leading to classification of a prediction to its respective node which denotes the outcome in a transparent manner. Spanning of a decision trees through the many branches and nodes can be viewed as a complex formation of multiple if-else statements. The idea behind decision tree algorithms is built on recursive division of training data until classification of elements belonging to similar classes together is achieved, making the decision tree algorithms pure. Generalization can be achieved in the algorithms through pruning of the trees leading to accuracy improvement on predictions by decision tree algorithms. Decision trees use statistical algorithms including Gini index, information gain (entropy) and Chi-square statistics to reduce bias in their predictions.

3.3.2 Regression Classifiers

Regression models (RM) are statistical tools used for outcome prediction through identification of variable relationships and useful patterns in datasets. They are data oriented in nature and thus considered as data fitting models [13]. This is because gathered data elements are compared with each other in these models without any close consideration of processes behind the data elements. Linear regression is a specific form of regression in which linear predictor functions are used to predict unknown parameters in subjected data [3] [14]. Where multiple predictor variables exist another form of regression christened multiple regression is considered, with the general formula:

$$Y = a_0 + \sum_{j=1}^{m} aj\ xj \quad (3)$$

Where Y is the model's output, xj = various input variables, a = partial regression coefficients

In the equation k is the number of folds, CV is the cross validation accuracy and A is the fold-wise accuracy

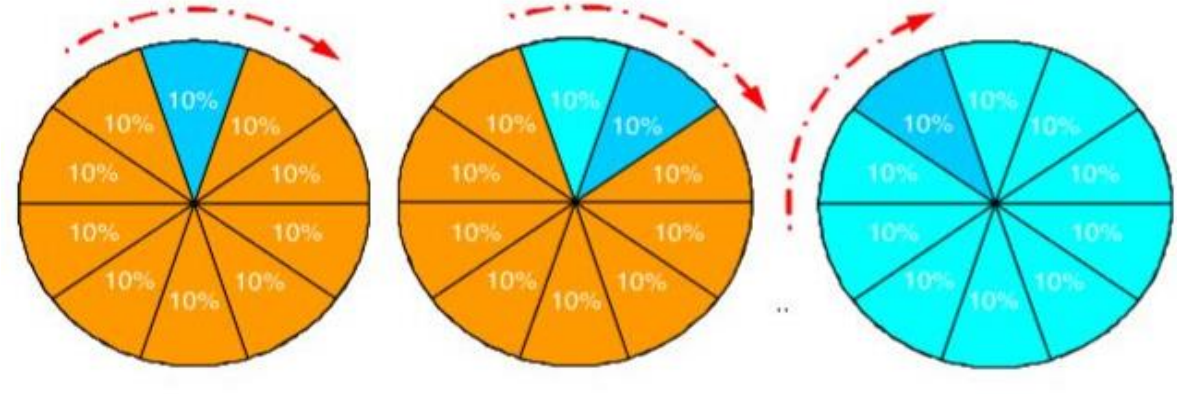

Figure 2 10-fold cross-validation approach conceptual view

Bagging regression is an acronym derived from Bootstrap aggregating, an ensemble approach to classification through utilization of different classification methods and regression methods with the aim of reducing prediction process variance [13]. The foundation concept in bagging is development of individual regression models that make use of random distributed training set to train a single algorithm. For each of the regression models there is a random training set of N instances (N = size of original training set). A significant number of the original instances may be repeated or completely ommited in the test. Upon construction of several regression models iteratively, average prediction values are used to give the final prediction. The approach has for long been associated with a proper response to handling of missing values in datasets due for use in prediction. [16], suggests splitting corresponding instances into pieces as one of the approaches deployed by bagging regression trees in handling missing values besides other approaches.

3.4 Model Evaluation

The fifth step in CRISP-DM involves ensuring effective performance of the model through evaluation. K-fold cross-validation, provided a good approach to the developed model evaluation by eliminating sampling biasness in the segments with a k-value of 10. The approach split the dataset randomly into k equal estimated subsets. The common form of the approach is the 10 fold approach where the value of k is 10 and the classification model is trained and tested these k number of times ensuring training on all but one of the folds each time. The remaining fold is used for testing and a confusion matrix formed out of all the test outcomes. In the approach overall accuracy cross-validation is calculated using equation 4:

$$CV=\frac{1}{k}\sum_{i=1}^{k} Ai \qquad (4)$$

measure. Accuracy, sensitivity and specificity can be used to compare the individual classifier performance over the developed ensemble model with the formulas in equation 5. Parameters in the equation are true positive (TP), true negative (TN), false positive (FP) and true positive (TP).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN,} \qquad (5)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN,} \qquad (6)$$

$$\text{Specificity} = \frac{TN}{TN+FP,} \qquad (7)$$

Additional evaluation

Root mean square error (RMSE) and mean absolute percentage error (MAPE) offered additional approaches to not only evaluation but also establishment of model accuracy levels associated with respective models. They are based on the sum of square errors in terms of how far the data is from mean and model predicted values. The equations used in the two approaches were:

$$\text{MAPE} = \frac{1}{N}\left|\frac{x-y}{x}\right| X\ 100 \qquad (8)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}}\ \sum (\text{A-P})^2 \qquad (9)$$

4 RESULTS

4.1 Discussion of Outcomes

A comparison of the single classifiers and the proposed ensemble 10-fold cross validation results are demonstrated in Table 2. The single classifiers evaluated included decision trees, regression trees as well as STRIM and STRIM classifier with PSO. Various aspects of comparison were factored in comparing the individual classifier outcomes including training time, training and testing accuracies as well as the prediction accuracies for each of the classifiers. In

addition to the above results based on the evaluation techniques provided. Accuracy levels for the individual classifiers were as summarized in Table 3. The model accuracy levels were further compared to each other leading to the outcomes demonstrated in Table 2. . The results demonstrate low accuracies in decision trees as well as low training time for decision trees. However their input once involved in ensemble modelling is very significant as demonstrated in the evaluation of STRIM model. Involvement of PSO in the model yielded better results boosting model accuracy as well as reducing training and testing time. Based on the findings and evaluations in Table 2, it can be concluded that ensemble approaches to classifier development lead to accuracy enhancement. It also has been described in the table that parameter tuning approaches such as particle swarm optimization can assist in improving model performance as well as significantly and positively impacting time series modeling. These evaluations match those of [6] in modelling for prediction of infectious diseases spread where regression trees were demonstrated as highly effective and transparent in predicting outcomes.

Table 2 individual classifier evaluation results (cross-validation, RMSE, MAPE)

Performance measure	Predictive technique	Min	mean	Max	Best fold
RMSE	Decision trees	86.78	88.02	89.11	2, 7, 9
	Regression tree (CART)	87.62	90.43	96,78	4
	STRIM Ensemble	92.16	94.11	97.54	1, 4, 7
	STRIM with PSO	96.28	96.84	98.14	3, 5
MAPE	Decision trees	85.74	87.07	90.60	5 and 6
	Regression tree (CART)	89.18	92.04	95.62	9
	STRIM Ensemble	93.51	94.01	98.06	2, 8
	STRIM with PSO	96.16	97.42	98.72	4
Training time (s)	Decision trees	3.6400	5.8200	6.4000	2
	Regression tree (CART)	0.8200	2.3400	5.0000	9
	STRIM Ensemble	0.7200	2.1200	4.3200	7
	STRIM with PSO	0.6200	2.8420	3.5400	4, 6, 8,
Testing time (s)	Decision trees	0.8881	2.3400	4.5200	4
	Regression tree (CART)	0.7460	1.8640	3.7500	9
	STRIM Ensemble	0.8240	2.1500	3.2420	5 and 6
	STRIM with PSO	0.6800	2.1000	3.1500	4, 7, 9

Table 3 classifier models accuracy summary

Model	Accuracy		Time	
	Training %	Testing %	Training (s)	Testing (s)
Decision trees	0.8409	0.8881	5.8200	2.3400
Regression tree (CART)	0.7825	0.8102	2.3400	1.8640
STRIM Ensemble	0.9714	0.9810	2.1200	2.1500
STRIM with PSO	0.9987	0.9999	2.8420	2.1000

Figure 3 model accuracy vs time comparison

4.2 Paper Contributions

This paper's contributes to enhancement of existing models in the research domain by using hybrid tree based regressor ensemble for infectious diseases spread modelling whose prediction capability was compared to previous models (SIR models and ANFIS). The research paper further explores the possibility of enhancing prediction accuracy through time series modelling and ensuring model transparency in prediction thus making the resultant model both a classifier and knowledge source for its users and potential domain experts.

5 CONCLUSION

Ensemble machine learning can be useful in prediction of Covid-19 and other viral infectious diseases' spread patterns within human populations. This paper proposed a tree based regressor ensemble model for prediction of COVID-19 spread patterns. The ensemble is based on common classifier models ensembled together through the bagging approach with particle swarm optimization added in effort to optimize prediction accuracy and robustness in the proposed ensemble. Compared to the individual select classifiers, the ensemble classifier was confirmed as more robust, interpretable and enhanced for predicting the spread patterns.

There were several limitations to this research paper including Covid-19 prediction data dynamism which has seen drastic changes in patterns on related data in a short time span. It is important to note that countries keep changing tactics associated with managing the

pandemic in a non-uniform way with several countries such as European countries imposing complete lockdowns of between 14 and 21 days, others imposing partial lock downs, cessation of movement, social distancing, and improved hygiene procedures. All these factors may alter infectious diseases' spread patterns greatly and therefore future research within these lines is highly proposed with semi supervised learning techniques proposed. However the developed ensemble model was found effective in predicting the disease spread trends through regional accurate prediction of periodic infections and fatalities. This demonstrated the underlying potential in tree based regressors and ensemble machine learning in classifier modelling.

References

[1] Dupan, From Man to Machine; Neural Control of Fine Hand Movement, Nijmegen: Donders Graduate School for Cognitive Neuroscience, 2018.

[2] M. H. M. Armaghani, "Application of Several Non-linear Prediction Tools for Estimating Uniaxial Compressive Strength of Granitic Rocks and Comparison of their Performances," *Eng. Computing Journal*, 2016.

[3] X. M. Luan, "Ensemble Learning Regression for Estimating Unconfined Compressive Strength of Cemented Paste Backfill," *IEEE Journal*, vol. 7, pp. 72125 - 72133, 2019.

[4] Chintalapudi, "COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach," *Journal of Microbiology, Immunology and Infection, 2020.04.004*, vol. 4, no. 4, 2020.

[5] I. F. A. M. P. G. Gergo, "COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach," *Journal of Microbiology, Immunology and Infection*, vol. 1, no. 1, 2020.

[6] M. N. Koike F., "Supervised forecasting of the range expansion of novel non-indigenous organisms: Alien pest organisms and the 2009 H1N1 flu pandemic," *Global Ecol. Biogeogr*, vol. 10, pp. 991 - 998, 2018.

[7] M. Imran, M. Usman, M. Dur-e-Ahmad and Khan, "Transmission Dynamics of Zika Fever: A SEIR based model," *Bulleting of Infectious Disease Modelling* , vol. 12, no. 4, 2020.

[8] R. Burke, M. Shah, M. Wikswo, L. Barclay, A. Kambhampati, Z. Marsh and J. Cannon, " The Norovirus Epidemiologic Triad: Predictors of Severe Outcomes in US Norovirus Outbreaks," *Journal of Infectious Deseases*, pp. 1364-1372, 2019.

[9] Shearer, "The CRISP-DM model: The new Blue Print for Data Mining," *Journal od Data Warehousing*, vol. 5, pp. 13 - 22, 2000.

[10] D. D. E. T. Ramesh Sharda, Business Intelligence, Analytics, and Data Science; A Managerial Perspective, Pearson, 2014.

[11] C.-U. B. S. Song, "The Comparative Accuracy of Judgemental and Model Forecasts of American FootbalL Games," *International Journal of Forecasting*, vol. 23, no. 3, pp. 405 - 413, 2007.

[12] L. Nderu, "An Ensemble Filter Feature Selection Method and Outlier Detection," *ACM Journal*, 2016.

[13] S. M. Sadrmomtazi, "Modelling Compressive Strengh of EPS Lightweight Concrete Using Regression, Neural Networks and ANFIS," *Construction and Building Materials Journal*, vol. 42, pp. 205 - 216, 2013.

[14] S. M. Sadrmomtazi, "Modelling Compressive Strengh of EPS Lightweight Concrete Using Regression, Neural Networks and ANFIS," *Construction and Building Materials Journal*, vol. 42, pp. 205 - 216, 2013.

[15] S. Rebecca, Tree Based Methods:Decision and Regression Trees, Third Edition ed., New York: Oxford Publishers, 2017.

[16] K. a. Eberhart, "Particle Swarm Optimization Proceedings, IEEE International Conference on Neural Networks," in *IEEE Journal*, Perth, Australia, 1995.

[17] G. a. Venu, "Comparison of Particle Swarm Optimization and BackPropagation as Training Algorithms for Neural Networks," *IEEE Journal*, vol. 18, no. 4, pp. 110 - 113, 2003.

[18] R. E. Shi, "Empirical Study of Particle Swarm Optimization Proceedings of the 1999 Congress on Evolutionary Computation," in *CEC Journal*, Madrid, 1999.

[19] A. Nashat, "Ensemble Machine Learning for Leukemia Cancer Diagnosis Based on Microarray Datasets," *International Journal of Applied Engineering Research*, vol. 19, no. 21, pp. 4077 - 4084, 2019.

[20] S. Anno, T. Hara, H. Kai, M. Lee, Y. Chang, K. Oyoshi, Y. Mizukami and Tadono, "Spatiotemperal dengue fever hotspots associated with climatic factors in taiwan including outbreak predictions based on machine-learning," *Geospatial Health 2019*, vol. 14, pp. 183-194, 2019.

[21] D. Raja, R. Mallol, C. Ting, F. Kamaludin, R. Ahmad, S. Ismail, V. Jayaraj and B. Sundram, "Artificial intelligence model as predictor for

dengue outbreaks," *J. Public Health Med.,* vol. 19, pp. 103-108, 2019.

[22] O. Titus Muurlink, P. Stephenson, M. Islam and A. Taylor-Robinson, "Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach," *Infectious Deseases Journal,* vol. 3, pp. 322-330, 2018.

[23] X. M. Luan, "Ensemble Learning Regression for Estimating Unconfined Compressive Strength of Cemented Paste Backfill," *IEEE Journal,* vol. 7, pp. 72125 - 72133, 2019.

[24] D. C. N. K. Dursun Delen, "A comparative analysis of data mining methods in predicting NCAA bowl outcomes," *Journal of Forecasting,* vol. 28, pp. 543 -552, 2012.