

Addressing Sustainability Issues through Transfer Learning

Vaibhav Gupta

DS440 – DS Capstone Course
Pennsylvania State University
State College, USA
vvg5125@psu.edu

Ally Racho

DS440 – DS Capstone Course
Pennsylvania State University
State College, USA
axr402@psu.edu

Divya Rustagi

DS440 – DS Capstone Course
Pennsylvania State University
State College, USA
dpr5375@psu.edu

Abstract — The team was assigned with the task of using transfer learning to address sustainability issues. Air quality has become more evident in recent years due to climate change and pollution becoming monitored more. The EPA tracks many cities' air quality index (AQI) and keeps historic records of the data. There are data insufficient cities however that the EPA does not keep information on. To overcome the issue of data insufficient cities and the lack of AQI data, we proposed a neural network which learns on data connected to AQI for a source city – Manhattan, NY – and then transfers the knowledge to target cities – Chicago, IL and Philadelphia, PA – and predicts the AQI measurements for a given week and year between 2018 – 2022. It was found that normalized RMSE for the Chicago model was 0.224 and the Philadelphia model was 0.200. Given these results, our models are predicting relatively accurately. We also noted that the COVID-19 pandemic may have been a factor in the accuracy of our model for years 2020 and 2021. A Streamlit web application was created to demonstrate the model's findings and allow for user-interaction with the model.

Keywords—transfer learning, air quality, neural network

I. INTRODUCTION

Air quality is a crucial metric for building sustainable cities. Smaller or less developed regions do not have enough data, or have poor quality of data, lacking labels to correctly identify ground truth of these given metrics. Hazleton, PA is a city we observed to have this issue occur. Other regions – New York City, Los Angeles, Tokyo, etc. – have a plethora of data available and can aid in bridging the gap between these smaller areas. Transfer learning can be used in helping lessen the impact of data insufficiency as we can use information from a source city to predict the sustainability factors, specifically air quality data (AQI), of a data insufficient target city.

A neural network which is trained on source city data is transferred to our target cities to predict the AQI measurements. The data which this model was trained on includes weather metadata (maximum/minimum temperature, average wind speed, precipitation), pollution emission values, traffic values (vehicle/bus/pedestrian volumes, gas prices, average volumes). The granularity of this data was weekly; thus, our overall model makes weekly predictions for the years 2018-2021.

Some challenges the team was faced with include identifying what the granularity of the data should be along

with how to correct datasets with different granularity and recognizing the need for some substantial data in data insufficient cities for the model to use in prediction. Huge datasets were also a factor with RAM and time-complexity. Additionally, new methods and technology had to be learned and understood throughout the project which took time.

The team's neural network resulted in an overall relatively accurate and efficient transfer learning model. The normalized RMSE for all final models including source and target cities was approximately 0.2 indicating an effective model.

The paper will be presented as follows:

- Abstract
- Introduction
- Related Work
- Methodology
- Results
- Discussion
- Conclusions
- Future Work
- Contributions
- References
- Appendix

II. RELATED WORK

Transfer Learning-based LSTM

Missing data is a commonly occurring issue with water quality data due to the expensive infrastructure needed to implement and inadequate sampling procedures. This has already been evident in our research of not only our data insufficient cities but also the cities of Philadelphia and Chicago. This paper proposes TrAdaBoost-LSTM which incorporates the LSTM model and transfer-learning technique to leverage related knowledge from complete datasets to overcome missing data.

The methodology behind this model includes data preprocessing and algorithm execution. Firstly, the data is cleaned and normalized and then target and source domains are determined based on similarities. TrAdaBoost-LSTM is then employed on the data. Because of the large scale of the missing data, transfer learning is applied to complete the incomplete sequences of the target domains. The paper states, "The motivation of the TrAdaBoost algorithm is based on the

concept that a portion of the inherent data information in the source domain with a similar distribution as the target training data may be more valuable than other portions of data” [3]. TrAdaBoost-LSTM constructs a series of weak learners using LSTM for regression-based prediction and TrAdaBoost for the encoding layer. As the iterations increase, the weights of similar source to target domains increase and vice versa. To evaluate the model, RMSE, mean absolute percentage error, mean absolute error, and R2 are calculated. When compared with other baselines, TrAdaBoost performs the best on datasets in which large amounts of data are missing.

Overall, this proposed method would be a beneficial starting point in our project for predicting air and water quality. Our end goal is predicting these sustainability issues on a data insufficiency city – Hazleton, PA– which has a plethora of missing historical data. Using this method’s technique would be a great advantage to overcoming this missing data. This research paper is also a good reference point in helping us begin the process of analyzing air quality data and enhancing our transfer learning technique in predicting air quality of data insufficient cities.

Transfer Learning for Thermal Comfort Prediction

There is an unavailability of comprehensive labeled thermal comfort data. Because of this, the team adopts transfer learning to leverage knowledge from a source domain (same climate zone) to a data insufficient target domain (different climate zone). The proposed model is a transfer learning based convolutional neural networks long short-term memory. (TF CNN-LSTM) Transductive transfer learning is the type of transfer learning used because it gains the well-learned knowledge from the source domain and utilizes it to improve the performance of prediction in the target domain. Likewise, we will also be employing this method in our project.

TF CNN-LSTM contains a convolutional layer, two LSTM layers, two dense layers, and is followed by an output layer to learn the spatio-temporal characteristics of the data. [4]. The transfer learning technique can gain information at different levels of the model. The target classifier is then able to better predict thermal comfort using the knowledge of the weights from the source domain. Since the team uses different climate zones, the lower layer of the model is affected, and thus retrained on the target domain. The higher layers of the model are then retained and aid in the prediction of source to target [4]. The baselines used in comparing this model include PMV factors, machine learning models, deep learning models, and deep learning models with transfer learning. These are based on confusion matrix, accuracy, precision, F1 score, and MCC. It is observed that TF CNN-LSTM outperforms other state-of-the-art models. Since the model can analyze both spatial and temporal data while also transferring higher-order relations from source to target domains, the model demonstrates an enhanced performance.

One issue with this model stated is that the data sets may not sufficiently cover the different climate zones. I believe using a model like this to assist in our prediction of air quality would be beneficial in enhancing this state-of-the-art model.

The model is using a transductive transfer learning technique. Implementing this approach to transfer learning and extending it to our dataset of different climate zones may further progress this model.

III. METHODOLOGY

Data Collection

The team collected data from various sources, EPA.gov for the historic AQI measurements, AirNow.gov for historic pollution emission values (pm25, co2, o3, no2, so2), NOAA.gov for historic weather metadata, and open data sites for each city (opendata.cityofnewyork.us, opendataphilly.org, data.cityofchicago.org). The open data sites include data on all types of topics ranging from traffic, business, environment, education, health, etc. We focused mainly on data involving traffic and construction as this more closely pertains to air quality factors. Due to the extremity of data available, we filtered the dates between Jan 1, 2018, to Jan 1, 2022. We chose these dates as we may also be able to learn more about how a cities’ air quality was possibly affected from COVID-19. Additionally, the data was based on certain counties, New York – Manhattan, Chicago – Cook County, Philadelphia – Philadelphia County, Hazleton – Luzerne.

To quickly scrape data for our source city, we wrote an R script which downloaded datasets from the open data site. This code is available on our GitHub [A.1]. Datasets pertaining to the target cities of Philadelphia, Chicago, and Hazleton were manually downloaded.

Data Preprocessing

Once all the data was collected, the data for each city was cleaned, analyzed, and compiled into master datasets for each. The cleaning included removing columns with more than 30% NAs, transforming the date column into day, month, year, adding a week column depending on that date for each year as well as adding city and county columns, and fixing data types if needed. EDA for the datasets was conducted after collection to determine which variables appeared most important via PCA and to see if any data appears invalid to include in the master datasets. EDA reports also gave us an overview of the dataset’s characteristics and statistics. The AQI data for Philadelphia was not weekly thus aggregations had to be completed before compilation. Other datasets also had values taken multiple times per week, so the averages were calculated. Prior to completing the master datasets, distributions comparing the AQI of the target with source cities were generated.

The datasets for all cities are organized by the weeks of the year (2018-2021) for a total of 212 records (53 weeks x 4 years). New York City dataset is comprised of 31 features, Chicago has 20 features, Philadelphia has 18 features, and Hazleton (data insufficient city) has 11 features. The features include week, month, year, county, city, weather data (temperature max, temperature min, average wind speeds, precipitation), pollution emissions (pm25, co2, o3, no2, so2), traffic information (vehicles speeds, vehicle/bus/pedestrian/bike volumes, taxi trip miles, and gas averages) and the average AQI index for the

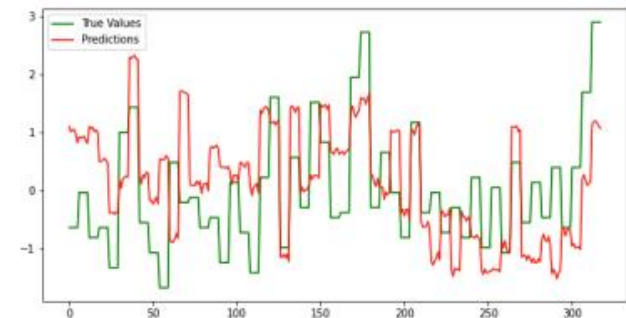
given week (which will be used to determine overall accuracy of model). After all data was compile to a master set, it was uploaded to Microsoft Azure for management and storage.

Modeling

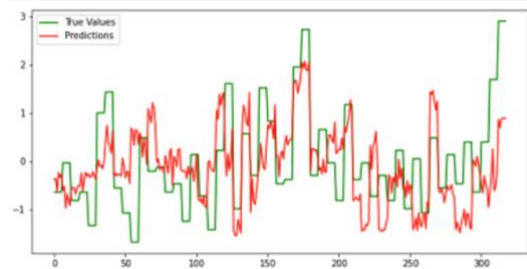
The New York dataset was first split into train and test sets. We used 2018-2020 as our training data and 2021 for the testing. It was then scaled using Standard Scalar. Three models were used – RandomForestRegressor, SVM and Custom Neural Network. The metrics used and given scores for each are presented below with the best outcome of the test data bolded.

Random Forest	R2	3.9
Regressor		
SVM	R2	-18.8
Custom Neural	MSE	1.1399
Network		

We then trained the data on the neural network since it had the best outcomes. Below is the plot of the true AQI values (green) against our predicted values (red).

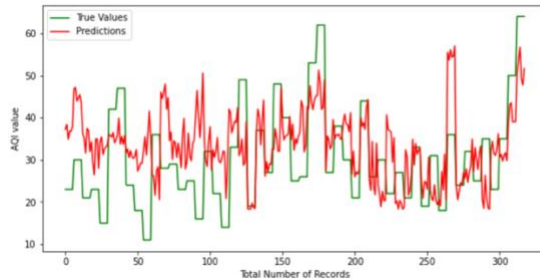


Afterwards, we tried a SDV Gaussian Copula on the model. The distributions found for this model were quite like our train set. The data was scaled and fit to the neural network. Similar outputs as previously, synthetic data MSE = 1.242 (compared to previously found 1.139). The synthetic and original training sets were then combined and shuffled for variability. We find the MSE of the model to have improved to 0.8811. The plot of true/predicted values was once again calculated.

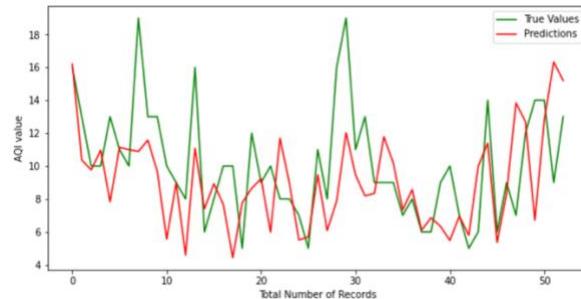


We can note that the predictions appear to follow the true values more closely. This is possibly due to the increase in data (936 original rows to 1872 rows).

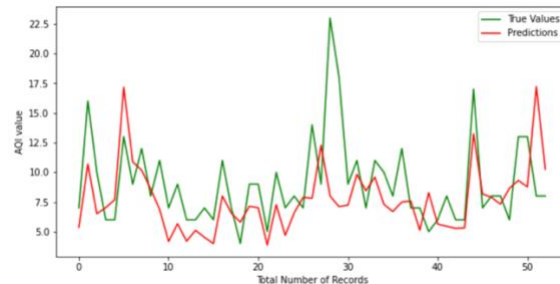
The neural network was optimized and then used in creating the transfer learning model. The plot for the transfer learning model is shown below. This proved to be our most accurate model with a 0.201 normalized RMSE.



From this model, we were able to transfer the leaned knowledge to the target cities of Philadelphia and Chicago. Below is the plot found for the AQI values of 2021 in Chicago. Our normalized RMSE was found to be 0.247.



We then transferred the model to Philadelphia, with the plot shown below. The AQI values being predicted are for 2021. Our normalized RMSE for was found to be 0.200.



We did then try to run the model on Hazleton, but ran into the issues of extreme data insufficiency, and thus could not get any accurate predictions. We also would not be able to have verified any predictions made due to the lack of historic data.

To overcome the issue of the model only being tested with data of 2021, we created synthetic data which was used to create test sets from 2018-2021. The chi square distribution of these models was checked and plotted. This data was then what was used for web application.

All this code is available on our Google Colab. [A.2]

Web-Application

The final step in the project was to create a website which would be an interactive site for users to play with the model and find out certain AQI predictions given their input. This application was built using Streamlit.

This site is a dashboard which gives a description of the project and then asks the user to input the desired city (Philadelphia or Chicago), desired week (1-53) and year (2018 – 2022). Additionally, there is a dropdown that lets the user choose if they would like to instead predict multiple weeks of AQI values or if they would like a monthly averaged AQI measurement for a specific year. Once the user chooses their preferences, a map of where the specific AQI value was calculated from is displayed, the actual and predicted values are displayed along with a chart of what the AQI value indicates, the plots of the model, normalized RMSE of the model, and the dataset used in the prediction.

Below is a snippet of the site's home screen and the second image is the beginning of what is shown when the user hits predict.

Prediction of Air Quality Index (AQI)

This application has been designed using transfer learning to predict the AQI values for specific cities. New York City is the source city of the model and the weights of this city were transferred to similar cities of Chicago, IL and Philadelphia, PA to predict the AQI. Below the user will get the chance to choose a city, week, and year and find out the AQI for the inputs.

This application has been developed as part of Pennsylvania State University D5440 Capstone Project.

Please enter the city you would like to predict:

Chicago

Please enter the week of the year you would like to predict:

1


Please enter the year you would like to predict:

2018

If you would rather choose multiple weeks of information or a monthly average for a given year, please select one of the following, else keep blank

-

Predict



The predicted AQI: [8.080795]

The actual AQI: [9.62651297]

As stated previously, synthetic data was created for this process so that the user can predict for more years than just 2021. The target city models used this data as input and then outputted the predictions for the years 2018-2021. The predictions were combined with the sample data to then allow for easy usage of one dataset for displaying of real and predicted values. Below are the first ten rows of this dataset

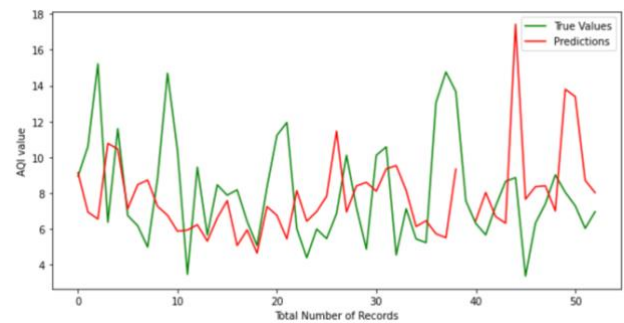
for Philadelphia. We can observe that predictions ('predictions') closely follow the real values ('AQI_weekly_avg').

	predictions	week	month	year	County	City	AQI_weekly_avg
1	8.886074	1	1	2018	Philadelphia	Philadelphia	9.60100502512563
2	8.661436	1	1	2019	Philadelphia	Philadelphia	6.76766826923077
3	8.938316	1	1	2020	Philadelphia	Philadelphia	9.12906976744186
4	15.2811165	1	1	2021	Philadelphia	Philadelphia	7.0555333998006
5	6.568886	2	1	2018	Philadelphia	Philadelphia	12.5432291666667
6	7.42284	2	1	2019	Philadelphia	Philadelphia	6.76248477466504
7	10.608303	2	1	2020	Philadelphia	Philadelphia	6.938
8	11.09194	2	1	2021	Philadelphia	Philadelphia	16.814328358209
9	7.0937376	3	1	2018	Philadelphia	Philadelphia	15.3846245530393
10	13.488091	3	1	2019	Philadelphia	Philadelphia	11.8868891537545

The dataset used (ie. Chicago or Philadelphia) is determined by user-input on the application and the specified week and year the user would like are subset out. The prediction and real values are then also returned. Depending on if the users would like to do a different function as described earlier, aggregations are run or different subsets are outputted.

IV. EVALUATION

Comparing our original models – Random Forest Regressor, SVM, and our custom neural network – we found the neural network to be the most efficient and produced the best results. From this, we optimized and enhanced the model which started with an RMSE of 1.13 and finished with an RMSE of 0.200. Once we seemed to no longer be able to improve the RMSE of the neural network, we used this model in creating a transfer learning model which will then be used on our target cities. The normalized RMSE found for the target cities was calculated. For Chicago, the model to be 0.247 and for Philadelphia it was found to be 0.200. Generally, a normalized RMSE found between 0.2-0.5 indicates the model is predicting relatively accurately. We also noted that COVID-19 may have been a factor with the models ran on the synthetic data used for the web application. The plot for 2020 of Philadelphia's AQI predicted values / real values is shown below. We see that our model predicted a significant spike around week 45; however, the true values are relatively low after week 40.



We believe our model's predictions are higher due to what the model was trained on and appear inaccurate because of the COVID-19 pandemic and its aid in improving air quality. Including more data of pre/post pandemic to train the model on may be beneficial in improving the overall accuracy in the future.

V. DISCUSSION

Overall, we were successful in predicting the AQI value for a given week and year for the cities of Philadelphia and Chicago from the transfer of the source city (NYC) model. We were not successful in being able to predict cities which do not have historic AQI data. As stated previously, we found due to Hazleton simply not having any substantial data and no AQI values transferring the model would have led to very poor, inaccurate results. A front-end implementation was also created to allow for user input and outputted predictions found from the model. The website also includes separate mock-up pages for each city which we hope to enhance with visualizations and statistics found from both datasets. The website was stress-tested manually by running through possible user-inputs. It was found that at times, the application would not predict if a user continually changed cities, weeks, and year multiple times. We were unsuccessful in trying to debug this issue, but we did find that simply refreshing the page resolved the issue. More stress tests will be run in the future if the application is deployed online to ensure the application works when multiple users are using it simultaneously. For more information pertaining to our future work of the project, please see section VII. Future Work.

VI. CONCLUSIONS

We were successful in our tasks and were able to create a neural network based on New York City data which predicts the air quality of a specific week and year and transfers it to the cities of Philadelphia and Chicago. We realized later the transfer of the model to Hazleton would not be feasible as there is no documented air quality index history for the city, causing the model to not be runnable due to the extremity of the data insufficiency and inability to verify any predictions. Future work in adjusting and optimizing the model to be able to handle these types of cities would be vital. We plan to also see if the model can still be enhanced and predict values even more accurately. Additionally, we successfully created a front-end dashboard which allows for user input of the city, week, and year and outputs the AQI measurement calculated from the model. Having this web application allows us to expand our audience to others who are interested in air quality information and would like an easy to use, comprehensive application. The website also completes the project and ties all the information together in one spot.

VII. FUTURE WORK

Open Source

One task we would like to complete in the future is to open-source our code and allow others to use it. Open sourcing it will

be beneficial for future scientists interested in this line of work and aiming to improve our modeling and ideas. It will be free to use, thus making it cost efficient.

Publish model to Python Package Index (PyPi)

We also plan to publish the source model to PyPi so that users looking for this type of neural network can implement it. PyPi is a python repository for python code and allows users to find specific packages based on keywords. This is another way of open sourcing the model for others to use. We originally planned to only publish the transfer model created on New York City because then researchers have the chance to transfer this model to other source cities and evaluate how well theirs works. After consulting with our client, we may also publish our models created on Chicago and Philadelphia after the transfer.

Include more functionality to front-end

Another option for future work is to extend the front-end to include more functionality. Possible additions include allowing the user to choose multiple months of a given year or solely a year and get the average AQI values for them. Another improvement would be included more cities and years for the users to pick from. Cosmetic adjustments to allow for easier user interaction and create a cleaner site are other ideas to enhance the website.

VIII. CONTRIBUTIONS

I was responsible for doing data collecting, data preprocessing, and early data analysis on the target cities. I also took the initiative in creating the Streamlit web application to showcase our project. Throughout the semester, I have taken on the role for team manager at points to make sure my team is also making progress with their parts.

I began by first researching and learning more on sustainability issues in cities, how transfer learning can be used to resolve these issues, and different projects that utilize similar ideas. After the team determined which cities would be interesting to investigate and seemed like the source city characteristically. I began collecting data for the source cities of Philadelphia, Hazleton, and Chicago. This data was found from multiple sources as stated previously. I then cleaned, aggregated – averaged AQI and other datasets if not weekly – and analyzed these sets – ran PCA to feature select, and determined if beneficial in using for the model. Distributions of AQI, weather, and pollution were also investigated to make sure these cities could be used as target cities of NYC. I also helped Divya with EDA on some NYC datasets and reported the results to her. After EDA was conducted on the datasets and variables were selected, I created a master dataset of the features and AQI values for each target city. The datasets were uploaded to our Azure dataset to then be able to be used with SQL and stored.

Modeling was conducted once all datasets were complete. During this time, I began learning more about Tableau for deploying the model. I later realized due to time constraints and minimal documents pertaining to deploying models on the

software, it would be not feasible for this project. I then began researching and working with Streamlit as I had not been familiar with it before this project. I created the dashboard from scratch which uses data calculated from the model and outputs the AQI predictions based on the user's input. I trained and tested and model in python on synthetic data, which then predicted the AQI values for each year. I also had plots created using the function we used throughout the model for each city and year. I then concatenated the predictions and sample data used and uploaded the dataset to my python Streamlit script. Additionally, I created functions which give an average monthly AQI value for a user-inputted city and year and one which allows the user to select multiple weeks of data and display the found results. We also planned to include statistics and visualizations found from the data on target cities to the site, however that was not completed in time. I did however include a multipage layout to the application in case they were to be created. The multi-page coding was adapted from a source code on GitHub and reproduced for this project. Adaptions made to the 'app.py' [1] file, including changing imports, name of title page, and add page selections. The 'multipage.py' [2] file was replicated for the project. Coding portions can be found at the links below:

- Data preprocessing:
<https://github.com/vbgupta/DS440-Transfer-Learning-Address-Sustainability-Issues/tree/main/prep>
 - o This includes folders ('Chicago', 'Hazelton', 'Philly' of each target city with R-markdowns of creating the master datasets along with aggregations and some data used (not all due to memory restrictions). Dataset created from predictions on from synthetic data for website also included.
 - o EDA folder contains reports of datasets which were inspected for feature selection and overall quality of the data, R script used in checking source and target cities in 'Distribution' folder. R markdowns created for target cities' AQI, pollution, WQI also in folder.
 - o AQIdata.R is an R script used for combining multiple years of data in JSON format retrieved from the EPA's API into a single CSV file.
- Web-Application:
<https://github.com/vbgupta/DS440-Transfer-Learning-Address-Sustainability-Issues/tree/main/pages>
 - o Chi_info.py, philly_info.py mock-up pages which can be added to in the future with visualization and statistics
 - o Dashboard.py code used in creating the functions throughout the website along with

app function used in creating what is displayed on the site

- o Pred_actual.R, Rscript used in combining predictions found from model for years 2018-2021 with master datasets for target city.
- o Streamlit-app-....webm is a video of me playing with the website

REFERENCES

- [1] Rathi, P. (2022) app.py [Source code]
<https://gist.github.com/prakharrathi25/74925f95291f002a6d999824c072db53>
- [2] Rathi, P. (2022) multipage.py [Source code]
<https://gist.github.com/prakharrathi25/ee6fb3fb3d02a282179a2c0d42293063>
- [3] Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., & Liu, Z. (2021, July 29). *A transfer learning-based LSTM strategy for imputing large-scale consecutive missing data and its application in a water quality prediction system*. Journal of Hydrology. Retrieved February 11, 2022, from
https://www.sciencedirect.com/science/article/pii/S0216942100620X?casa_token=b1A280ongAAAAA%3ASPU2HSz8KdcAU6iZ7ZtxB20ngT6h_Qdop0XBuxg07Py5pDCWk_oR2fWzY1w72nkTckgW2i5qx_A
- [4] Somu, N., Sriram, A., Kowli, A., & Ramamritham, K. (2021, July 10). *A hybrid deep transfer learning strategy for thermal comfort prediction in buildings*. Building and Environment. Retrieved February 11, 2022, from
https://www.sciencedirect.com/science/article/pii/S0360132321005345?casa_token=ej7Fl_8Wc_QAAAAA%3AOKrLhyFJVNjLsahswzvFGYbg0z7ZKgA--gdjzvbyx13XipQbkWcHKAd9CYNffmLbhi2_IMUxJ4g

APPENDIX

- [A.1] Gupta, V., Racho, A., Rustagi, D. (2022) *DS440-Transfer-Learning-Address-Sustainability-Issues* [Source code]
<https://github.com/vbgupta/DS440-Transfer-Learning-Address-Sustainability-Issues>
- [A.2] Gupta, V., Racho, A., Rustagi, D. (2022) *DS440 – Multiple Models* [Source code]
<https://drive.google.com/file/d/1h8nm-zm1ZPU-yheLi1AF-xhxmeTN0Ybz/view?usp=sharing>