PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Predicción temporal de calidad del aire en Lima a partir de datos de estaciones de bajo costo y Aprendizaje Automático: una revisión de literatura

TRABAJO DE INVESTIGACIÓN PARA LA OBTENCIÓN DEL GRADO DE BACHILLER EN CIENCIAS CON MENCIÓN EN INGENIERÍA INFORMÁTICA

AUTOR

Diego Jose Paredes Salazar

ASESOR:

Edwin Rafael Villanueva Talavera

Lima, enero, 2021

Resumen

El presente trabajo explora los estudios en los cuales se utilizan técnicas de aprendizaje profundo para realizar predicción temporal de calidad del aire, de manera que se pueda comprender que características tendrían los modelos de aprendizaje profundo que tienen un mejor rendimiento con para realizar esta tarea y puedan utilizarse como línea base para desarrollar modelos similares en el contexto de la ciudad de lima. Esta revisión de literatura se realiza con el objetivo de poder obtener los modelos de aprendizaje profundo que estén teniendo un mejor rendimiento en la actualidad al predecir temporalmente la calidad del aire mediante un procedimiento que garantice objetividad y reproducción de resultados. Para ello, se realiza una revisión sistemática de literatura que garantiza el uso de procedimientos estructurados y definidos para conocer las preguntas de investigación que guían la exploración de los estudios de predicción temporal de calidad del aire, los motores de búsqueda considerados para la revisión y las cadenas de búsqueda asociadas tanto a las preguntas de investigación como los motores de búsqueda, de manera que estas se puedan ejecutar y reproducir la obtención de estudios. Las respuestas se reportan en un formulario de extracción con datos relacionados a las arquitecturas de aprendizaje profundo, limitaciones de los modelos empleados y el rendimiento obtenido por cada modelo en cada estudio. Al finalizar el estudio, se concluye que se puede desarrollar un modelo basado en una arquitectura adecuada de aprendizaje profundo para poder atacar el problema de la predicción inadecuada de calidad del aire en Lima al percatar su efectividad reportada en la literatura para otras localidades en el mundo, considerando que dichos modelos deben tomarse únicamente como una línea base y que deben ajustarse a la localidad de Lima para obtener predicciones adecuadas a su entorno.

Tabla de Contenido

Índice de Tablas

1 Introducción

En ciencias atmosféricas, uno de los temas que ha recibido más atención en las últimas décadas es el de calidad del aire, el cual está referido a la medida de contaminación de este en un espacio geográfico determinado (Godish, 2003). Debido a la estrecha relación que existe entre la contaminación del aire con la adquisición de enfermedades respiratorias como el asma es que ha sido de interés de la comunidad científica desarrollar modelos predictivos que permitan determinar la calidad del aire en el futuro, pues esto permitiría planificar actividades diarias previendo el riesgo de interactuar con aire contaminado que perjudique la salud de la población en una zona urbana (Baumann et al., 2012).

En el caso particular de la ciudad de Lima, la capital de Perú, el problema de la calidad del aire es crítico: es la segunda ciudad más contaminada de América Latina según un estudio de la organización AQAir (AQAir, 2018), la cual se dedica a realizar estudios de calidad del aire en todo el mundo. Según el estudio, Lima tiene 28 μg/m³ de concentración promedio anual de PM2.5 (AQAir, 2018), lo cual sobrepasa el límite establecido por la Organización Mundial de la Salud: 10 μg/m³(World Health Organization, 2005). A pesar del conocimiento de esta situación, aún no se cuenta con un sistema de predicción de calidad del aire que permita a los pobladores de la ciudad planificar sus actividades con el conocimiento del nivel de contaminación del aire que presenta la zona urbana donde realizan sus actividades(Reátegui-romero et al., 2018). Si bien ya ha habido intentos por parte de instituciones estatales y académicas de atender este problema mediante modelos numéricos físico-químicos, ninguno ha podido ser desplegado con éxito debido a que estaban basados en datos de inventarios de emisiones imprecisos y estaciones de calidad del aire limitadas (Reátegui-romero et al., 2018).

Actualmente, la Pontificia Universidad Católica del Perú, en colaboración con la empresa Qaira y la Universidad Católica San Pablo, ha logrado desarrollar una red de estaciones de calidad del aire de bajo costo, la cual se encuentra operativa en la ciudad de Lima (Cobarrubias, 2020). Las estaciones de calidad del aire son dispositivos que poseen herramientas para realizar mediciones automáticas de las concentraciones de diversos contaminantes del aire en la zona donde se ubica (Environmental Protection Agency, 2020). En consecuencia, el problema de disponibilidad de datos que limitó los intentos anteriormente mencionados ha sido solucionado, por lo cual se podrían aprovechar para construir modelos de predicción temporal mediante técnicas de ciencias de la computación.

El presente trabajo de investigación permitirá comprender el proceso de revisión de literatura ejecutado con la finalidad de averiguar sobre estudios de predicción de calidad del aire mediante el uso de aprendizaje profundo. Se describirá el objetivo de la revisión realizada en función del problema de predicción de calidad del aire que se estudiará en este proyecto de fin de carrera, el cual permite plantear las preguntas de revisión que se apreciarán en esta sección. A continuación, se describe la Estrategia de búsqueda de estudios primarios en función de los motores de búsqueda seleccionados, las cadenas de búsqueda diseñadas para cada motor, las características de los documentos encontrados y los criterios de inclusión y exclusión, los cuales guardarán relación con el ámbito de predicción de calidad del aire. Se continuará con la presentación de los formularios de extracción utilizados para recopilar información de los documentos encontrados y se presentarán las respuestas

a las preguntas de revisión a partir del contenido de los formularios de extracción. Finalmente se concluirá respecto a cómo los resultados obtenidos a partir de la revisión fundamentan el desarrollo de este proyecto de fin de carrera mediante su relación con la problemática descrita en este documento. De esta manera, las respuestas a las preguntas de la revisión permitirán comprender cómo se puede resolver el problema de calidad del aire mediante soluciones similares en otros estudios actuales de la comunidad académica mediante aplicación de aprendizaje profundo.

2 Método

Se realizará una revisión sistemática que tendrá como objeto de revisión las arquitecturas de aprendizaje profundo y su aplicación en modelos predictivos en el ámbito de la calidad del aire. Como lo que interesa conocer es la manera en la cual se emplean dichas arquitecturas para resolver el problema de predicción de calidad del aire en el ámbito científico, se realizará una revisión empírica de los estudios primarios relacionados a esta temática. La presente revisión tiene como objetivo principal identificar cuáles son las arquitecturas de aprendizaje profundo para entrenar modelos de predicción de calidad del aire cuya función es predecir la contaminación de este en los últimos 6 años. Para ello, se realizará una revisión de los métodos y resultados obtenidos en los estudios de predicción de calidad del aire mediante el uso de aprendizaje profundo.

2.1 Revisión Sistemática

2.1.1 Preguntas de Investigación

- P1. ¿Cuáles son las características de las arquitecturas de aprendizaje profundo utilizadas para realizar modelos predictivos de la calidad del aire actualmente?
- P2. ¿Qué limitantes se han encontrado en los estudios que emplean arquitecturas de aprendizaje profundo para realizar modelos predictivos de la calidad del aire actualmente?
- P3 ¿Cuál es el grado de exactitud que han mostrado los modelos recientes de predicción de calidad del aire basados en aprendizaje profundo?

2.1.2 Proceso de Búsqueda

Para agilizar el proceso de búsqueda, se emplearán bases de datos de estudios científicos, los cuales serán extraídos a partir de una misma cadena de búsqueda adaptada a la sintaxis de cada motor de búsqueda de las bases de datos mencionadas. Para garantizar la calidad mínima de los estudios recopilados se emplearán 2 bases de datos altamente utilizadas por la comunidad académica: “Web of Science” y “Scopus”

Debido a la naturaleza del proyecto, las respuestas a las preguntas planteadas deben corresponder al mismo estudio realizado. Como lo que se busca es comprender cómo se están implementado las soluciones de aprendizaje profundo para realizar la tarea de predecir la calidad del aire, las respuestas de a la pregunta no pueden ser simples hallazgos encontrados en fuentes distintas, sino que deben complementar sus respuestas mediante el mismo estudio. Por ejemplo: Si un estudio muestra las características de una arquitectura de aprendizaje profundo para predecir la calidad del aire, pero no

presenta el grado de exactitud de dicha arquitectura, entonces no se tendrá experiencia empírica para sustentar la validez de dicha arquitectura. De igual manera, un estudio puede mostrar el grado de precisión de una arquitectura, pero si no la describe al detalle no se podrá replicar el experimento realizado. Finalmente, las limitaciones de un estudio, como se busca responder en la segunda pregunta, no son útiles si no se conoce a detalle la arquitectura empleada y los resultados de su empleo. Por eso es necesario comprender la arquitectura empleada, las limitaciones de la arquitectura y del estudio realizado y los resultados que permitan comprender la validez empírica de dicha arquitectura en el contexto de sus limitaciones. Es por esta razón que en el proceso de búsqueda se utilizará una sola cadena de búsqueda que permita recolectar artículos relacionados al uso de modelos de predicción de calidad del aire basados en arquitecturas de aprendizaje profundo en vez de una cadena por cada pregunta de investigación. Las palabras claves a considerar en esta revisión son las siguientes: “Deep learning”, “Neural Networks”, “prediction”, “forecasting”, “Air quality”, “Air pollution”. Las cadenas de búsqueda a utilizar son las siguientes:

Scopus:

- TITLE ("air pollution" OR "air quality") AND TITLE ("prediction" OR "forecasting") AND TITLE-ABS-KEY ("deep learning" OR "neural network") AND TITLE-ABS-KEY ("architecture" OR "model") AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015)) AND (LIMIT-TO (LANGUAGE , "English"))

Web of Science:

- ((TI=("air pollution" OR "air quality") AND TI=("prediction" OR "forecasting") AND TS=("deep learning" OR "neural networks")) AND PY=(2015 OR 2016 OR 2017 OR 2018 OR 2019 OR 2020)) AND LANGUAGE: (English)

Se revisarán los estudios que se extraigan a partir de las cadenas mencionadas y los datos a analizar, que se verán con detalle en la sección “Datos Analizados”, serán reportados en un formulario de extracción que se incluirá junto con el presente documento.

2.1.3 Criterios de Inclusión y Exclusión

Para la presente revisión se incluyeron los estudios que cumplieron con los siguientes criterios de inclusión:

1. El estudio emplea arquitecturas de aprendizaje profundo para realizar modelos de predicción del nivel de calidad del aire en función de la contaminación generada por sustancias contaminantes, pues es la subárea de ciencias de la computación que se desea desarrollar en la presente tesis.
2. El estudio explica el proceso empleado para la elección de la arquitectura y el modelo utilizado, pues así podrá realizar un proceso similar para el proyecto a desarrollar.

3. El estudio muestra al menos una métrica que permita evaluar el grado de exactitud del modelo realizado, pues es necesario que lo propuesto tenga prueba empírica de su grado de exactitud para replicar dicho modelo en el proyecto.
4. El estudio se encuentra escrito en inglés, pues es un idioma más empleado por los autores de las publicaciones científicas.
5. El estudio debe estar realizado en los últimos 6 años, pues al ser una revisión de estado del arte es necesario que los estudios sean recientes.

Para la presente revisión se excluyeron los estudios que cumplieron con los siguientes criterios de exclusión:

1. El estudio emplea métodos numéricos para realizar la predicción del nivel de calidad del aire, pues a pesar de ser otra técnica utilizada para este tipo de tareas, es más afín a la física y estadística que a las ciencias de la computación, por lo cual no es de interés en una tesis de ingeniería informática.
2. El estudio emplea un modelo de aprendizaje profundo únicamente para comparar resultados con el modelo principal, el cual se ha desarrollado con otras técnicas, pues es necesario que se explique el proceso de selección y desarrollo del modelo para poder utilizarlo como base para el proyecto.
3. El estudio está relacionado a la predicción de la calidad del aire, pero el modelo desarrollado predice variables no relacionadas con la concentración de contaminantes en la intemperie, como la tasa de mortalidad asociada al aire contaminado, pues esas predicciones no están contempladas en el alcance del proyecto.
4. El estudio reporta una investigación que ha sido mejorada en otro estudio escrito por los mismos autores, pues ese estudio tendría la misma información y no agregaría información adicional.

2.1.4 Datos Extraídos

Qué datos extraerá de cada artículo, por ejemplo: resumen, área, fuente del artículo, entre otros.

De cada artículo obtenido, se extraerán los siguientes datos:

- Fuente del artículo
- Resumen

2.1.5 Datos Analizados

Datos que recolectará para poder responder a las preguntas de investigación. Por ejemplo: años o países donde se ha encontrado el mismo problema o soluciones.

Datos que analizar para la pregunta 1:

- Tipo de arquitectura
- Número de neuronas en capa de entrada
- Número de capas encondidas
- Número de neuronas en capa de salida

- Temporalidad de predicción
- Implementación pública

Datos que analizar para la pregunta 2:

- Limitaciones en los datos de entrada
- Limitaciones en el entrenamiento del modelo
- Limitaciones en los resultados del modelo
- Limitaciones en el entorno del modelo

Datos que analizar para la pregunta 3:

- Métrica utilizada para medir el rendimiento
- Valor de la métrica
- Se considera aceptable para predicción

2.2 Resultados

2.2.1 Resultados de Búsqueda

La lista de estudios primarios recopilados para esta revisión de literatura se encuentra en el archivo "20151107_DiegoParedes_EdwinVIllanueva_EstudiosPrimarios.pdf" incluido junto a este documento. Los resultados de la búsqueda se encuentran en el archivo "20151107_DiegoParedes_EdwinVillanueva_FormularioExtracción.xlsx", donde cada celda corresponde a los datos a analizar especificados en la sección anterior para cada estudio recopilado por medio de las búsquedas realizadas en las bases de datos especificadas en la sección "Procesos de Búsqueda" del presente documento, constituyendo de esta manera el detalle de los Resultados de Búsqueda.

Tabla 1. Estudios primarios obtenidos en la revisión sistemática

Autores	Título del estudio	Año de publicación
Abdullah, S., Ismail, M., Ahmed, A., & Abdullah, A.	Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support.	2019
Abdullah, S., Ismail, M., Ahmed, A., & Mansor, W.	Big data analytics and artificial intelligence in air pollution studies for the prediction of particulate matter concentration.	2019
Ao, D., Cui, Z., & Gu, D.	Hybrid model of air quality prediction using k-means clustering and deep neural network.	2019
Athira, V., Geetha, P., Vinayakumar, R., & Soman, K.	DeepAirNet: Applying Recurrent Networks for Air Quality Prediction.	2018

Dedovic, M., Avdakovic, S., Turkovic, I., Dautbasic, N., & Konjic, T.	Forecasting PM10 concentrations using neural networks and system for improving air quality.	2016
Freeman, B., Taylor, G., Gharabaghi, B., & Thé, J.	Forecasting air quality time series using deep learning.	2018
Ge, L., Zhou, A., Li, H., & Liu, J.	Spatially Fine-Grained air quality prediction based on DBU-LSTM.	2019
Honarvar, A., & Sami, A.	Towards Sustainable Smart City by Particulate Matter Prediction Using Urban Big Data, Excluding Expensive Air Pollution Infrastructures.	2019
Huang, M., Zhang, T., Wang, J., & Zhu, L.	A new air quality forecasting model using data mining and artificial neural network.	2015
Jin, X., Yang, N., Wang, X., Bai, Y., Su, T., & Kong, J.	Deep hybrid model based on EMD with classification by frequency characteristics for long-term air quality prediction.	2020
Jo, B., & Khan, R.	An internet of things system for underground mine air quality pollutant prediction based on azure machine learning.	2018
Kang, Z., & Qu, Z.	Application of BP neural network optimized by genetic simulated annealing algorithm to prediction of air quality index in Lanzhou.	2017
Kaya, K., & Gündüz Öğüdücü, Ş.	Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting.	2020
Keerthana, R., & Chooralil, V.	Forecasting of the air pollution based on meteorological data and air pollutants using deep learning: A novel review.	2020
Kök, I., Şimşek, M., & Özdemir, S.	A deep learning model for air quality prediction in smart cities.	2017
Korunoski, M., Stojkoska, B., & Trivodaliev, K.	Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization.	2019

Li, J., Shao, X., & Sun, R.	A DBN-based deep neural network model with multitask learning for online air quality prediction.	2019
Li, Y., Shen, X., Han, D., Sun, J., & Shen, Y.	Spatio-temporal-Aware Sparse Denoising Autoencoder Neural Network for Air Quality Prediction.	2019
Lim, Y., Aliyu, I., & Lim, C.	Air pollution matter prediction using recurrent neural networks with sequential data.	2019
Lin, C.-Y., Chang, Y.-S., Chiao, H.-T., & Abimannan, S.	Design a hybrid framework for air pollution forecasting.	2019
Lin, Y., Mago, N., Gao, Y., Li, Y., Chiang, Y.-Y., Shahabi, C., & Ambite, J.	Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning.	2018
Liu, T., Ying, Y., Xu, Y., Ke, D., & Su, K.	Fine-Grained Air Quality Prediction using Attention Based Neural Network.	2018
Luo, Z., Huang, J., Hu, K., Li, X., & Zhang, P.	Accuair: Winning solution to air quality prediction for KDD Cup 2018.	2019
Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G.	A machine learning model for air quality prediction for smart cities.	2019
Oprea, M., Mihalache, S., & Popescu, M.	A comparative study of computational intelligence techniques applied to PM2.5 air pollution forecasting.	2016
Patni, J., & Sharma, H.	Air Quality Prediction using Artificial Neural Networks.	2019
Pawul, M., & Śliwka, M.	Application of artificial neural networks for prediction of air pollution levels in environmental monitoring.	2016
Qi, Z., Wang, T., Song, G., Hu, W., Li, X., & Zhang, Z.	Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality.	2018

Septiawan, W., & Endah, S.	Suitable Recurrent Neural Network for Air Quality Prediction with Backpropagation Through Time.	2019
Sinnott, R., & Guan, Z.	Prediction of Air Pollution through Machine Learning Approaches on the Cloud.	2019
Siwek, K., & Osowski, S.	Data mining methods for prediction of air pollution.	2016
Soh, P.-W., Chang, J.-W., & Huang, J.-W.	Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations.	2018
Srikamdee, S., & Onpans, J.	Forecasting Daily Air Quality in Northern Thailand Using Machine Learning Techniques.	2019
Sun, X., & Xu, W.	Deep Random Subspace Learning: A Spatial-Temporal Modeling Approach for Air Quality Prediction.	2019
Tamas, W., Notton, G., Paoli, C., Nivet, M., & Voyant, C.	Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks.	2016
Tao, Q., Liu, F., Li, Y., & Sidorov, D.	Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU.	2019
Udaya Bharathi, R., & Seshashayee, M.	Intelligent air pollution prediction system using internet of things (Iot).	2019
Wang, B., Kong, W., Guan, H., & Xiong, N.	Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things.	2019
Wu, Z., Wang, Y., & Zhang, L.	MSSTN: Multi-Scale Spatial Temporal Network for Air Pollution Prediction.	2019

Xayasouk, T., Lee, H., & Lee, G.	Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models.	2020
Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y.	Deep distributed fusion network for air quality prediction.	2018
Zhao, C., Van Heeswijk, M., & Karhunen, J.	Air quality forecasting using neural networks.	2017
Zhao, C., Van Heeswijk, M., Karhunen, J., Kök, I., Şimşek, M., Özdemir, S., . . . Karhunen, J.	Suitable Recurrent Neural Network for Air Quality Prediction with Backpropagation Through Time.	2019
Zhao, G., Huang, G., He, H., He, H., & Ren, J.	Regional Spatiotemporal Collaborative Prediction Model for Air Quality.	2019
Zheng, Y., & Ch, E.	Casting Fine-G Grained Air Quality Based on Big D.	2015
Zhenghua, W., & Zhihui, T.	Prediction of air quality index based on improved neural network.	2018

2.3 Discusión

2.3.1 Respuesta a pregunta P1: ¿Cuáles son las características de las arquitecturas de aprendizaje profundo utilizadas para realizar modelos predictivos de la calidad del aire actualmente?

De la revisión que se ha realizado, se han encontrado 29 arquitecturas distintas. La más utilizada ha sido la arquitectura LSTM (Long Short Term Memory), empleada en 14 de los 40 estudios revisados hasta el momento. Dicha arquitectura es considerada pertinente por la comunidad científica debido a su habilidad para considerar la influencia de datos tomados en temporalidades anteriores en las próximas (Freeman et al., 2018). Otra arquitectura que tiene presencia considerable en los estudios revisados es la ANN clásica con 9 estudios. Sin embargo, 5 de esos estudios fueron publicados entre el 2015 y 2016 y 4 fueron publicados en entre el 2017 y el 2019, mientras que solo uno de los que emplean LSTM fue publicado en 2017 y el resto entre 2018 y 2020. Esto mostraría que la tendencia actual es el uso de arquitecturas orientadas a series temporales como LSTM. Otra arquitectura que se ha encontrado en estudios recientes es la GRU (Gated recurrent unit) con 4 estudios entre el 2018 y el 2020.

En cuanto a la constitución de las capas, Se han encontrado 15 estudios que han empleado capas escondidas híbridas, lo cual quiere decir que emplean una combinación de arquitecturas con el fin de lograr un mejor grado de exactitud (Tao et al., 2019). Por ejemplo, la arquitectura CBGRU cuenta con 5 capas escondidas: 3 1D-CNN y 2 Bi-GRU. Las otras arquitecturas híbridas encontradas fueron DFS (Kaya & Gündüz Öğüdücü, 2020) y ST-DNN(Soh et al., 2018). De los 40 estudios revisados,15 utilizan arquitecturas de una sola capa y de esos 15 estudios5 corresponden a la arquitectura LSTM, 3 al clásico ANN, 3 utilizan MLP y los otros 4 corresponden a arquitecturas híbridas, como RBF, DCRNN, ELM y Jordan RNN.

Respecto a las capas de salida de las arquitecturas en los estudios encontrados, 28 de los 40 estudios emplean una sola neurona en la salida. De estos 28 estudios, 9 se emplean para predecir PM2.5 de la siguiente unidad temporal, ya sea horas o días. Existen 12 estudios en los cuales se utilizan múltiples neuronas en la capa de salida, de los cuales 5 se utilizan para predecir el valor de la siguiente temporalidad en varias estaciones o con varias variables a predecir, como concentración de PM2.5 y PM10. Los otros 7 permiten predecir el valor de una variable, como concentración de PM2.5, pero en varias temporalidades.

Finalmente, en el caso de las capas de entradas se denotó que el número de neuronas utilizadas está más condicionado por el contexto del estudio realizado que las otras configuraciones de arquitectura. Por ejemplo, un estudio tenía 40 neuronas debido a las 5 muestras realizadas en el estudio, 2 unidades de temporalidad, en este caso horas, por muestra y en cada unidad de temporalidad se miden 4 variables (Freeman et al., 2018). En otros estudios donde se utilizan datos más cualitativos que cuantitativos, se emplean múltiples neuronas para representar un dato en particular, como la dirección del aire (Kaya & Gündüz Öğüdücü, 2020). Hasta el momento, no se han encontrado estudios que posean una implementación pública de su arquitectura.

2.3.2 Respuesta a pregunta P2: ¿Qué limitantes se han encontrado en los estudios que emplean arquitecturas de aprendizaje profundo para realizar modelos predictivos de la calidad del aire actualmente?

De la revisión que se ha realizado, 21 No especifican alguna limitación en lo que refiere a la estructura de los datos de entrada, 16 no especifican alguna limitación correspondiente al entrenamiento que realiza el modelo de aprendizaje profundo, 26 no lo hacen respecto a los resultados de los datos predichos y 16 en lo que respecta a limitaciones del entorno, como la fuente de los datos trabajados.

De los 40 estudios, 5 solo consideran datos relacionados a condiciones meteorológicas, como la rapidez del aire, el nivel de precipitación, la dirección del aire, entre otros (Xayasouk et al., 2020). Por otro lado, 6 estudios solo consideran la concentración de contaminantes, como el CO2, PM2.5, PM10, SO2, gases NOx, O3, entre otros (Septiawan & Endah, 2019). Cuatro estudios aclaran que han tenido datos incompletos en sus fuentes de datos y han intentado arreglarlos con técnicas de limpieza de datos (Abdullah et al., 2019).

En lo que respecta a limitaciones en el entrenamiento de los modelos, la limitación que más se ha presentado es el tiempo prolongado de entrenamiento que presentan las arquitecturas híbridas y del tipo LSTM en comparación con las redes neuronales clásicas, teniendo así 20 de 40 estudios que

afirman tener esta limitación. En particular, Septiawan afirma debido a que estas arquitecturas deben emplear el método “Back propagation through time” para su entrenamiento, el cual es más costoso computacionalmente respecto al “Back propagation” clásico (Septiawan & Endah, 2019). Existen 2 estudios que utilizaron algoritmos genéticos con el fin de maximizar la exactitud de los valores predichos por las redes neuronales. Sin embargo, la gran limitación en estos casos es que los algoritmos genéticos tienen un alto costo de procesamiento, por lo cual se necesitan equipos con mayor poder de procesamiento que en un estudio de aprendizaje profundo simple (Zhenghua & Zhihui, 2018).

En cuanto a los resultados del modelo, 9 estudios han decidió reducir la predicción a un dato que resume las características de los contaminantes, como el AQI o el MEI. Esto se realiza para disminuir la complejidad del modelo (Jo & Khan, 2018). Sin embargo, como son datos resumidos se pierde la oportunidad de representar de una mejor manera las condiciones de calidad del aire en una zona (Keerthana & Chooralil, 2020). Adicionalmente existen 4 estudios que al haber utilizado una arquitectura de redes neuronales simple, han ignorado la dependencia entre contaminantes o estaciones continuas, lo cual introduce error a la hora de realizar futuras predicciones (Sun & Xu, 2019).

En cuanto al entorno del modelo, en 6 estudios se utilizaron datos provenientes de repositorios web, lo cual puede evitar analizar los datos que existen en la actualidad, situación que podría evitar considerar los cambios climáticos que han ocurrido en una determinada zona (Ao et al., 2019). Otro caso que es importante mencionar es en el que se utilizan datos provenientes de diversas localidades lejanas, como ciudades de diferentes países(Lin et al., 2018). Se han encontrado 2 estudios en los cuales se ha dado esta situación para cubrir la insuficiencia de datos. Si bien de esta manera se garantizan más datos de entrenamiento, las diferencias meteorológicas de las regiones, pues ya no son cercanas en el caso de países distintos, y la diferencia de fechas puede inducir sesgo a la hora de realizar la predicción (Oprea et al., 2017).

2.3.3 Respuesta a pregunta P3: ¿Cuál es el grado de exactitud que han mostrado los modelos recientes de predicción de calidad del aire basados en aprendizaje profundo?

De la revisión que se ha realizado, se han encontrado 11 métricas distintas para medir el grado de exactitud de los modelos de predicción de calidad del aire basados en aprendizaje profundo. De los 40 estudios, 18 emplean la métrica RMSE para medir el grado de exactitud del modelo desarrollado. En 16 estudios se emplea MAE para realizar la misma tarea. Es común que en los estudios se emplee más de una métrica con el fin de poder validar de mejor manera los resultados observados, lo cual ocurre en 13 de los 40 estudios, teniendo a la combinación MAE, RMSE como la más empleada con 11 de los 13 estudios mencionados. Otras métricas empleadas incluyen la razón o porcentaje de exactitud con 5 estudios, R2 con 3 estudios, SMAPE con 2 estudios, precisión con un estudio y MAPE con 4 Estudios. En 5 estudios esta métrica no está especificada, pues se limitan a mostrar gráficos en los cuales se compara el resultado predicho con el resultado real (Zhao et al., 2017).

Los valores de las métricas son variados y se emplean para sustentar que su modelo es mejor comparado a los otros métodos propuestos en el momento en el cual se realizó el estudio(Lin et al., 2018). De los estudios revisados, 11 tenían la métrica definida pero no fue presentada de manera cuantitativa sino cualitativa, ayudándose de un gráfico que ayude a comparar el valor de dicha métrica en el modelo desarrollado con los valores en los otros modelos. Por ejemplo, en el estudio de Wang se compara en diagramas de barras los valores de los modelos LSTM y LSTM&GRU (Wang et al., 2019). Estos casos se denotan en el formulario de extracción al tener el campo de métrica utilizada completado y el campo valor de la métrica en blanco. Para el caso de RMSE, el resultado más bajo encontrado fue en el estudio de Qi con 0.0877 (Qi et al., 2018). Si bien es cierto que no se deberían comparar directamente los resultados, pues los factores de estudio pueden variar, el resultado es importante para tener una referencia a que grado de exactitud se ha podido llegar bajo una combinación de factores determinada (Freeman et al., 2018).

En el caso de la aceptabilidad de la predicción, se han encontrado 5 estudios donde el estudio no se ha encontrado con un resultado aceptable. Tres de estos casos son por no tener otros modelos para ver si se ha realizado alguna mejora, en uno el autor menciona que no encontró un modelo adecuado para entrenarlo (Wang et al., 2019) y en otro su modelo de aprendizaje profundo resultó inferior a uno de aprendizaje de máquina clásico (Mahalingam et al., 2019). El resto de los estudios ha tenido un estudio cuyas métricas han superado a los estudios pasados, por lo cual se consideran aceptables para la predicción. Esto lo lograron debido a que adaptaron mejor el modelo a las condiciones climatológicas que exigía el ambiente estudiado (Kaya & Gündüz Öğüdücü, 2020).

3 Conclusiones

En la introducción se habló de cómo el problema de la predicción de calidad del aire en Lima no podía realizarse en el tiempo por falta de datos hasta el momento y que ello era posible en este momento debido a la presencia de nuevas estaciones de calidad que suplían esa precondición. Como se aprecia en los resultados de la respuesta 2, este es un problema que se ha enfrentado en varias partes del mundo y que, si bien los diseños de modelos predictivos pueden ayudar a solventar estos problemas, constituye una fuerte limitación el hecho de no tener datos de las regiones en las cuales se va a realizar la predicción, ya sea de manera temporal o espacial (Oprea et al., 2017). Para el caso de Lima no se tendrán las limitaciones de falta de datos, como se exploró en la introducción, por lo cual se podrá desarrollar la solución del modelo predictivo adaptado a Lima de manera adecuada. Además, se reafirma, tal como lo indica la pregunta 2, que trabajar con datos de otras localidades es una seria limitación por el hecho de que las predicciones de calidad del aire son dependientes de las condiciones ambientales de las zonas a predecir (Lin et al., 2018). Por esta razón, no bastaría simplemente con implementar una de las soluciones existentes en los estudios científicos, sino que hay que realizar el modelo acorde con el entorno donde va a operar. En lima aún no existe un modelo predictivo de calidad del aire de estas características (Reátegui-romero et al., 2018), por lo cual es necesario desarrollar un modelo adaptado a Lima, lo cual se propone realizar a partir de arquitecturas de aprendizaje profundo con los mejores rendimientos encontrados en la presente revisión.

Como se observa en la pregunta 1, el desarrollo de una correcta arquitectura de modelos predictivos, en el caso de la revisión orientados a aprendizaje profundo y no los modelos fisicoquímicos que no pudieron implementarse adecuadamente en Lima (Sánchez-ccoyllo et al., 2018), permite predecir valores de contaminantes, así como medidas resumen de estos. Con ello se observa que efectivamente se puede aprovechar la capacidad predictiva de los modelos para obtener indicadores de los contaminantes a través del tiempo y el espacio. Como se observa además en la respuesta a la pregunta 3, se tiene que el diseño personalizado de las arquitecturas permite obtener mejores resultados de exactitud, lo cual demuestra que el hecho de no contar con modelos de predicción adaptados a la realidad geográfica no permite realizar predicciones exitosas, por lo cual se debe implementar un modelo de predicción personalizado para lima como se desea realizar en base a la información recopilada en este trabajo. Con dicho modelo desarrollado se atendería no solo el problema de predicción de calidad del aire mediante herramientas informáticas, como se realiza en los estudios presentados, sino que se aprovecharán los valores de predicción para presentar los indicadores correspondientes a la concentración de los contaminantes predichos a la comunidad que habita en la zona de lima, así como la comparación de las predicciones con las mediciones establecidas en los estándares de calidad del aire (El Peruano, 2017) para que sea más sencilla su fiscalización.

4 Referencias

Abdullah, S., Ismail, M., Ahmed, A. N., & Abdullah, A. M. (2019). Forecasting particulate matter concentration using linear and non-linear approaches for air quality decision support. *Atmosphere*, *10*(11). https://doi.org/10.3390/atmos10110667

Ao, D., Cui, Z., & Gu, D. (2019). Hybrid model of air quality prediction using k-means clustering and deep neural network. *Chinese Control Conference, CCC*, *2019-July*, 8416–8421. https://doi.org/10.23919/ChiCC.2019.8865861

Baumann, L. M., Robinson, C. L., Combe, J. M., Romero, K., Gilman, R. H., Cabrera, L., & Nadia, N. (2012). *Effects of distance from a heavily transited avenue on asthma and atopy in a peri-urban shanty-town in Lima, Peru*. *127*(4), 875–882. https://doi.org/10.1016/j.jaci.2010.11.031.Effects

El Peruano. (2017). *Aprueban Estándares de Calidad Ambiental (ECA) para Aire y establecen Disposiciones Complementarias*. 6–9.

Freeman, B. S., Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. *Journal of the Air and Waste Management Association*, *68*(8), 866–886. https://doi.org/10.1080/10962247.2018.1459956

Godish, T. (2003). *Air Quality* (4th Editio). Lewis Publishers.

Jo, B. W., & Khan, R. M. A. (2018). An internet of things system for underground mine air quality pollutant prediction based on azure machine learning. *Sensors (Switzerland)*, *18*(4). https://doi.org/10.3390/s18040930

Kaya, K., & Gündüz Öğüdücü, Ş. (2020). Deep Flexible Sequential (DFS) Model for Air Pollution Forecasting. *Scientific Reports*, *10*(1), 3346. https://doi.org/10.1038/s41598-020-60102-6

Keerthana, R., & Chooralil, V. S. (2020). Forecasting of the air pollution based on meteorological data and air pollutants using deep learning: A novel review. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 801–807. https://doi.org/10.30534/ijatcse/2020/115912020

Lin, Y., Mago, N., Gao, Y., Li, Y., Chiang, Y.-Y., Shahabi, C., & Ambite, J. L. (2018). Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 359–368. https://doi.org/10.1145/3274895.3274907

Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019). A machine learning model for air quality prediction for smart cities. *2019 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2019*, 452–457. https://doi.org/10.1109/WiSPNET45539.2019.9032734

Oprea, M., Mihalache, S. F., & Popescu, M. (2017). Computational intelligence-based PM<inf>2:5</inf> air pollution forecasting. *International Journal of Computers, Communications and Control*, *12*(3), 365–380. https://doi.org/10.15837/ijccc.2017.3.2907

Qi, Z., Wang, T., Song, G., Hu, W., Li, X., & Zhang, Z. (2018). Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-Grained Air Quality. *IEEE Transactions on Knowledge and Data Engineering*, *30*(12), 2285–2297. https://doi.org/10.1109/TKDE.2018.2823740

Reátegui-romero, W., Sánchez-ccoyllo, O. R., Andrade, M. D. F., Paulo, S., & Paulo, S. (2018). *PM2 . 5 Estimation with the WRF / Chem Model , Produced by Vehicular Flow in the Lima Metropolitan Area*. 215–243. https://doi.org/10.4236/ojap.2018.73011

Sánchez-ccoyllo, O. R., Ordoñez-aquino, C. G., Muñoz, Á. G., & Llacza, A. (2018). *Modeling Study of the Particulate Matter in Lima with the WRF-Chem Model : Modeling Study of the Particulate Matter in Lima with the WRF-Chem Model : Case Study of April 2016*. *June*.

Septiawan, W. M., & Endah, S. N. (2019). Suitable Recurrent Neural Network for Air Quality Prediction with Backpropagation Through Time. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 196–201. https://doi.org/10.1109/ICICOS.2018.8621720

Soh, P. W. P.-W., Chang, J.-W. J. W., & Huang, J.-W. J. W. (2018). Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations. *IEEE Access*, *6*, 38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820

Sun, X., & Xu, W. (2019). Deep Random Subspace Learning: A Spatial-Temporal Modeling Approach for Air Quality Prediction. *Atmosphere*, *10*(9), 560. https://doi.org/10.3390/atmos10090560

Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU. *IEEE Access*, *7*, 76690–76698. https://doi.org/10.1109/ACCESS.2019.2921578

Wang, B., Kong, W., Guan, H., & Xiong, N. N. (2019). Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things. *IEEE Access*, *7*, 69524–69534. https://doi.org/10.1109/ACCESS.2019.2917277

Xayasouk, T., Lee, H. M., & Lee, G. (2020). Air pollution prediction using long short-term memory (LSTM) and deep autoencoder (DAE) models. *Sustainability (Switzerland)*, *12*(6). https://doi.org/10.3390/su12062570

Zhao, C., Van Heeswijk, M., & Karhunen, J. (2017). Air quality forecasting using neural networks. *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*. https://doi.org/10.1109/SSCI.2016.7850128

Zhenghua, W., & Zhihui, T. (2018). Prediction of air quality index based on improved neural network. *2017 International Conference on Computer Systems, Electronics and Control, ICCSEC 2017*, *6*, 200–204. https://doi.org/10.1109/ICCSEC.2017.8446883

ET LUX IN TENEBRIS LUCET
MCMXVII