Clasificación Automatizada de Patrones Respiratorios para la Detección de Enfermedades Pulmonares mediante Inteligencia Artificial

Nava Fabricio[1,a,*], Juárez Mauricio[1,b,*], Chirre Luis[1,c,*], De la Cruz Lewis[2,d]
[1]Faculty of Sciences and Engineering- Biomedical Engineering,
Universidad Peruana Cayetano Heredia UPCH
[2]LID-UPCH, Laboratorios de Investigación y Desarrollo,
Universidad Peruana Cayetano Heredia UPCH
afabricio.nava@upch.pe, bmauricio.juarez@upch.pe, cluis.chirre@upch.pe, dumbert.de.la.cruz@upch.pe

***Abstract*—El proyecto propone desarrollar un sistema basado en inteligencia artificial para la detección automatizada y precisa de enfermedades pulmonares, con enfoque en la Enfermedad Pulmonar Obstructiva Crónica (EPOC), a partir de datos obtenidos mediante pruebas de espirometría. Los objetivos se centran en identificar los requerimientos del sistema, diseñar una solución técnica y funcional, y desarrollar una interfaz amigable para la clasificación de EPOC en pacientes. La metodología presenta etapas de identificación de requerimientos, desarrollo del software y validación del mismo. Se emplearon técnicas de preprocesamiento de datos, extracción de características y desarrollo de modelos de inteligencia artificial, como redes neuronales recurrentes, para analizar y clasificar los patrones respiratorios. La validación del modelo se realizó usando métricas de conjuntos de datos separados, evaluando precisión, sensibilidad y especificidad.**

Palabras clave: EPOC, enfermedades respiratorias, AI, software, diagnóstico

I. Introducción

La Enfermedad Pulmonar Obstructiva Crónica (EPOC) es un problema de salud global debido a su gran prevalencia de 10% en adultos mayores, su incidencia creciente con los años y su significante costo económico, personal y social [1]. Esta enfermedad es caracterizada como incurable la cual causa una obstrucción de flujo de aire de los pulmones y que puede tener síntomas como dificultad para respirar, tos, producción de moco, sibilancias y fatiga [2], [3]. Las causas son bastante variadas y pueden ser desde la exposición al tabaco (por fumar o exposición pasiva) o al polvo hasta eventos en la vida fetal, asma en la infancia y deficiencia en alfa-1 antitripsina [3]. Esta afección puede generar distintas complicaciones como infecciones respiratorias, problemas cardíacos, cáncer al pulmón, presión arterial alta en arterias pulmonares y en aspectos emocionales como la depresión [2].

Debido a este problema mundial, como lo es la EPOC, la Organización Mundial de la Salud adopta algunas medidas y abarcan esta complicación en su Plan de Acción Mundial para la Prevención y el Control de las Enfermedades No Transmisibles y la Agenda 2030 para el Desarrollo Sostenible de las Naciones Unidas [3]. Por otra parte, la misma organización ha creado la Global Initiative for Chronic Obstructive Lung Disease (GOLD) la cual está en constante actualización con esta enfermedad, aportando con reportes anuales sobre la definición de la EPOC, patogénesis, diagnósticos, tratamiento, etc. [4].

Las enfermedades pulmonares, como el asma, la enfermedad pulmonar obstructiva crónica (EPOC) y la fibrosis pulmonar, pueden afectar significativamente la calidad de vida de los pacientes. La detección temprana y el monitoreo continuo de los patrones respiratorios son fundamentales para el manejo efectivo de estas enfermedades. Sin embargo, la evaluación manual de los patrones respiratorios es subjetiva y puede ser propensa a errores.

Actualmente la prueba de espirometría es el gold standard en la evaluación para la identificación de obstrucción de flujo aéreo mediante la evaluación de las propiedades mecánicas presentes en el sistema respiratorio [5]. Los parámetros utilizados son la capacidad vital forzada (FVC), el volumen espiratorio forzado en el primer segundo (FEV1) y el cociente FEV1/FVC [5]. Sin embargo, el uso único de esta herramienta para la detección de EPOC no es suficiente pues influyen más variables ajenas a las propiedades mecánicas respiratorias, por ejemplo: peso, edad, género, historial de fumador, condición física, etc.

La espirometría para la evaluación de EPOC tiene buena sensibilidad pero baja especificidad y por ello es difícil diferenciar entre una EPOC y asma [6], [7]. Por ello se utiliza en conjunto con otras pruebas y análisis para tener un mejor diagnóstico [7].

Una inclusión adecuada es la aplicación de modelos computacionales de Inteligencia Artificial (IA), esta ha sido fundamental en el avance de la automatización en diversas áreas. En la salud, la IA se emplea para analizar datos médicos, diagnosticar enfermedades y personalizar tratamientos. La capacidad de la IA para analizar grandes cantidades de datos de manera rápida y precisa ha revolucionado la automatización en múltiples industrias, mejorando la eficiencia y la calidad del trabajo realizado [8].

II. METODOLOGÍA

A. Base de datos

Para el desarrollo de nuestro sistema de detección de Enfermedad Pulmonar Obstructiva Crónica (EPOC), se utilizó una base de datos compuesta por datos clínicos y parámetros fisiológicos obtenidos por bases de datos libres (Physionet), de acceso no restringido. Esta base de datos consta de pruebas de espirometría realizadas a 20 adultos jóvenes universitarios sin presencia de EPOC. En el estudio estos fueron sometidos a diferentes niveles de Presión Positiva al Final de la Espiración (PEEP), siendo los niveles de 0, 4 y 8 cmH2O. Por cada uno de estos niveles de PEEP fueron simulados cuatro niveles de EPOC, siendo 0, 200, 250 y 300mL. En total 12 muestras por sujeto con 240 muestras conformando la base de datos [9].

Adicionalmente, los datos clínicos incluyen información relevante sobre el historial médico de los pacientes, como antecedentes de tabaquismo/vaping, exposición a factores de riesgo ambientales y comorbilidades asociadas (como presencia de asma o complicaciones cardiacas). Además, se recopilaron mediciones específicas de la función pulmonar, como la capacidad vital forzada (FVC), el volumen espiratorio forzado en el primer segundo (FEV1) y el cociente FEV1/FVC.

Esta base de datos fue fundamental para definir los criterios de detección de la EPOC y para entrenar y validar nuestros modelos de inteligencia artificial. Se utilizaron técnicas de preprocesamiento de datos para eliminar artefactos y ruido, y se realizó una cuidadosa normalización de las señales respiratorias para garantizar la calidad de los datos utilizados en el desarrollo del modelo para predicción de niveles de EPOC.

B. Diseño de software

El diseño del software para la detección automatizada de EPOC se llevó a cabo siguiendo un enfoque basado en inteligencia artificial y aprendizaje automático. Durante la etapa de desarrollo del producto, se implementaron varios módulos para procesar y analizar los datos respiratorios obtenidos de las pruebas de espirometría.

El proceso de desarrollo del software incluyó las siguientes etapas:

- **Preprocesamiento de Datos**: Se aplicaron técnicas de preprocesamiento para limpiar y normalizar los datos respiratorios, eliminando artefactos y ruido para mejorar la calidad de las señales. Para esto, se utilizó `StandardScaler` para normalizar las señales de presión, flujo y volumen tidal así como sus características numéricas (Edad, Altura, Peso, etc) Para normalizar características categóricas (Género, Historial de Fumador/Vapeador, Frecuencia de Tabaquismo, etc) se utilizó `OneHotEncoder`.
- **Desarrollo de Modelos de Inteligencia Artificial**: Se implementaron modelos de aprendizaje automático, evaluando diferentes algoritmos y ajustando sus hiperparámetros para optimizar el rendimiento del sistema. Los modelos evaluados incluyeron RandomForestRegressor, GradientBoostingRegressor, SVR y XGBRegressor. Adicionalmente se utilizó la herramienta de Autogluon para poder optimizar la búsqueda de modelos óptimos adecuados a la data y tarea presente siendo específicamente dos: Clasificación (detectar si el paciente presenta EPOC o no) y Regresión (detectar el nivel de EPOC de los sujetos que sí tengan EPOC).
- **Entrenamiento del Modelo**: Se entrenaron los modelos utilizando conjuntos de datos etiquetados que contenían ejemplos de patrones respiratorios asociados con diferentes enfermedades pulmonares, incluida la EPOC. Los datos se dividieron en conjuntos de entrenamiento y prueba usando `GroupShuffleSplit` tomando al 'Subject Number' como indicador, esto con la finalidad de estratificar la data para que ambos grupos (train y test) presenten por lo menos una muestra de cada sujeto.
- **Validación y Evaluación del Modelo**: Se realizaron pruebas de validación utilizando conjuntos de datos separados para evaluar la precisión, sensibilidad y especificidad del modelo en la detección de patrones respiratorios anormales. Se reportaron las mejores configuraciones de hiperparámetros para cada modelo y se analizaron los resultados obtenidos.

Esta metodología combina técnicas de preprocesamiento de datos, extracción de características y ajuste de modelos de machine learning para desarrollar un sistema eficaz de detección de EPOC basado en datos de espirometría.

III. RESULTADOS

A. Clasificación para Determinar la Presencia de EPOC

Utilizando AutoGluon, se realizó una clasificación para determinar la presencia de EPOC. Se probaron diversos modelos, incluyendo KNeighbors, LightGBM, RandomForest, CatBoost, ExtraTrees, NeuralNet, y XGBoost. El modelo final seleccionado fue un ensamblado ponderado denominado *WeightedEnsemble_L2*. Este modelo utilizó el modelo LightGBMLarge en su totalidad.

Los resultados en el conjunto de prueba fueron excepcionales, con métricas de evaluación como se detalla a continuación:

- **Accuracy:** 0.9962
- **Balanced Accuracy:** 0.9924
- **MCC:** 0.9899
- **ROC AUC:** 1.0
- **F1 Score:** 0.9975
- **Precision:** 0.9950
- **Recall:** 1.0

B. Modelo de Regresión para Predecir la Severidad de EPOC

Para la predicción de la severidad de EPOC, se utilizó nuevamente AutoGluon. Los modelos evaluados incluyeron KNeighbors, LightGBM, RandomForest, CatBoost, ExtraTrees, NeuralNet, y XGBoost. El modelo final seleccionado fue un ensamblado ponderado denominado *WeightedEnsemble_L2*. Este modelo también combinó los resultados de múltiples modelos base para mejorar la precisión de las predicciones.

Modelos Base Combinados:

- LightGBMLarge (con un peso de 0.786)
- KNeighborsDist (con un peso de 0.214)

Los resultados obtenidos al evaluar el modelo en el conjunto de prueba fueron:

- **Root Mean Squared Error (RMSE):** 16.2177
- **Mean Squared Error (MSE):** 263.0132
- **Mean Absolute Error (MAE):** 7.6598
- **R^2:** 0.9797
- **Pearson Correlation Coefficient:** 0.9899
- **Median Absolute Error:** 3.2117

IV. DISCUSIONES

Clasificación para Determinar la Presencia de EPOC

El desempeño del modelo *WeightedEnsemble_L2* en la clasificación para determinar la presencia de EPOC fue excepcional, con una precisión de 0.9962 y un área bajo la curva ROC perfecta de 1.0. Esto indica que el modelo es extremadamente efectivo para diferenciar entre pacientes con y sin EPOC. La alta precisión, F1 score, y recall demuestran que el modelo no solo es preciso sino también robusto en la detección de casos positivos de EPOC.

El uso del ensamblado ponderado combina las fortalezas de varios modelos, lo que reduce la probabilidad de errores de predicción y mejora la generalización del modelo. Este enfoque es especialmente útil en problemas médicos donde las implicaciones de una clasificación errónea pueden ser significativas.

Modelo de Regresión para Predecir la Severidad de EPOC

El modelo de regresión para predecir la severidad de EPOC también mostró un desempeño notable, con un R^2 de 0.9797, indicando que el modelo explica casi el 98% de la variabilidad en la severidad de EPOC. El RMSE de 16.2177 y el MAE de 7.6598 son bastante bajos, lo que sugiere que las predicciones del modelo están muy cerca de los valores reales.

El uso del ensamblado ponderado permitió que el modelo lograra un equilibrio entre la precisión y la capacidad de generalización, aprovechando las fortalezas de LightGBMLarge y KNeighborsDist. Esto es crucial en la práctica médica, donde la precisión en la predicción de la severidad de una enfermedad puede influir significativamente en las decisiones de tratamiento.

Comparación con otros estudios

La inteligencia artificial ha sido ampliamente utilizada en el ámbito de la espirometría para poder diagnosticar diferentes clases de EPOC. Rosaly Moreno, et al. utilizaron variables como edad, sexo, numero de cigarros fumados diariamente, años de fumador, FVC, FEV1 y su cociente (FEV1/FVC) de un dataset con 1190 pacientes luego del EDA (Exploratory Data Analysis) para entrenar al modelo de machine learning. Se estimó las variables usando un gradient tree boosting (GTB) y un Decision Tree. Dicho modelo obtuvo métricas favorables para el diagnóstico de EPOC (sensitivity: 93%, specificity: 97%, accuracy: 95%, precision: 94%) [8]. X. Wang et al. usaron 38 variables en total como la edad, sexo, presencia de algún problema respiratorio, tos, presión, etc. para evaluar el riesgo de contraer EPOC en la población fumadora. En este estudio se elaboraron 7 modelos diferentes como Support Vector Machine, regresión logística, Random Forest, XGBoost, NGBoost, LGBM y CatBoost y luego se interpretaron los resultados de dichos modelos con Shapley additive explanations (SHAP) y Partial Dependence Plot (PDP). Luego de balancear la data el modelo de SVM demostró un sensitivity: 0.608, accuracy: 0.736, specificity: 0.8, F1: 0.372, G-mean values: 0.646 [10].

V. CONCLUSIONES

1) **Eficacia de AutoGluon:** La herramienta AutoGluon demostró ser eficaz tanto en tareas de clasificación como de regresión, permitiendo la automatización de la selección y ajuste de modelos, lo cual es particularmente útil para problemas complejos de salud.
2) **Modelos Ensamblados:** El uso de modelos ensamblados ponderados (*WeightedEnsemble_L2*) mejoró significativamente la precisión y la capacidad de generalización en ambos problemas, aprovechando las fortalezas de múltiples algoritmos.
3) **Desempeño en Clasificación:** El modelo de clasificación alcanzó un desempeño casi perfecto, sugiriendo que es altamente confiable para determinar la presencia de EPOC, lo cual puede tener un impacto positivo en la detección temprana y el manejo de la enfermedad.
4) **Desempeño en Regresión:** El modelo de regresión mostró una excelente capacidad para predecir la severidad de EPOC, lo cual es crucial para la planificación del tratamiento y el monitoreo de la enfermedad.
5) **Implicaciones para la Práctica Médica:** La alta precisión y robustez de estos modelos pueden mejorar significativamente el diagnóstico y la gestión de EPOC, ofreciendo a los profesionales de la salud herramientas más precisas y confiables para la toma de decisiones.

REFERENCES

[1] Rosaly Moreno, et al. "Artificial Intelligence Applied to Forced Spirometry in Primary Care — Open Respiratory Archives," Elsevier.es, 2022.

[2] X. Wang et al., "An explainable artificial intelligence framework for risk prediction of COPD in smokers," BMC public health, vol. 23, no. 1, Nov. 2023.

[3] Organización Mundial de la Salud, Plan de Acción Mundial para la Prevención y el Control de las Enfermedades No Transmisibles.

[4] Global Initiative for Chronic Obstructive Lung Disease (GOLD).

[5] "Espirometría para la evaluación de EPOC," Elsevier.es.

[6] E. Andreeva, M. Pokhaznikova, A. Lebedev, I. Moiseeva, O. Kuznetsova, and J.-M. Degryse, "Spirometry is not enough to diagnose COPD in epidemiological studies: a follow-up study," npj Primary Care Respiratory Medicine, vol. 27, no. 1, p. 62, 2017/11/14 2017, doi: https://doi.org/10.1038/s41533-017-0062-6.

[7] C. E. Bolton, A. A. Ionescu, P. H. Edwards, T. A. Faulkner, S. M. Edwards, and D. J. Shale, "Attaining a correct diagnosis of COPD in general practice," Respiratory Medicine, vol. 99, no. 4, pp. 493-500, 2005/04/01/ 2005, doi: https://doi.org/10.1016/j.rmed.2004.09.015.

[8] "Artificial Intelligence Applied to Forced Spirometry in Primary Care — Open Respiratory Archives," Elsevier.es, 2022. *https://www.elsevier.es/en-revista-open-respiratory-archives-11-resumen-artificial-intelligence-applied-forced-spirometry-S265966362400016X* (accessed May 29, 2024).

[9] Clifton, J. A., Guy, E. F. S., Caljé-van der Klei, T., Knopp, J., Chase, J. G. (2023). Simulated Obstructive Disease Respiratory Pressure and Flow (version 1.0.0). PhysioNet. https://doi.org/10.13026/xczc-3662.

[10] X. Wang et al., “An explainable artificial intelligence framework for risk prediction of COPD in smokers,” BMC public health, vol. 23, no. 1, Nov. 2023, doi: *https://doi.org/10.1186/s12889-023-17011-w*.

[11] Autor 3, ”Aplicación de Inteligencia Artificial en Salud,” Journal of Medical AI.