Isolation Forest Based Multi-Source Unsupervised Transfer Learning for Missing GDP Prediction

Sandeep Kumar **Amit K. Shukla** **Pranab K. Muhuri**
Department of Computer Science, South Asian University
Akbar Bhawan, Chanakyapuri, New Delhi-110021, India
sandeepkumar@students.sau.ac.in, amitkshukla@ieee.org, pranabmuhuri@cs.sau.ac.in

Abstract— The rapid growth in industrialization has proportional effect on the increase in carbon emission as well as economic growth of a nation. Nevertheless, there are many nations with unavailable information on their gross domestic products (GDPs). Therefore, primarily, this paper addresses the problem of predicting missing GDP of these nations with the help of their carbon emission data. However, the available data of these countries are insufficient for training a predictive machine learning model. So, we have focused on the emerging yet under-explored area of multi-source unsupervised transfer learning to enlarge the training domain by introducing the detection and removal of anomalies in order to build a robust prediction framework. This is empirically evaluated over the carbon emission and per capita GDP data, collected from the World Bank repository, of a number of developing countries as well as over a set of mixed (developed and developing) countries. Five different domains generated using multi-source unsupervised transfer learning framework are evaluated using three different machine learning models. The best among them is then used to predict the missing per capita GDP of a nation.

Keywords: *Multi-source unsupervised transfer learning (UTL), anomaly detection, GDP prediction, carbon emission, environmental kuznets curve, isolation forest.*

I. INTRODUCTION

GDP (Gross domestic product) is one of the quantitative measures which define the status of a nation's economy. It is dependent on several other measures such as: gross investment, exports, imports, consumption, industrialization, etc. By Industrialization, almost every nation has recorded a gain in their GDP's (GDP values) and also in their carbon emissions. It has been empirically established that GDP and carbon emissions of a nation follows Environmental Kuznets Curve (EKC). EKC claims that a monotonically increasing relationship exist among GDP and CO_2 emission. EKC postulates that GDP and CO_2 emission follow an inverted U-shaped relationship. It means that at the initial stage of economic development (industrialization), GDP is lower, and environment degradation is significant (i.e. CO_2 and other pollutants increases). However, later, when GDP reaches a threshold the quality of the environment improves (i.e. with better GDP, goal shifts towards reducing the CO_2 and other pollutants).

It has already been established in the literature that irrespective of being developed or developing country, economic growth of a country is directly proportional to its carbon emission [22]. This relationship could be easily exploited to estimate GDP of a country using its CO_2 emission. Ironically, there are many countries like, Afghanistan, Syria, Myanmar and Yemen etc. with no available information of their GDP for some particular duration of time. It is mainly because they were either war-torn or inaccessible. For such countries, it is impossible to estimate their GDP since their basic macro-economic data is insufficient.

However, these war-torn or inaccessible countries don't have sufficiently available data in order to train a machine learning (ML) model to accurately predict their missing GDP. This limitation can be efficiently handled using TL techniques in which data of some other countries of similar standards can be used for training. Since training and testing are from two different domains with different data distribution, thus it formulates to the classical case of transfer learning (TL).

As the name suggests, transfer learning uses the previously learned knowledge in a different domain, and utilizes (i.e., transfers) it to some other domains with possibly different conditions. TL has been used with some other terminologies such as: transfer of learning, domain adaptation, multi-task learning, meta-learning, etc. [4-6]. Its practical contribution was explored in various real-time and real-world applications, e.g., intelligent systems [2] and failure prediction [3] etc.

This paper uses multi-source UTL to pre-process the dataset for efficient prediction of missing per capita Gross domestic product (GDP) of a country. The term multi-source symbolizes here that a single country is not used as source domain, although numerous countries are collectively used as source domain. The target countries whose missing GDP are to be predicted are all developing countries, therefore two different kind of multi-source domains are used in this paper, one is from the collection of 20 developing countries and the other multi-source domain is from the collection of 60 countries comprising both developing and developed. The paper generates five different combinations of source domains using these two multi-source domains. Which are then used to train three supervised ML models to predict missing per capita GDP of target domain countries.

As in a real-world application framework, any of the ML approaches are highly prone to the quality of data used for training the model. These ML approaches fails to generalize in the TL framework where the training and testing occur in two different domains. In this paper, we have pre-processed the source domain data with respect to target domain data

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

using anomaly detection approach i.e. by considering the non-relevant source domain data as outliers. The term Anomalies or outliers, are interchangeably used in this paper. They are generally the points in the dataset which doesn't follow the distribution of rest of the datasets and lay far-distant from them. Presence of such anomalies may affect the prediction or estimation results very badly. Anomaly detection is widely explored data mining problem. In the literature, there are several available methods for detecting these anomalies such as: statistical [13], clustering [14], graph based [15], information measure based [16] etc. Recently, a new approach called Isolation Forest (IF) was proposed which has shown significant improvements in detecting correct anomalies.

Therefore, in this paper, for the first time we propose to use the anomaly detection, in the framework of multi-source UTL to predict the missing GDP using only the carbon emission data. Initially, anomaly detection using IF is employed to keep only those source domain data points that have similar distribution as that of the target domain data points. This selection helps in building a more robust and more accurate ML model for GDP prediction. This method of prediction is much effective than the traditional prediction approaches as it removes the anomalies in the carbon emission which lie far away from the normal behaviour of other datasets, which, if included, may substantially degrades the prediction preciseness of the ML model. Thus this paper proposes a novel multi-source UTL approach in which anomaly detection is introduced for the first time in order to force the source and target domains to have similar distribution for building an efficient machine learning model.

The rest of the paper is structured in the following manner. The related works in the area of this novel approach are discussed in Section II. Then, our proposed approaches with detailed step-by-step explanation are given in Section III. Section IV explains the datasets and the experimental results. Finally, Section V discusses the conclusion.

II. Basics and Related Work

In this section, we first briefly introduce the basic transfer learning notations and symbols, and then provide a concise review of the existing literatures.

a) Basics of TL

Some common notations and symbols which are used in TL are as follows [3] [5]:

Domain: Every domain (D) in a TL framework comprises of a feature space F which itself consists of a marginal probability distribution P(X). Mathematically, it may be represented as $\{F, P(X)\}$, $X = \{x_1, \dots \dots, x_n\} \in F$. Therefore, there are as many as F and P(X) as there are different domains.

Practically, there are two different domains to be considered as a case of TL. For computing the marginal probability distribution, features are considered constant.

Task: A domain has a task T defined by $\{Y, f(.)\}$, where $Y = \{y_1, \dots, y_m\}$. This Y is the label space and $f(.)$ represents the objective function which is estimated by training over the source domain data. The training data has two components which form a pair (x_i, y_i), where $x_i \in X$ and $y_i \in Y$. The learned function $f(.)$ predicts labels of new unseen instances.

Transfer learning: In a typical TL framework, there are source domain D_s and a target domain D_t both of which comprises their respective learning tasks T_s and T_t, respectively. A learning function $f_t(.)$ in D_t is constructed by utilizing D_s and T_s, when $D_s \neq D_t$ or $T_s \neq T_t$. Note that source domain data is used to train, and for testing the prediction performance, target domain data is used. However, if the domains and their corresponding learning tasks are similar (i.e. $D_s = D_t$ and $T_s = T_t$), then it is just the classical machine learning model.

Transductive transfer learning: When source domain (D_s) is not equal to target domain (D_t), which means either source domain feature space (F_s) and target domain feature space (F_t) are unequal or source domain marginal probability ($P_s(X)$) and target domain marginal probability ($P_t(X)$) are unequal or both of the inequalities coexist.

Domain Adaptation: In transductive transfer learning, $F_t = F_s$, though $P_t(X) \neq P_s(X)$. This special scenario is categorized as Domain Adaptation.

Inductive transfer learning: In the inductive transfer learning setting, source task (T_s) and target task (T_t) are unequal, irrespective of whether or not the source domain (D_s) is equal to the target domain (D_t). In target domain (D_t), some labeled data are definitely required to induce the objective function ($f(.)$) for precise prediction of the target domain.

Multi-Task learning: It is a special case of inductive transfer learning in which $T_s \neq T_t$, which means $Y_s \neq Y_t$ and/or $f_s(.) \neq f_t(.)$. This also implies that labels are present in source and target domains both.

Self-taught learning: It is also a special case of inductive transfer learning. However, in this case, labels are not present in the source domain.

Unsupervised transfer learning (UTL): It occurs when $D_s \neq D_t$ and both the domains have no labels.

b) Review of Related Works

We now present the related work in the area of the proposed methodology. It discusses works on carbon emission vs. GDP prediction, relevant TL approaches and the combination of both. The discussion on carbon emission related papers with the effect of economic fronts are presented next.

One of the initial works was presented by Wagner [17], in which the author utilized the relationship between carbon emission and GDP to address various econometric challenges. The same author extended the work to highlight the monotonic behavior and carbon emission elasticity of the GDP [18]. Later in 2011, Pao et al. [19] discussed the relationship between carbon emission, energy consumption,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

GDP, and Foreign Direct Investment among Brazil, Russia, India and China. An extremely detailed study on the effect of energy consumption and carbon emission on GDP of a large number of countries (i.e. 51) was studies by Chaabouni and Saidi [20]. They stated that 1% increase in CO2 emissions causes economy to increase by 0.011%.

Acheampong [21] depicted the positive effect of carbon emission on economic growth in a study of 116 countries. Establishing the same, Govindaraju et al. [12] proposed the proportionality between the carbon emission and economic strength of a country due to the modern industrialization. Similar type of relationship was discussed by Stern [22] which established that any increase in carbon emission increases GDP by showing that the carbon-income elasticity is 1.509 globally. A related work on our proposed approach was presented by Marjanovic et al. [23] in which they predicted the GDP of European Union by the using its carbon emission.

Similar problem was addressed by Shukla et al. [24] [25], with the focus on uncertainty modeling on the input dataset with the help of fuzzy sets and random fuzzy variable. In [24], they predicted GDP using carbon emission data while in [25], they predicted human development index using the same data. Aghamaleki and Baharlou [31] used IF for noise reduction in noisy web data classification using TL. Kumar and Muhuri [26] recently predicted the missing GDP's using carbon emission data but they didn't utilize anomaly detection approach. In this work they used only a single country's (developing/developed) carbon emission data for training a ML model directly.

III. Transfer Learning based GDP Prediction

In this section, we present our proposed anomaly based multi-source UTL approach to robustly predict the missing per-capita GDP of a nation. Firstly, in the first sub-section the dataset used in this paper is explained. Later, the proposed multi-source UTL methodology is described.

a) Dataset Description

This paper proposes a multi-source UTL approach to predict missing per capita GDP of Afghanistan, Myanmar, Syria, Vietnam, and Yemen. Among them Afghanistan, Syria, Vietnam, and Yemen are war-torn countries, whereas Myanmar was largely an isolated country. Due to these reasons their GDP's are unknown for a larger duration as depicted in table 1 [30].

TABLE 1: MISSING PER CAPITA GDP'S OF WAR-TORN/ISOLATED COUNTRIES

War-torn / isolated countries	Duration of missing per capita GDP values
Afghanistan	1982 to 2000
Myanmar	1960 to 1999
Syria	2008 to 2014
Vietnam	1960 to 1984
Yemen	1960 to 1989

The dataset used in the paper for experimentation is extracted from the World Bank database [30], from year 1960 to 2014, as no carbon emission data is available after 2014. Carbon emission from gaseous, liquid, and solid are mentioned in percentage (%) in World Bank database [30], which are firstly converted to real values using CO2 emission (metric tons per capita). These input parameters used for training a supervised ML model, are explained in table 2 and the output parameter is described in Table 3.

TABLE 2: DETAILS OF INPUT PARAMETERS

INPUTS	Description
1	Carbon emissions from gaseous fuel consumption (metric tons per capita)
2	Carbon emissions from liquid fuel consumption (metric tons per capita)
3	Carbon emissions from solid fuel consumption (metric tons per capita)

TABLE 3: DETAILS OF OUTPUT PARAMETER

OUTPUT	Description
1	Per capita GDP at current US$

b) Proposed multi-source UTL methodology

The available dataset of the above mentioned countries (table 1) are not sufficient enough to train and validate a supervised ML model. So, to overcome this unavailability of sufficient data for training we used the TL approach as conceptualized in Algorithm 1.

Algorithm 1: TL methodology to overcome data shortage for training an ML model

Output: Prediction of missing GDP values of war-torn / isolated countries

Input: Source domain datasets, Target domain dataset

1. Source domain datasets generation:
   a. Collect Source domain datasets (table 2 and table 3) from various other countries of the world, excluding Target domain country.
   b. Pre-process the Source domain datasets using multi-source UTL to generate five different source domain datasets.
2. Train a supervised ML model over this source domain
3. Validates the prediction accuracy of the model over the target domain dataset
   a. Use RMSE for calculating prediction accuracy
4. Repeat step 2-3 for different kind of supervised ML model.
5. Repeat step 2-4 for each of the five Source domain datasets generated in step 1.
6. Select the most accurate model (step 1-5) to predict missing per capita GDP values of the target domain country.

The flow diagram of this algorithm is also depicted diagrammatically in Fig. 1 and is explained stepwise as follows:

Step 1: Multi-source domain datasets:

Multi-source domain data here represents the carbon emission and per capita GDP dataset (table 2 and table 3) collected from various other countries of the world, except the countries mentioned in table 1 (we have referred these

countries as target domain countries). These multi-source domain datasets are preprocessed using multi-source UTL methodology (*Algorithm 2*) to generate five different source domain datasets.

Step 2: ML Methodology

Three extensively used supervised machine learning regression models are utilized in the paper viz. KELM (kernel extreme learning machines) [7-8], GRNN (Generalized Regression Neural Network) [9], SVR (Support Vector Regression) [10-11]. These models are trained over each of the source domain datasets (generated in step 1) using carbon emission data for per capita GDP value predictions.

Step 3: Prediction precision evaluation (Validation)

These ML models once trained using one type of the source domain are now tested for their prediction preciseness in estimating per capita GDP of a target domain country using root mean square error (RMSE). The carbon emission dataset of a war-torn / isolated country corresponding to those years for which their per capita GDP values are available, is used for validation of these ML models. Step 4-5: for a Target domain country a total 15 different ML models are trained i.e. GRNN, ELM and SVR for each of the five source domains generated in step 1.

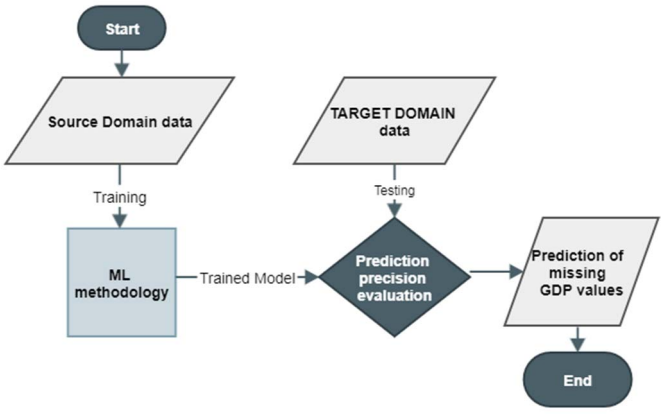

Fig.1: TL approach to train an ML model in case of data shortage

Step 6: Missing GDP Prediction

Out of these 15 trained ML models, the trained model that predicted the target domain's per capita GDP values with least RMSE value (in step 3), is then used to predict the missing per capita GDP values of that target domain country. Initially two different multi-source domain datasets are collected from [30], one dataset is from 20 developing countries and another one from 20 developed countries.

The first source domain of following 20 developing countries is henceforth termed as Source Domain 1:
{'Egypt', 'Tunisia' , 'Cameroon', 'Kenya', 'South Africa', 'Zimbabwe' , 'Nigeria', 'Burkina Faso', 'Indonesia' , 'India', 'Jordan', 'Thailand', 'Cuba', 'Haiti', 'Ethiopia', 'Pakistan', 'Bangladesh', 'Iran', 'Zambia', 'Sudan'}

Developed countries data comprises of following countries:
{'Austria', 'Australia', 'Canada', 'Japan', 'New Zealand', 'United States', 'France', 'Germany', 'Italy', 'United Kingdom', 'Norway', 'Sweden', 'Spain', 'Portugal', 'Netherlands', 'Greece', 'Finland', 'Hungary', 'Switzerland', 'Poland',}

However, at later stage in order to demonstrate the capabilities of our proposed UTL in selecting the right domain points for constructing an optimum training dataset, we mixed 20 developed countries data with an additional 40 countries carbon emission data, mostly from developing countries and termed this whole dataset of 60 countries as Mixed Countries data. Its reasoning is empirically stated in next section. These 40 additional countries are following:
{'European Union , 'Benin ', 'Burundi', 'Central African Republic ', 'Chad', 'Equatorial Guinea', 'Cambodia', 'China', 'Fiji', 'Malaysia', 'Papua New Guinea', 'Albania', 'Armenia', 'Belarus', 'Bosnia and Herzegovina', 'Bulgaria', 'Czech Republic', 'Georgia', 'Monaco', 'Tajikistan', 'Turkey', 'Ukraine', 'Uzbekistan', 'Argentina', 'Brazil', 'Mexico', 'Peru', 'Venezuela', 'Algeria', 'Morocco', 'Bhutan', 'Nepal', 'Sri Lanka', 'Angola', 'Eritrea', 'Ghana', 'Madagascar', 'Mauritius', 'Mozambique', 'Uganda'}

The choice of these countries is totally arbitrary. One can choose any country data. Now, it contains European Union data which itself comprises of 28 different countries and also 59 developed, developing or least developed countries. Henceforth, this multi-source domain dataset is termed as mixed counties dataset. A single country's data represents a single source domain, as it is unique in itself, and we have collected source domain data from numerous countries, which have enormous data distribution differences with each other, due to differences in their economic condition/policies or their geographical locations etc. That's why we called it a multi-source domain dataset.

Algorithm 2: Unsupervised Transfer Learning using Isolation Forest

Output: Source Domain 1, Source Domain 2, Source Domain 3, Source Domain 4 and Source Domain 5
Input: Source domain 1, Mixed countries dataset

1. Merge Source Domain 1 and Mixed countries dataset and term it as Source Domain 2
2. Train IF (an Unsupervised ML approach) on Source domain 1, using only table 2 parameters
3. Create Source Domain 3 after removing anomalies from Source Domain 1 through the trained IF (from step 2)
4. Create Temp domain by eliminating anomalies of Mixed countries dataset using the IF (trained in step 2).
5. Generate Source Domain 4 by merging Source Domain 1 and Temp domain (from step 4)
6. Build Source Domain 5 by merging Source domain 3 (step 3) and Temp domain (step 4)
7. Delete Temp Domain

Return Source Domain 1, Source Domain 2, Source Domain 3, Source Domain 4 and Source Domain 5 for *Algorithm 1*.

The above mentioned multi-source domains (especially the mixed countries) are from different distribution than the target domains (countries depicted in table 1). This may cause performance degradation of the model trained on it for

solving target domain task, if used as it is. To overcome this issue of performance degradation due to difference in data distributions, this paper employs multi-source UTL using IF for boosting the efficient utilization of source domain during TL. In the proposed multi-source UTL, IF detects these anomalous points of source domain that are not significant during TL. From now on we term all these points of source domain as anomalies for being of no use in building a target domain specific predictive model.

We have used IF for anomaly detection because it isolates anomalies based on tree path length, which is more efficient than density or distance based algorithm. As it can easily detect both scattered and clustered anomalies without much parameter adjustments, while density or distance based anomaly detection approaches require too much fine tuning of parameters in such cases. Moreover, IF is highly robust to swamping and masking; swamping occurs when normal data points are wrongly detected as anomalies, as either they are heavily scattered or their numerical strength increases; masking of anomalies takes place when anomalies are very large and occurs in dense clusters, masking themselves as normal data points.

Algorithm 2 explains the proposed multi-source UTL approach incorporating IF methodology. The flow diagram of this algorithm is depicted in Fig. 2. In this IF is trained on developing nation dataset (Source domain 1) for anomaly detection, as the target countries whose missing GDP are to be predicted are also developing countries. Then it predicts the anomalies not only in mixed nation dataset, but also in developing nation dataset (i.e. it removes those data instances which do not conforms to the majority of the developing countries data instances). More importantly, it introduces five different source domains. The reason of construction several of these domains is to validate the use of anomaly detection for its necessity in our proposed approach.

Algorithm 2 generates five different types of source domain data, explained as follows step by step:

Source Domain 1 (Input): It is just the collection of the carbon emission and GDP data of 20 developing countries without any modification.

Source Domain 2 (Step 1): It is a collection of the 20 developing countries data (source domain 1) and the mixed 60 countries dataset (both developed and developing). IF Training (Step 2): IF is an unsupervised ML methodology and the target countries whose missing GDP are to be predicted are also developing countries. So, IF is trained using only Source Domain 1 features as shown in Table 2.

The rest of the three domains are created after the process of anomaly detection and removal by the trained IF.

Source Domain 3 (Step 3): This is the processed 20 developing countries data, after removal of its anomalies as detected by IF (i.e. trained in step 3).

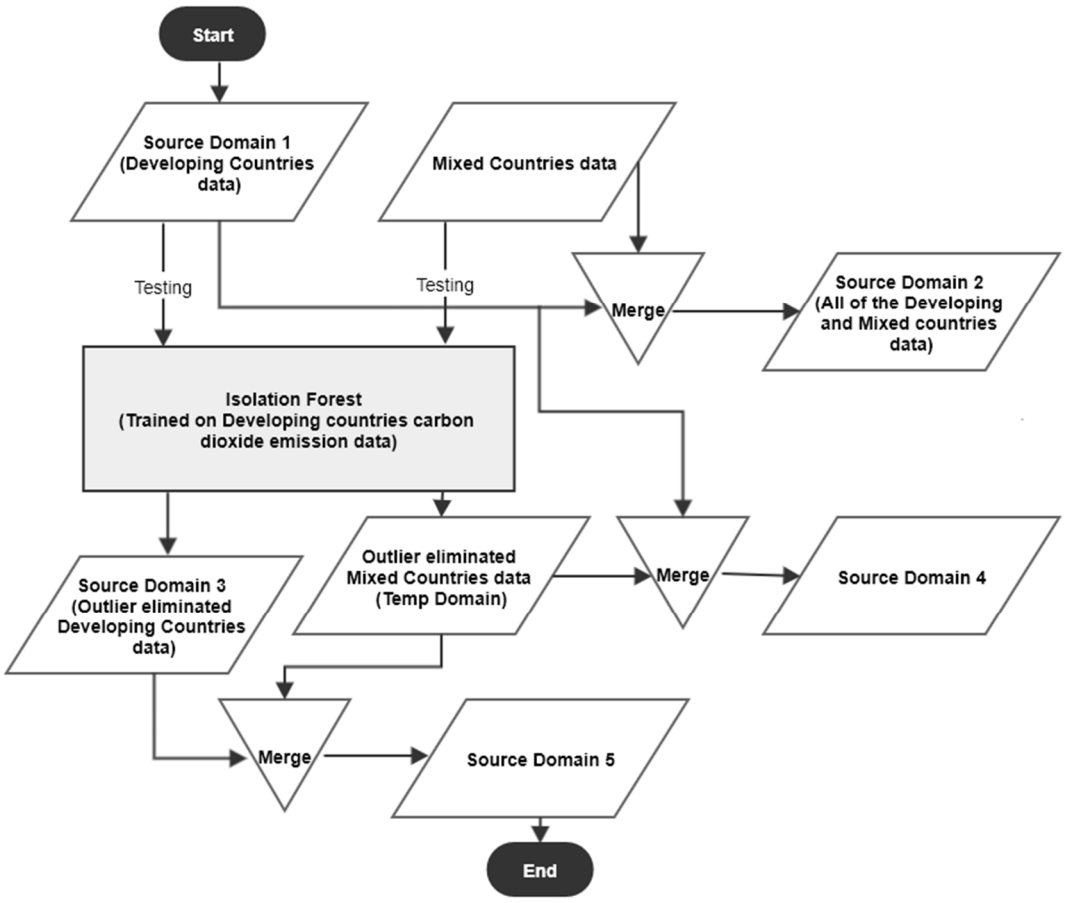

Fig. 2: Unsupervised Transfer Learning using Isolation Forest

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Temp Domain (Step 4): It is created from Mixed Countries dataset, after removal of its anomalies which are detected by this trained IF.

Source Domain 4 (Step 5): It is the combination of the 20 developing countries (Source domain 1) data and Temp domain dataset of step 4.

Source Domain 5 (Step 6): This domain data combines the source domain 3 (step 3) and Temp Domain dataset (step 4).

In step 6 of the *Algorithm 2*, Temp Domain is deleted as it is not required anymore. In step 7 these five different Source Domains (i.e. Source Domain 1, Source Domain 2, Source Domain 3, Source Domain 4 and Source Domain 5) are then further used in *Algorithm 1* (Fig. 1) for training an ML model.

In step 2 of *Algorithm 2*, IF is trained using only Source Domain 1 (i.e. 20 developing countries data). It is because the performance of ML approach on target domain (*Algorithm 1*), heavily depends on the source domain dataset used for its training. That's why the source domain dataset should be of similar distribution as that of the target domain dataset. Since, all the target domain countries mentioned in table 1, are also developing. So, we trained IF on Source Domain 1 (i.e. 20 developing countries data). So this trained IF now classify all those points as anomalies which don't conform to the data distribution of developing countries.

Now from this universal behaviour, it may be assumed that even in developing nation there could be few countries having comparatively higher GDP's in later years, which needs to be detected as anomalies; as their inclusion during training of an ML model in predicting the GDP of low-income developing nations may degrade its efficacy. Similarly, in mixed nations data, there will be a large number of such anomalies. Therefore, to detect and eliminate them we trained an IF on developing nation's carbon emission data and then it is employed to predict and remove the anomalies both in the developing nations as well as in mixed nations carbon emission data. By this this way the resultant dataset will turn out to be much more useful for training an ML model for precise estimation of missing per capita GDP of developing countries.

IV. Empirical Results

This section empirically shows the validity of our proposed approach to predict the missing GDP of nations with carbon emission data. There are three sub-sections: Details for data pre-processing, experimental results and predictions, and strength and weakness of the proposed approach.

a) Multi-source UTL based data pre-processing

Fig. 2 shows the visual representations of the input dataset (Table 2) for developing and developed countries. Developing countries have total of 1025 data points while developed countries have total of 998 points (after removing missing values from it).

As can be seen from Fig. 3, developed countries have very large values as compared to developing countries. Therefore, all of the developed countries dataset cannot be used to train an ML model to predict developing countries missing GDP, as it will cause drastic performance degradation of the model.

It is because data distributions of developed nation data lie away from the data distribution of developing countries data. Also, as can be seen in Fig. 3, developing countries data (denoted as circles), is mostly clustered at origin, however there are a few point which are scattered far away due to their high values. For a ML model to precisely predict a low-income developing country's GDP these anomalies must be removed from training samples. Therefore, here we used IF for detection and removal of these anomalies. IF was first trained on developing countries data in an unsupervised way (training without output parameters). Then both the datasets are passed through the trained IF [32] to detect anomalies. We observe that 308 and 981 data points are detected as anomalies from the developing and developed countries, respectively.

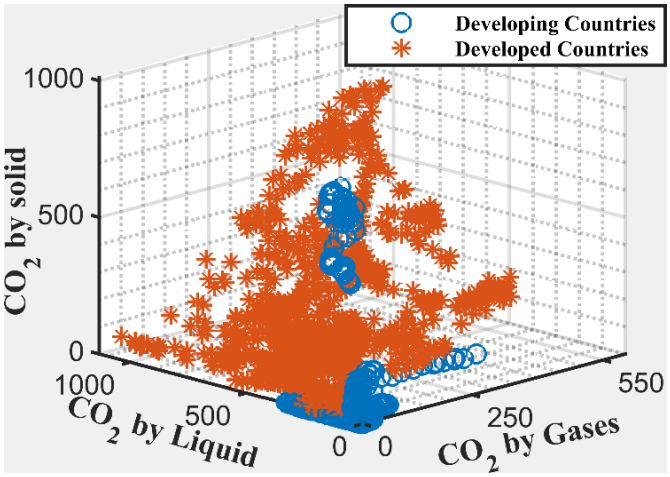

Fig.3: Representation of developing and developed nation data

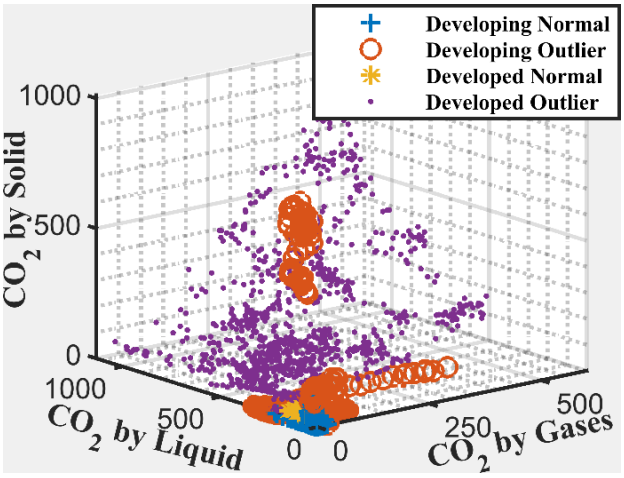

Fig.4: Developing and developed nation CO2 emission data with outliers.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Fig. 4 shows the visual representation of the normal points and detected anomalous points in the developed and developing nation's data by IF. Further, it can be clearly seen from the figure that data points away from the origin are detected as anomalies, irrespective of whether they are from developing or developed. Based on these results, it can be deduced that if we have a large dataset of many countries regardless of being developed or developing, necessary data points may be extracted using the above trained IF. Note that after the anomaly detection, we are left with only 17 useful data points from the developed countries out of 998 total data points. Therefore, to demonstrate the capabilities of our proposed multi-source UTL in selecting the right domain points for constructing an optimum training dataset, we mixed these 20 developed countries data with another 40 countries carbon emission data, as described in section III.

b) Results and Predictions

In this Section, missing per capita GDP of five developing countries are predicted using different ML models, separately . trained on each of the five source domains as shown in Fig. 1 using multi-source UTL. These countries are: Afghanistan, Myanmar, Syria, Vietnam, and Yemen. The available per capita GDP's of these countries is used for validation,in order to choose the best performing ML model and the respective source domain on which it is trained. Table 4 compiles the results (best values in bold) for all of the above mentioned five source domains on GRNN, SVR and ELM.

TABLE 4: RMSE ERROR USING DIFFERENT SOURCE DOMAINS

Countries	Source Domain	GRNN	SVR	ELM
Afghanistan	Source Domain 1	780.05	318.07	704.20
	Source Domain 2	5156.62	975.75	1097.92
	Source Domain 3	299.14	211.44	297.43
	Source Domain 4	635.19	244.55	565.83
	Source Domain 5	418.15	**207.65**	406.34
Myanmar	Source Domain 1	592.99	434.99	549.05
	Source Domain 2	4843.13	748.95	1074.37
	Source Domain 3	440.09	484.06	440.24
	Source Domain 4	500.84	461.86	471.84
	Source Domain 5	428.27	487.04	**426.65**
Syria	Source Domain 1	504.16	645.60	473.36
	Source Domain 2	4774.61	591.76	4038.25
	Source Domain 3	665.87	751.68	657.09
	Source Domain 4	501.66	705.39	**427.16**
	Source Domain 5	583.43	753.17	529.07
Vietnam	Source Domain 1	654.03	558.55	611.91
	Source Domain 2	4848.34	788.17	1371.48
	Source Domain 3	566.01	610.54	565.14
	Source Domain 4	584.04	586.65	541.69
	Source Domain 5	544.71	612.95	**537.87**
Yemen	Source Domain 1	490.32	489.14	476.66
	Source Domain 2	4752.21	631.96	1813.95
	Source Domain 3	502.42	577.45	500.89
	Source Domain 4	433.69	541.39	**429.59**
	Source Domain 5	443.47	580.85	436.01

Fig. 5 shows the two set of sub-figures for five countries. Figures in left column shows countries with the duration of their missing per capita GDP. While the right column figures depict the predicted per capita GDP values for each of these five countries using best performing models from Table 4 (mentioned with bold values). Except Syria, predicted GDP's of missing duration appears to be almost constant, suggesting a stagnant economy during these years. However, in the case of Syria, the prediction of the missing GDP values is decreasing over the years, suggesting that economy gradually collapsing due to the civil war during that period.

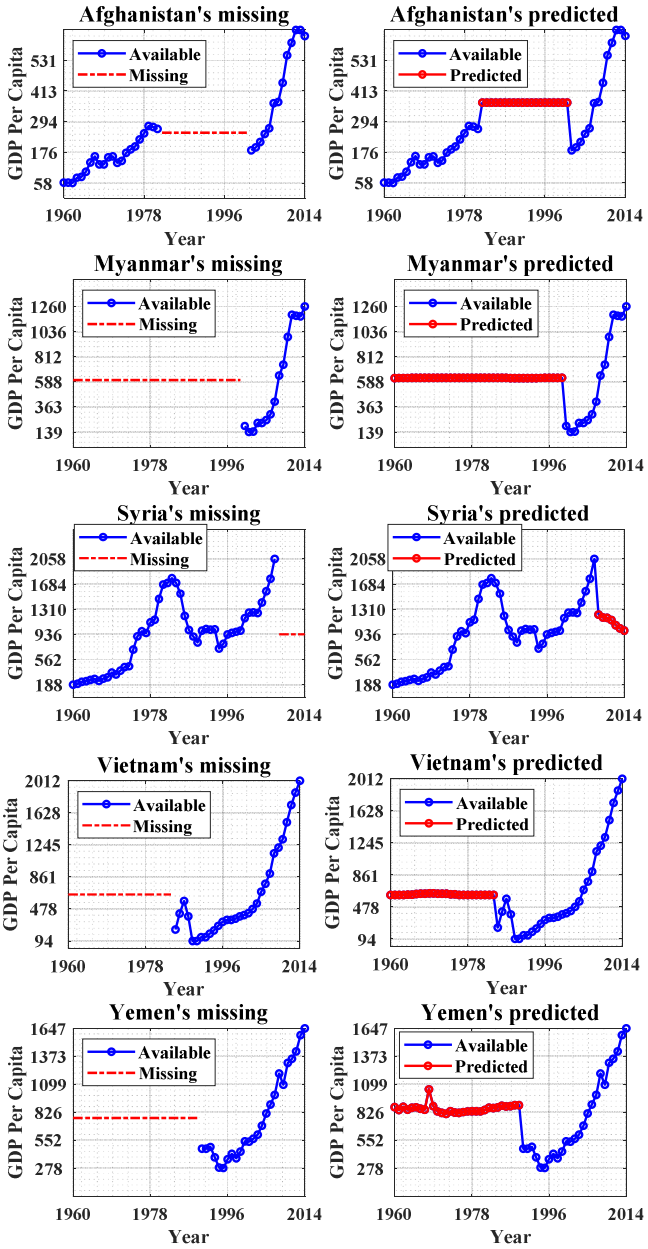

Fig. 5: Missing and predicted per capita GDP of five nations: Afghanistan, Myanmar, Syria, Vietnam, and Yemen

V. DISCUSSION AND CONCLUSION

This paper has proposed a novel approach for prediction of missing per capita GDP of various war-torn/isolated countries using their carbon emission data in the framework of transfer learning technique. In this approach, the model is trained over different countries dataset where data availability is in abundance. However, abrupt use of dataset of these countries may be detrimental if they have highly

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

dissimilar data distribution. In order to tackle this problem this paper proposed a multi-source unsupervised transfer learning (UTL) approach using isolation forest (IF). Firstly, IF is trained in an unsupervised way over 20 low-income countries dataset, and then this trained IF is used to remove all those values that are outlier to majority of these datasets. Also in order to increase our dataset and depict the pros and cons of the proposed multi-source UTL approach, it is also applied to a mixture of 60 developed, developing countries dataset in order to select the most relevant useful data instances and discard the non-useful instances as anomalies. These anomalies otherwise degrade the preciseness of an ML model in predicting missing GDP of a developing country, if included during training.

At the end, we may firmly conclude that our novel approach for GDP prediction using carbon emission data in the framework of transfer learning technique is highly effective. The inclusion of anomaly detection technique makes the approach more robust and pragmatic to real-world applications. Moreover, it provides an effective and automated way to increase relevant dataset repository by extracting the useful data from a related domain. Hence, it provides ability to reuse the available domains data in a more effective way to deal with a newer related domain problem. In future, we would enhance our approach to estimate the missing GDP's of such developed countries like Switzerland and Poland.

ACKNOWLEDGEMENT

First author gratefully acknowledges the financial support received from the Department of Science and Technology, Government of India in the form of INSPIRE fellowship.

REFERENCES

[1] P. K. Muhuri, A. K. Shukla, and A. Abraham. "Industry 4.0: A bibliometric analysis and detailed overview," *Engineering applications of artificial intelligence,* vol. 78, pp. 218-235, 2019.

[2] V. Behbood, J. Lu, and G. Zhang. "Text categorization by fuzzy domain adaptation," In 2013 IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE), pp. 1-7. IEEE, 2013.

[3] V. Behbood, J. Lu, G. Zhang, and W. Pedrycz. "Multistep fuzzy bridged refinement domain adaptation algorithm and its application to bank failure prediction," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 1917-1935, 2015.

[4] J. Shell, S. Coupland, "Fuzzy transfer learning: methodology and application," *Information Sciences*, vol. 293 pp. 59–79, 2015.

[5] S. J. Pan, Q. Yang, "A survey on transfer learning," *IEEE Tran. on knowledge and data engineering*, vol. 22, no. 10, pp. 1345-1359, 2009.

[6] S. Kumar, A. K. Shukla, P. K. Muhuri, and Q. M. D. Lohani. "Atanassov Intuitionistic Fuzzy Domain Adaptation to contain negative transfer learning," In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2295-2301. IEEE, 2016.

[7] G. Huang, Q. Zhu, and C. Siew. "Extreme learning machine: a new learning scheme of feedforward neural networks," In Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, vol. 2, pp. 985-990. IEEE, 2004.

[8] G. Huang, G. Zhu, and C. Siew. "Extreme learning machine: theory and applications," *Neurocomputing,* vol. 70, no. 1-3, pp. 489-501, 2006.

[9] J. Zhai, L. Zang, and Z. Zhou. "Ensemble dropout extreme learning machine via fuzzy integral for data classification," *Neurocomputing,* vol. 275 pp. 1043-1052, 2018.

[10] V. Vapnik. "The nature of statistical learning theory," *Springer Science & Business Media*, 2013.

[11] J. Bi and K. P. Bennett, "A geometric approach to support vector regression," *Neurocomputing*, vol. 55, no. 1, pp. 79–108, 2003.

[12] V. G. R. C. Govindaraju, and C.F. Tang. "The dynamic links between CO2 emissions, economic growth and coal consumption in China and India," *Applied Energy*, vol. 104, pp. 310–318, 2013.

[13] M. Kemmler, E. Rodner, E.-S. Wacker, and J. Denzler, One-class classification with gaussian processes," Pattern Recognition, vol. 46, no. 12, pp. 3507-3518, 2013.

[14] S. S. Bama, M. I. Ahmed, and A. Saravanan. "Network intrusion detection using clustering: a data mining approach," *International Journal of Computer Applications*, vol. 30, no. 4, pp. 14-17, 2011.

[15] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey*," Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626-688, 2015.

[16] W. Lee and D. Xiang. "Information-theoretic measures for anomaly detection," In *Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001*, pp. 130-143. IEEE, 2000.

[17] M. Wagner. "The carbon Kuznets curve: a cloudy picture emitted by bad econometrics?" *Resource and Energy Economics*, vol. 30, no. 3, pp. 388-408, 2008.

[18] M. Wagner. "The environmental Kuznets curve, cointegration and nonlinearity," *Journal of Applied Econometrics*, vol. 30, no. 6, 2015.

[19] H.T. Pao, C.-M. Tsai, Multivariate Granger causality between CO2 emissions, energy consumption, FDI and GDP: evidence from a panel of BRIC (Brazil, Russian Federation, India, and China) countries, *Energy,* vol. 36, pp. 685–693, 2011.

[20] H.-T. Pao, and C.-M. Tsai. "Multivariate Granger causality between CO2 emissions, energy consumption, FDI (foreign direct investment) and GDP (gross domestic product): evidence from a panel of BRIC (Brazil, Russian Federation, India, and China) countries," *Energy*, vol. 36, no. 1, pp. 685-693, 2011.

[21] S. Chaabouni, and K. Saidi. "The dynamic links between carbon dioxide emissions, health spending and GDP growth: a case study for 51 countries," *Environmental research*, vol. 158, pp. 137-144, 2017.

[22] A. O. Acheampong. "Economic growth, CO2 emissions and energy consumption: What causes what and where?," *Energy Economics,* vol. 74, pp. 677-692, 2018.

[23] D. I. Stern. "Between estimates of the emissions-income elasticity," *Ecological Economics*, vol. 69, no. 11, pp. 2173-2182, 2010.

[24] V. Marjanović, M. Milovančević, I. Mladenović. "Prediction of GDP growth rate based on carbon dioxide (CO2) emissions," *Journal of CO2 Utilization,* vol. 16, pp. 212-217, 2016.

[25] A. K. Shukla, S. Kumar, R. Jagdev, P. K. Muhuri, and Q. M. D. Lohani. "Interval Type-2 Fuzzy weighted Extreme Learning Machine for GDP Prediction," In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2018.

[26] A. K. Shukla, S. Kumar, B. Mor, and P. K. Muhuri. "Random Fuzzy Variable based Uncertainty Modelling for the Prediction of Human Development Index using CO2 Emission Data," In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2117-2124. IEEE, 2018

[27] S. Kumar, and P. K. Muhuri. "A novel GDP prediction technique based on transfer learning using CO2 emission dataset," *Applied Energy,* vol. 253, pp. 113476, 2019.

[28] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413-422. IEEE, 2008.

[29] D. H. Wolpert, and W. G. Macready. "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67-82, 1997.

[30] F. T. Liu, K. M. Ting, and Z.-H. Zhou. "Isolation-based anomaly detection," ACM *Transactions on Knowledge Discovery from Data (TKDD),* vol. 6, no. 1, pp. 1-39, 2012.

[31] U. nation N. World Bank; https://data.worldbank.org/ accessed on June 2019, GDP of the World, (n.d.).

[32] J. A. Aghamaleki, and S. M. Baharlou. "Transfer learning approach for classification and noise reduction on noisy web data." *Expert Systems with Applications*, vol. 105, pp. 221-232, 2018.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel et al. "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.