INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

Prediagnóstico de enfermedades respiratorias mediante algoritmos de Cómputo Inteligente

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

Ing. Osvaldo David Velázquez González

DIRECTORES DE TESIS:

Dr. Cornelio Yáñez Márquez

Dra. Yenny Villuendas Rey

Centro de Investigación
en Computación
Instituto Politécnico Nacional

Ciudad de México, México

Junio 2022

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL
SECRETARIA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 15 de mayo del 2022

El Colegio de Profesores de Posgrado del **Centro de Investigación en Computación** en su Sesión
(Unidad Académica)

Ordinaria No 11 celebrada el día 30 del mes noviembre de 2021, conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	VÉLAZQUEZ	Apellido Materno:	GONZÁLEZ	Nombre (s):	OSVALDO DAVID

Número de registro: B 2 0 0 4 4 1

del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Referente al registro de su tema de tesis; acordando lo siguiente:

1.- Se designa al aspirante el tema de tesis titulado:

"Prediagnóstico de enfermedades respiratorias mediante algoritmos de Cómputo Inteligente"

Objetivo general del trabajo de tesis:

Realizar un estudio comparativo del desempeño de algoritmos de clasificación inteligente de patrones en el prediagnóstico de enfermedades respiratorias.

2.- Se designa como Directores de Tesis a los profesores:

Director: **Dr. Cornelio Yáñez Márquez** 2° Director: **Dra. Yenny Villuendas Rey**

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

Centro de Investigación en Computación

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis	2° Director de Tesis
Dr. Cornelio Yáñez Márquez	**Dra. Yenny Villuendas Rey**
Aspirante	**Presidente del Colegio**
C. Osvaldo David Velázquez González	**Dr. Francisco Hiram Calvo Castro**

INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN
EN COMPUTACIÓN
SUBDIRECCIÓN
IPN-CIC

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 10:00 horas del día 02 del mes de junio del 2022 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado del: **Centro de Investigación en Computación** para examinar la tesis titulada: ***"Prediagnóstico de enfermedades respiratorias mediante algoritmos de Cómputo Inteligente"*** del (la) alumno (a):

Apellido Paterno:	VELÁZQUEZ	Apellido Materno:	GONZÁLEZ	Nombre (s):	OSVALDO DAVID

Número de registro: B 2 0 0 4 4 1

Aspirante del Programa Académico de Posgrado: **Maestría en Ciencias de la Computación**

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene _3_ % de similitud. **Se adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** [ ] **NO** [x] **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

. El 3% de similitud corresponde a metodologías adecuadamente referidas a fuente original, y a pequeños fragmentos de textos comunes en este tipo de documentos.

****Es responsabilidad del alumno como autor de la tesis la verificación antiplagio, y del Director o Directores de tesis el análisis del % de similitud para establecer el riesgo o la existencia de un posible plagio.**

Finalmente, y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** [x] **NO APROBAR** [ ] la tesis, en virtud de los motivos siguientes:

Cumplió satisfactoriamente con lo especificado en el Reglamento de Estudios de Posgrado

COMISIÓN REVISORA DE TESIS

Dr. Cornelio Yáñez Márquez
Director de tesis

Dr. Grigori Sidorov

Dra. Gudrun Kalmbach

Dra. Yenny Villuendas Rey
2° Director de tesis

Dr. Amadeo José Argüelles Cruz

Dra. Ana María Magdalena Salgado Reyes

Dr. Francisco Hiram Calvo Castro
PRESIDENTE DEL COLEGIO DE PROFESORES

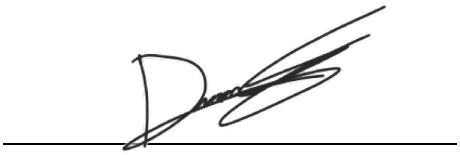
INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día 09 del mes de junio del año 2022, el (la) que suscribe Osvaldo David Velázquez González alumno(a) del programa Maestría en Ciencias de la Computación con número de registro B200441, adscrito(a) a Centro de Investigación en Computación manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección del Dr. Cornelio Yáñez Márquez y la Dra. Yenny Villuendas Rey y cede los derechos del trabajo intitulado Prediagnóstico de enfermedades respiratorias mediante algoritmos de Cómputo Inteligente, al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo vgos1097@hotmail.com. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.



Osvaldo David Velázquez González

RESUMEN

El uso de técnicas de cómputo inteligente aplicados al prediagnóstico médico de enfermedades se está convirtiendo en un área cada vez más importante de investigación a nivel mundial, debido a sus destacados resultados, así como su fácil accesibilidad y manejo.

El principal aporte de esta investigación es un novedoso algoritmo asociativo para la clasificación inteligente de patrones aplicado al prediagnóstico de las enfermedades respiratorias más comunes. El nuevo algoritmo es capaz de lidiar adecuadamente con el desbalance de clases, lo cual favorece el prediagnóstico de las enfermedades respiratorias, debido a que la mayoría de los conjuntos de datos de enfermedades respiratorias son desbalanceados. Asimismo, el algoritmo propuesto es interpretable y transparente, ya que se conocen las razones por las que un patrón de prueba fue clasificado como perteneciente a una clase específica.

Se presenta un estudio comparativo del desempeño de los algoritmos de clasificación inteligente de patrones del estado del arte en el prediagnóstico de enfermedades respiratorias. Los resultados experimentales se validaron con el propósito de encontrar posibles diferencias significativas en el rendimiento; para ello, se usaron pruebas estadísticas. Es preciso enfatizar que las pruebas experimentales realizadas, permiten verificar que el nuevo algoritmo propuesto es competitivo contra los algoritmos más utilizados en el estado del arte para el prediagnóstico médico de las enfermedades respiratorias más comunes.

ABSTRACT

The computational intelligent algorithms applied to the medical pre-diagnosis diseases have become an increasingly important area of research worldwide, due to their outstanding results, as well as their easy accessibility and handling.

The main contribution of this research is a novel associative algorithm for pattern classification applied to the pre-diagnosis of common respiratory diseases. The novel algorithm is able to deal with the imbalanced data, which favors the pre-diagnosis of respiratory diseases due to most respiratory disease datasets being imbalanced. Also, the proposed algorithm is interpretable and transparent, since the reasons why a test pattern was classified as belonging to a specific class are known.

A comparative study of the performance of state-of-the-art pattern classification algorithms in the pre-diagnosis of respiratory diseases is presented. The experimental results were validated with the purpose of finding possible significant differences in performance; for this purpose, statistical tests were used. It should be emphasized that the experimental tests performed allow verifying that the new proposed algorithm is competitive against the most used algorithms in the state-of-the-art for the medical pre-diagnosis of the most common respiratory diseases.

AGRADECIMIENTOS

Agradezco al Centro de Investigación en Computación (CIC), al Instituto Politécnico Nacional (IPN) y al CONACyT por todo el apoyo brindado para realizar esta investigación.

A mis profesores que me acompañaron en este camino de aprendizaje y superación profesional, sobre todo al Dr. Cornelio Yañez Marquez y a la Dra. Yenny Villuendas Rey, por la confianza y por sus enseñanzas, así como todo su apoyo brindado.

A mis padres, porque gracias a ellos no sería posible.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1. INTRODUCCIÓN

Ee el presente capítulo se describe una introducción detallada sobre este trabajo de tesis. Asimismo, se incluyen los antecedentes, la justificación, los orígenes del problema a resolver, su importancia en los diferentes ámbitos, los objetivos, la contribución científica y la organización del presente documento.

La detección temprana de enfermedades ha sido de suma importancia durante los últimos años, debido a los diferentes beneficios que pueden impactar en la sociedad, como lo es aumentar las posibilidades de supervivencia en pacientes que sufren enfermedades potencialmente mortales [1]. Actualmente, las investigaciones realizadas en el prediagnóstico de enfermedades son notablemente relevantes, específicamente con gran interés por minimizar los errores en la detección temprana de las enfermedades pulmonares; esto, debido a los diferentes beneficios, como aumentar la supervivencia en los pacientes, lograr una mejor recuperación gracias a la detección en una fase prematura de la enfermedad, implementar un mejor manejo clínico del paciente, adoptar medidas de salud pública y controlar posibles brotes [1]. Por otro lado, también las pruebas de diagnóstico confiables son de vital importancia en el manejo y seguimiento de brotes epidemiológicos; por ejemplo, los resultados inexactos pueden perjudicar enormemente los esfuerzos por lograr contener una posible pandemia respiratoria, como fue el caso de la pandemia de COVID-19 ocasionada por el virus SARS-CoV-2 en el año 2020 [2].

Recientemente, las técnicas de cómputo inteligente aplicadas al ámbito de la medicina se han convertido en un área de investigación cada vez más importante a nivel global, fomentando que frecuentemente en la literatura surjan trabajos relacionados al desarrollo de novedosos y avanzados modelos especializados en el prediagnóstico de enfermedades, lo cual lo convierte en un tema de investigación activo. Un aspecto muy importante relacionado con el prediagnóstico médico de enfermedades es que la mayoría de los conjuntos de datos relacionados con este tipo de problemas son desbalanceados, exhibiendo la categoría “enfermo” como la clase minoritaria. Es un hecho conocido que este fenómeno de desbalance de clases no es favorable para los algoritmos de cómputo inteligente [3].

En este contexto, el presente trabajo de tesis consiste en el prediagnóstico médico de las enfermedades respiratorias más comunes utilizando un nuevo algoritmo de cómputo inteligente propuesto, el cual será descrito más adelante en el capítulo 4 del documento. Por otro lado, se presenta en esta investigación un estudio detallado sobre el desempeño de los algoritmos de clasificación inteligente de patrones más utilizados en la literatura, aplicándolos al prediagnóstico de enfermedades respiratorias. Esto, con el fin de poder hacer un análisis comparativo y presentar las ventajas en contraste con el nuevo modelo propuesto, con respecto al estado del arte. Asimismo, el nuevo modelo es capaz de trabajar adecuadamente con conjuntos de datos severamente desbalanceados, disminuyendo el error ocasionado por dicho problema; esta es una ventaja debido a que, como se mencionó anteriormente, es común la presencia del desbalance de clases en temas relacionados al prediagnóstico médico de enfermedades.

Por otro lado, el modelo propuesto es interpretable y transparente, debido a que se conocen las razones por las que un patrón fue clasificado como perteneciente a una clase específica.

1.1 Antecedentes

De acuerdo con el Instituto Nacional del Cáncer de los Institutos Nacionales de la Salud de EE. UU. [4] las enfermedades respiratorias o enfermedades pulmonares son condiciones patológicas que afectan los pulmones y otras partes del sistema respiratorio. Existen dos tipos de enfermedades respiratorias [5]: infecciosas y crónicas, las cuales van desde sintomatologías leves, como el resfriado común, la gripe y la faringitis, hasta enfermedades potencialmente mortales como la neumonía, embolia pulmonar, tuberculosis, asma, cáncer de pulmón, fibrosis pulmonar, enfermedad pulmonar obstructiva crónica (EPOC) y síndromes respiratorios agudos graves, como lo es la enfermedad COVID-19 [6, 7].

De acuerdo con el Foro de Sociedades Respiratorias Internacionales (FIRS, por sus siglas en inglés) y datos de la Organización Mundial de la Salud (OMS), las enfermedades respiratorias se ubican entre las más importantes causas de muerte y discapacidades en todo el mundo [5, 8]. Aproximadamente 200 millones de personas (4% de la población mundial) padecen la enfermedad pulmonar obstructiva crónica, y lamentablemente 3.2 millones de las personas afectadas pierden la vida por esta enfermedad cada año; esto provoca que la enfermedad pulmonar obstructiva crónica sea la tercera causa principal de muerte a nivel mundial [5]. Por otro lado, el asma es una enfermedad que afecta a más de 350 millones de personas en todo el mundo, además de ser la enfermedad infantil crónica más frecuente. Asimismo, la neumonía provoca la pérdida de más de 2.4 millones de vidas anualmente, siendo la principal causa de muerte en niños menores de 5 años después del periodo neonatal y en adultos mayores de 65 años [5]. Adicionalmente, la neoplasia letal más común en el mundo es el cáncer de pulmón [5]. Según la Organización Mundial de la Salud (OMS), en 2019, las enfermedades respiratorias fueron tres de las diez causas principales de muerte, causando más de 8 millones de muertes anualmente [8]. Al momento de escribir este documento de tesis, la pandemia de COVID-19 iniciada en el año de 2020 ha afectado a alrededor de 400 millones de personas, cobrándose la vida de más de 6 millones de personas en todo el mundo [9, 10]. Del mismo modo, solamente en México la pandemia provocada por la enfermedad COVID-19 ha dejado un saldo de más de 320 mil muertos en un lapso de dos años [9], convirtiéndola en la principal causa de muerte a nivel nacional durante el primer semestre del año 2021 [11].

Por otro lado, el diagnóstico de enfermedades respiratorias suele realizarse aplicando diferentes métodos, tanto invasivos como no invasivos; por ejemplo, uno de los más comunes es a través del diagnóstico asistido por computadora (CAD, por sus siglas en inglés). El CAD es un conjunto de técnicas que obtienen imágenes internas del cuerpo, las cuales permiten detectar diferentes tipos de anomalías en dichas imágenes médicas [12]. Algunas de las técnicas más frecuentes utilizadas dentro del CAD para diagnosticar enfermedades respiratorias son: la radiografía de tórax, la tomografía computarizada y la resonancia magnética [12].

Por otra parte, es común realizar pruebas de función pulmonar para diagnosticar enfermedades pulmonares crónicas, tales como: asma, enfermedad pulmonar obstructiva crónica (EPOC) y enfisemas. Para ello existen varios tipos de pruebas, como la espirometría, que consiste en medir la cantidad de aire que sale del pulmón durante la respiración. La prueba de volumen pulmonar, que consiste en medir la capacidad de aire de los pulmones y la cantidad de aire que queda después de exhalar. La prueba de difusión de gases, que mide la eficiencia del traslado de gases de los pulmones al torrente sanguíneo; y la broncoscopia, que consiste en la visualización de las vías respiratorias (laringe, tráquea y bronquios) [13, 14].

De igual forma, existen métodos de diagnóstico microbiológicos para las enfermedades respiratorias infecciosas, los cuales se basan en métodos convencionales que incluyen cultivos artificiales para el aislamiento de bacterias, hongos y cultivos celulares por virus usando técnicas de detección de antígeno; sin embargo, estos métodos cuentan con un principal inconveniente, que es el tiempo necesario para obtener el diagnóstico etiológico de la infección [15]. No obstante, recientemente han surgido técnicas basadas en biología molecular, como la prueba de reacción en cadena de la polimerasa (PCR) para diagnóstico rápido de las infecciones. Este tipo de pruebas logran mejorar ligeramente el tiempo en la entrega del diagnóstico; sin embargo, en ciertas circunstancias este factor sigue siendo un problema para el adecuado tratamiento de los pacientes. Asimismo, lamentablemente algunas de estas técnicas suelen tener un costo monetario y de operación muy elevado, presentando cierto grado de complejidad debido a que requieren personal altamente entrenado y preparado para su correcta realización, provocando que personas de escasos recursos o residentes en localidades marginadas no puedan tener acceso a este tipo de diagnósticos, lo cual pone en peligro la salud pública [15].

Como se puede observar, los métodos de diagnóstico presentados anteriormente tienen desventajas y limitaciones, causando resultados negativos al implementar estas técnicas en el diagnóstico de enfermedades. Por lo tanto, es necesario continuar investigando nuevos métodos o tecnologías que ayuden a realizar un mejor diagnóstico temprano.

Es por ello que las técnicas de cómputo inteligente aplicadas a la medicina se han convertido en un área de investigación cada vez más importante en todo el mundo, así como la aplicación y desarrollo de nuevos modelos para el prediagnóstico de enfermedades, el cual es un tema de investigación activo [2, 7, 13, 16-18].

Por otro lado, el teorema de *No Free Lunch* [19] demuestra y establece que no existe ningún clasificador inteligente de patrones que sea el mejor en cualquier conjunto de datos; es decir, que logre clasificar los patrones de manera óptima sin importar el tipo de problema a resolver. La existencia del teorema *No Free Lunch* ha llevado a los investigadores a concluir que es inútil buscar el mejor clasificador. Asimismo, es innecesario pretender que un clasificador logre clasificar correctamente todos los casos o patrones del conjunto de datos, obteniendo cero errores. Entonces el trabajo de los investigadores es tratar de minimizar estos errores [20], por lo que algunos de ellos han decidido hacer uso de modelos y algoritmos de cómputo inteligente que auxilien en esta tarea de minimizar los errores de clasificación.

Dado que los modelos asociativos han mostrado ser eficaces y eficientes para lograr esta minimización de errores, en el presente trabajo de tesis se propone un nuevo modelo de clasificación especializado para el prediagnóstico de enfermedades respiratorias. Los procesos de diseño e implementación del nuevo modelo, llamado *Subtractive Threshold Associative Classifier* (*STAC*), incluyen algoritmos de cómputo inteligente, especialmente dos modelos asociativos: la Lernmatrix [21] y el Correlograph [22]. Las pruebas experimentales realizadas con el STAC, permiten verificar que el nuevo modelo es competitivo en el estado del arte.

1.2 Justificación

Actualmente, al menos 2,400 millones de personas están expuestas a contaminación del aire dentro de los edificios [5]. Además, casi toda la población mundial (el 99% de las personas) respiran aire que excede los límites de las directrices de la OMS, especialmente en países de bajos recursos, los cuales contienen altos niveles de contaminación, causando anualmente la muerte de aproximadamente 7 millones de personas en todo el mundo [23]; asimismo, más de 1,300 millones de personas están expuestas a humo de tabaco [5], provocando que estos niveles de contaminación en el aire tan elevados influyan radicalmente en la salud respiratoria a nivel mundial. Un resultado nocivo de lo anterior, es que las enfermedades respiratorias se ubican entre las más importantes causas de muerte y discapacidades en todo el mundo [5, 8]. Aunado a lo anterior, con la aparición cada vez más frecuente de nuevas enfermedades respiratorias infecciosas, se incrementa el riesgo de sufrir epidemias globales, afectando la salud pública y la vida de millones de personas.

En este contexto, el presente trabajo de tesis se justifica plenamente, porque la detección temprana de las enfermedades respiratorias potencialmente mortales puede ayudar a mejorar la esperanza de vida de los seres humanos y, al mismo tiempo, puede aportar información útil para el desarrollo de tratamientos adecuados.

1.3 Hipótesis

Es posible generar un nuevo algoritmo asociativo para la clasificación inteligente de patrones basado en dos modelos de memorias asociativas pioneras, *Lernmatrix* y *Correlograph*, el cual sea capaz de trabajar adecuadamente con el desbalance de clases, y logre así competir con los diferentes modelos de la literatura aplicados a los conjuntos de datos pertenecientes al prediagnóstico de las enfermedades respiratorias más comunes.

1.4 Objetivos

Objetivo General

Proponer e implementar un nuevo algoritmo de clasificación inteligente de patrones, *Subtractive Threshold Associative Classifier* (STAC), el cual está basado en la fusión de dos modelos asociativos: Lernmatrix y Correlograph, para el prediagnostico médico de enfermedades respiratorias.

Objetivos Específicos

- Recolectar conjuntos de datos que abarquen las enfermedades respiratorias más comunes.
- Analizar el comportamiento de los algoritmos de clasificación inteligente de patrones basados en memorias asociativas, así como los algoritmos clásicos más usados en la literatura.
- Implementar el algoritmo propuesto STAC y aplicarlo a los conjuntos de datos de enfermedades respiratorias.
- Comparar el desempeño del algoritmo STAC con el de los algoritmos del estado del arte.

1.5 Contribuciones

Las contribuciones del presente trabajo de tesis son las siguientes:

- Un nuevo algoritmo asociativo de clasificación inteligente de patrones que compite con los algoritmos más usados en la literatura.
- Un estudio experimental que permite evaluar el comportamiento de diversos algoritmos de clasificación aplicados sobre datos de enfermedades respiratorias

1.6 Organización del documento de tesis

El resto del documento de tesis está organizado de la siguiente manera. En el capítulo 2 se presentan los trabajos relacionados, donde se hace una breve descripción de los diferentes trabajos publicados sobre la aplicación de algoritmos de calificación inteligente de patrones para el prediagnóstico de enfermedades respiratorias presentes dentro del estado del arte. Por otro lado, en el capítulo 3 se presentan los materiales y métodos, donde se destacan los conceptos teóricos que darán soporte al presente trabajo, así como una descripción de los diferentes conjuntos de datos bajo estudio. Asimismo, la propuesta de este trabajo se aborda en el capítulo 4, y en el capítulo 5 se muestran los resultados alcanzados después de la fase experimental para evaluar de esa forma la viabilidad del nuevo modelo; finalmente, las conclusiones y las propuestas para trabajos futuros se presentan en el capítulo 6.

CAPÍTULO 2. TRABAJOS RELACIONADOS

En este capítulo se describen brevemente lo más recientes trabajos presentes en la literatura relacionados al tema de investigación sobre el prediagnóstico medico de enfermedades respiratorias mediante la aplicación de algoritmos de cómputo inteligente.

La detección temprana de enfermedades ha incrementado su relevancia en años recientes, a causa de los diversos beneficios que impactan provechosamente en la salud pública, como lo son, aumentar las posibilidades de supervivencia en pacientes que padecen enfermedades respiratorias graves [1], lograr una mejor recuperación gracias a la detección en una fase prematura de la enfermedad, un mejor manejo clínico del paciente, facilidad de poder adaptar medidas de salud pública y controlar posibles brotes epidemiológicos [1].

Las investigaciones enfocadas en el prediagnóstico de enfermedades respiratorias han tomado fuerza recientemente a nivel mundial, con amplio interés en mejorar la detección temprana de enfermedades respiratorias. Actualmente, para realizar este tipo de diagnóstico en enfermedades respiratorias se aplican distintos métodos, tanto invasivos como no invasivos. Algunos de los métodos más comunes son el diagnóstico asistido por computadora (CAD), las pruebas de función pulmonar (tales como la espirometría, volúmen pulmonar, difusión de gases, y broncoscopía), diagnósticos microbiológicos y diagnósticos basados en biología molecular [12-15]. Sin embargo, últimamente el uso de técnicas de cómputo inteligente aplicados en el ámbito del prediagnóstico de enfermedades se ha convertido en un área de investigación paulatinamente más importante a nivel global, debido a su facilidad de implementación y acceso [15]. Provocando que frecuentemente en la literatura surjan investigaciones relacionados al desarrollo de novedosos modelos especializados para la prediagnostico medico de todo tipo de enfermedades [2, 7, 13, 16-18].

Existen numerosas técnicas para realizar un prediagnóstico de enfermedades aplicando modelos de clasificación inteligente de patrones. Uno de los modelos más conocidos en la literatura es el clasificador k-NN (*K* vecinos más cercanos) [24], el cual se basa en métricas. De igual forma, existen clasificadores basados en probabilidades, como el Naïve Bayes [25] y basados en árboles de decisión, como el C4.5 [26]. Por otro lado, se ha destacado la implementación de los modelos de Máquinas de Soporte Vectorial (SVM) [27] y los modelos inspirados en el cerebro biológico, como lo son las Redes Neuronales Artificiales [28].

Dentro de la literatura diversos trabajos han abordado el tema del prediagnóstico de enfermedades respiratorias aplicando técnicas de Cómputo Inteligente. En estas investigaciones se han utilizado diversos modelos de clasificación, de los cuales algunos de los más destacados se describirán brevemente a continuación.

Maleki et al. [24] abordaron el prediagnóstico de cáncer de pulmón, una de las enfermedades más comunes entre los humanos a nivel mundial. Para la tarea de clasificación los autores utilizan datos generales referentes a pacientes que sufren cáncer de pulmón. El conjunto de datos bajo consideración para el estudio de esa investigación está conformado por 100 patrones con 23 características, las cuales describen información sobre los pacientes, tales como su edad, género,

contaminación de aire, consumo de sustancias, riesgo genético, peso, dolores crónicos, dieta alimenticia, actividad física, entre otras más. Este conjunto de datos se encuentra etiquetado en 3 diferentes clases (*low*, *medium*, *high*), los cuales representan el nivel de riesgo de sufrir cáncer de pulmón. Los autores proponen una nueva metodología para mejorar el diagnóstico, el cual consiste en varias etapas. La primera etapa consiste en aplicar un pre-procesamiento en caso de ser necesario, para así eliminar los valores perdidos que puedan tener los patrones. Posteriormente, para reducir las dimensiones del conjunto de datos y mejorar el desempeño del clasificador a utilizar, los autores aplican un método de selección de características empleando un algoritmo genético para encontrar la mejor combinación de características; en este proceso se obtiene un vector, el cual indica las características que fueron seleccionadas por el algoritmo. El algoritmo de selección de características escogió únicamente 6 de ellas, el cual convergió después de la cuarta iteración con una función de costo igual a 0.532. Después, para separar el conjunto de datos en dos diferentes particiones (entrenamiento y de prueba) los autores aplican el método de validación *10-fold cross validation*. Por último, en el proceso de clasificación, en este caso usado para diagnosticar cáncer de pulmón, se aplica el algoritmo kNN, obteniendo el valor de *k* de forma experimental, encontrando que el mejor valor de *k* corresponde a esta tarea es igual a 6, ya que con esta configuración se alcanzan valores de *accuracy* máximos, iguales al 100%. Finalmente, los autores demuestran que es posible obtener un método de ayuda médico para diagnosticar a pacientes que sufren cáncer de pulmón de forma temprana.

Por otro lado, Spathis et al. [18] estudiaron la prevención, diagnóstico y detección temprana de enfermedades respiratorias, tales como el asma y la enfermedad pulmonar obstructiva crónica (EPOC). Esto debido a que son enfermedades comunes entre la población, ocasionando millones de muertes prematuras a nivel mundial. El estudio realizado por los autores examina los factores que caracterizan el diagnóstico de asma y EPOC utilizando diversas técnicas de aprendizaje automático. Para ello, se realizó un estudio comparativo aplicando diferentes algoritmos, tales como: *Naïve Bayes*, regresión logística, redes neuronales de perceptrón multicapa (MLP), máquinas de soporte vectorial (SVM), vecinos cercanos (kNN), árboles de decisión, y *Random Forest*. El conjunto de datos bajo consideración para dicho estudio se encuentra conformado por 132 patrones los cuales están compuestos por 22 características. Con el objetivo de encontrar el algoritmo más adecuado para este tipo de problemas, en el estudio realizado por los autores se aplicó el método de validación *10-fold cross validation* y *5-fold cross validation*, así como la medida de desempeño *accuracy*. Como resultado del análisis comparativo se observó que el mejor algoritmo de clasificación para diagnosticar la enfermedad de asma y EPOC es el algoritmo *random forest*, el cual obtuvo los valores más altos de *accuracy*, superando significativamente a los demás algoritmos implementados.

En el trabajo presentado por Cardoso et al. [29] propusieron una nueva metodología para diagnosticar la enfermedad pulmonar intersticial (EPI) obtenido mejores resultados en el diagnóstico sobre los trabajos relacionados del esto del arte. En este caso, se hizo uso de un conjunto de datos compuesto por 3252 patrones, con 28 características cada uno; sin embargo, los autores aplicaron técnicas de extracción de características para reducir la dimensionalidad de dichos patrones, alcanzado reducir a 5 características por patrón. Para ello utilizaron el método de Análisis de componentes principales (PCA) y análisis discriminante lineal.

Por otro lado, los datos resultantes por los métodos de extracción de características obtenidos previamente son utilizados para entrenar modelos de aprendizaje automático, y de ese modo realizar un análisis comparativo, para ello se hizo uso de los modelos: SVM, kNN, *Gaussian Mixture Model* y una red neuronal profunda *Feedforward* (DFNN). Asimismo, se aplicó técnicas de redes neuronales convolucionales (CNN) directamente a los datos; es decir, sin la aplicación previa de las técnicas de extracción de características al conjunto de datos. Los autores aplicaron como método de validación el *10-fold cross validation* y como medida de desempeño el valor de *accuracy*. Como resultado de los experimentos realizados en esa investigación, se observó que el mejor algoritmo de clasificación para el diagnóstico de la enfermedad EPI fue el algoritmo DFNN, el cual obtuvo el mejor valor de *accuracy*.

Finkelstein et al. [30] utilizaron tres algoritmos de aprendizaje automático (*Naïve Bayes*, red bayesiana adaptativa, y máquinas de soporte de vectorial) para realizar un análisis comparativo en la detección temprana de exacerbaciones en pacientes adultos con asma. El conjunto de datos usado para el estudio exploratorio consiste de 7001 patrones conformados a partir de 20 características. Por otro lado, para la investigación presentada por los autores utilizaron los valores de *sensitivity* y *specificity* como medida de desempeño, asimismo se aplicó *hold-out* como método de validación, usando 70% del conjunto de datos para el entrenamiento y el resto 30% como conjunto de prueba. Finalmente, en base a los resultados obtenidos por la investigación realizada, los autores demostraron que es posible y viable diagnosticar, así como la detectar de forma temprana exacerbaciones del asma agudas, concluyendo que el mejor algoritmo para dicha tarea es la red bayesiana adaptativa con un valor de *sensitivity* y specificity igual a 100%.

Amaral et al. [31] desarrollaron un sistema de apoyo en la decisión médica para simplificar el uso clínico así como mejorar el diagnóstico de obstrucción de las vías respiratorias en pacientes que sufren asma. Para el estudio se utilizó un conjunto de datos de 75 patrones, de los cuales 39 representaban a pacientes que sufrían obstrucción en las vías respiratorias y 36 con pacientes que no sufrían tal condición. En el estudio realizado se aplicó el método de validación *10-fold cross-validation* y la medida de desempeño *accuracy*. Por otro lado, para el estudio comparativo se utilizaron técnicas de extracción de características, tales como el análisis de componentes principales (PCA) para que tratar de mejorar el desempeño en la clasificación; sin embargo, en base a los resultados obtenidos por los autores en dicha investigación, se concluyó que el uso de una técnica de extracción de características no beneficia significativamente el desempeño de los algoritmos en este caso particular. Por lo tanto, se demostró que el mejor algoritmo para diagnosticar la obstrucción de las vías respiratorias en pacientes con asma son el algoritmo kNN con un valor de *k=1* y el clasificador *AdaBoost*, los cuales permiten clasificar con un desempeño sobresaliente a los pacientes enfermos.

Con el objetivo de incrementar la tasa de supervivencia en pacientes que sufren cáncer de pulmón, Radhika et al. [32] proponen diagnosticar de forma temprana el cáncer de pulmón en pacientes afectados utilizando algoritmos de clasificación inteligente de patrones, tales como: Naïve Bayes, Maquinas de Soporte Vectorial (SVM) y regresión logística. Para ello, los autores mencionan que la clave para mejorar el diagnóstico de cáncer de pulmón es hacer el proceso de diagnóstico más eficiente y efectivo, lo cual se podría conseguir al mejorar la capacidad con la que los algoritmos logran aprender a reconocer a los pacientes en sus diferentes clases.

En esta investigación los autores utilizaron dos conjuntos reconocidos referentes a la clasificación de pacientes con 3 tipos de cáncer de pulmón (maligno, pre-benigno, y benigno).

En otra tendencia, recientemente han surgido novedosas técnicas de aprendizaje automático, las cuales trabajan adecuadamente utilizando imágenes como información de entrada, logrando superar fácilmente a los demás algoritmos en este tipo de tareas [33]. Estas nuevas técnicas son denominadas aprendizaje profundo (*Deep Learning*) o redes neuronales convolucionales (CNN, *convolutional neural networks*). Por ejemplo Xiong et al. [34] propusieron una modelo CNN especializado para reconocer *Mycobacterium* tuberculosis utilizando muestras de tejido tratadas con tinción acidorresistente, recolectados del Departamento de patología del primer hospital de la universidad de Pekín entre enero de 2016 y junio de 2017. Después de los experimentos realizados, el nuevo modelo CNN propuesto logró valores de *sensitivity* de 97.94% y *specificity* de 83.65%. Por tal motivo, los autores concluyeron que su nuevo modelo puede ser un sistema de apoyo prometedor para detectar Mycobacterium tuberculosis y ayudar en la toma de decisiones clínicas.

Otro ejemplo bajo el mismo grupo de algoritmos se encuentra Christe et al. [35] que presentan un estudio para evaluar el desempeño de un nuevo sistema de diagnóstico asistido por computadora basado en una red neuronal convolucional (CNN) para la clasificación automática de imágenes de tomografía computarizada de alta resolución en cuatro categorías de diagnóstico radiológico, así como realizar una comparación con el desempeño de radiólogos profesionales bajo la misma tarea de diagnóstico. Los autores encontraron que el nuevo sistema de diagnóstico asistido por computadora basado en aprendizaje profundo fue capaz de clasificar la fibrosis pulmonar idiopática con un valor de *accuracy* similar a un experto humano.

Asimismo, se encuentran trabajos relacionados donde se aplican las técnicas relacionadas al aprendizaje profundo para prediagnosticar a pacientes que sufren enfermedades respiratorias de COVID-19 o neumonía [7, 17]. En estas investigaciones se utilizan redes neuronales convolucionales pre-entrenadas, para clasificar imágenes de radiografías de tórax. Las investigaciones realizadas demuestran que es una herramienta prometedora, la cual ayuda a diagnosticar clínicamente a pacientes con este tipo de enfermedades.

Finalmente, dentro de la literatura relacionada se encuentra trabajos donde se han utilizado señales acústicas pulmonares de la ecografía torácica de los pacientes, con el fin de realizar diagnósticos de enfermedades vinculas al tórax, tales como derrame pleural, atelectasis, neumotórax y neumonía [36]. Por ejemplo, Pham et al. [16] hacen uso de redes neuronales convolucionales para detectar enfermedades respiratorias a partir de grabaciones de sonidos respiratorios. Para ellos los investigadores extraen las características de las grabaciones, transformando los sonidos en espectrogramas, las cuales posteriormente son utilizadas por la red neuronal para aprender y clasificar las anomalías presentadas por las grabaciones. De igual forma, Palaniappan et al. [37] detectan enfermedades pulmonares (patología de obstrucción de las vías respiratorias y patología del parénquima) mediante señales acústicas; pero, utilizando modelos tradicionales de aprendizaje automático, como lo son los algoritmos Máquinas de Soporte Vectorial (SVM) y Vecinos más cercanos (kNN), asimismo, los investigadores utilizan sonidos respiratorios del conjunto de datos R.A.L.E.

En este capítulo, se presentó una breve compilación de los resultados de otras investigaciones realizadas bajo el mismo tema de estudio, el prediagnóstico medico de enfermedades respiratorias utilizando técnicas de computó inteligente.

CAPÍTULO 3. MATERIALES Y MÉTODOS

En este capítulo se describen detalladamente los materiales y métodos que son usados con el objetivo de efectuar la propuesta de este trabajo de tesis. Dentro del capítulo se presenta una breve descripción de los conjuntos de datos seleccionados y los diferentes algoritmos de clasificación inteligente de patrones propuestos para el estudio comparativo; asimismo, en este capítulo se incluyen conceptos fundamentales de dos modelos pioneros de memorias asociativas, la *Lernmatrix* de *Steinbuch* [21] y el *Correlograph* de *Willshaw* [22], esto debido a que dichos modelos son base para el modelo propuesto presentado en el capítulo 4.

3.1 Conjuntos de datos seleccionados

Para este trabajo se seleccionaron 12 conjuntos de datos en tres repositorios diferentes, el repositorio *Knowledge Extraction base on Evolutionary Learning* (KEEL) [38] ubicado en https://sci2s.ugr.es/keel/datasets.php, el repositorio de Aprendizaje Automático de la Universidad de California en Irvine (UCI) [39] ubicado en https://archive.ics.uci.edu/ml/index.php, y por último, el repositorio de Kaggle ubicado en la dirección https://www.kaggle.com/datasets. De los 12 conjuntos de datos seleccionados, 10 tienen un índice de desbalance (IR) mayor a 1.5. El índice IR se calcula como en la expresión **1**. De acuerdo con [40], esta característica significa que los conjuntos de datos, el cual representa el 83.3% de ellos, son desbalanceados; es decir, existe un desequilibrio en la cantidad de patrones en las diferentes clases dentro de todo el conjunto de datos. En la **Tabla 1.** Descripción de los conjuntos de datos seleccionados se muestra información detallada sobre cada uno de los conjuntos de datos seleccionados.

Tabla 1. *Descripción de los conjuntos de datos seleccionados.*

Conjuntos de datos	Características		Patrones	IR	Clases
	Categóricas	Numéricas			
Post-operative	8	0	90	32.00	3
Thyroid	0	21	7200	40.10	3
Newt-thyroid1	5	0	215	5.14	2
Newt-thyroid2	5	0	215	5.14	2
Thoracic-Surgery	13	3	470	5.70	2
Lung-Cancer	0	52	32	1.40	3
Survey Lung-Cancer	14	1	309	6.90	2
ACPs Lung Cancer	38	0	901	31.25	4
Exasens-COPD	0	7	80	1.00	2
Lymphography	3	15	148	40.50	4
Lymphography-NF	3	15	148	23.60	2
Primary-tumor	16	1	336	42.00	18

El índice de desbalance de clases (IR) se calcula de la siguiente forma:

$$IR = \frac{Cantidad\ de\ patrones\ de\ la\ clase\ mayoritaria}{Cantidad\ de\ patrones\ de\ la\ clase\ minoritaria} \quad (1)$$

Por otro lado, los 12 conjuntos de datos mencionados anteriormente fueron seleccionados debido a que incluyen información sobre las enfermedades respiratorias más comunes [5], como la neumonía, embolia pulmonar, tuberculosis, asma, cáncer de pulmón, fibrosis pulmonar, enfermedad pulmonar obstructiva crónica (EPOC) y síndromes respiratorios agudos graves.

A continuación, se presenta una breve descripción sobre cada uno de los conjuntos de datos seleccionados.

Post-operative: Este conjunto de datos proviene de un estudio para determinar a dónde se debe enviar a un paciente después de una recuperación postoperatoria, debido a que la hipotermia es un riesgo importante posterior a la cirugía. Asimismo, las características representan aproximadamente las mediciones de la temperatura del cuerpo. Por último, el conjunto de datos fue recuperado del repositorio de KEEL en https://sci2s.ugr.es/keel/dataset.php?cod=179.

Thyroid: Este conjunto de datos fue obtenido del repositorio de la UCI en https://archive.ics.uci.edu/ml/datasets/thyroid+disease. La tarea de clasificación de este conjunto de datos consiste en determinar si un paciente dado está sano (normal) o sufre hipotiroidismo o hipertiroidismo.

Newt-thyroid1 y Newt-thyroid2: Estos conjuntos de datos fueron recuperados del repositorio de KEEL en https://sci2s.ugr.es/keel/dataset.php?cod=145 y https://sci2s.ugr.es/keel/dataset.php?cod=146, respectivamente. Ambos conjuntos de datos representan una versión desbalanceada del conjunto de datos original Thyroid. En el conjunto Newt-thyoid1, la clase positiva pertenece a la clase hipertiroidismo, y los patrones de la clase negativa está conformada por los patrones del resto de clases. Por otro lado, en Newt-thyroid2, la clase positiva pertenece a la clase hipotiroidismo, y la clase negativa se encuentra conformada por el resto.

Thoracic-Surgery: Este conjunto de datos fue recolectado del repositorio de la UCI en https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data, y representa a pacientes que se sometieron a resecciones pulmonares mayores por cáncer de pulmón primario entre los años 2007 y 2009 en el Centro de Cirugía Torácica de Wroclaw.

Lung-Cancer: Este conjunto de datos fue obtenido del repositorio de UCI en https://archive.ics.uci.edu/ml/datasets/lung+cancer y describe tres tipos de cánceres de pulmón patológicos. El objetivo del conjunto de datos es lograr clasificar a estos tres tipos de cánceres.

Survey Lung-Cancer: La tarea de clasificación de este conjunto de datos es detectar si un paciente dado sufre o no cáncer de pulmón, en base a diferentes variables recolectadas de una encuesta. El conjunto fue obtenido del repositorio de Kaggle en https://www.kaggle.com/mysarahmadbhat/lung-cancer.

ACPs Lung Cancer: Este conjunto de datos fue obtenido del repositorio de UCI en https://archive.ics.uci.edu/ml/datasets/Anticancer+peptides, el cual representa información sobre péptidos (código de aminoácidos) y la actividad anticancerígena en líneas celulares de cáncer de pulmón.

Exasens-COPD: Este conjunto de datos tiene como objetivo (a partir de información demográfica de la saliva) clasificar a pacientes dentro de cuatro clases según su pertenencia: enfermedad pulmonar obstructiva crónica, EPOC o COPD (por sus siglas en inglés), asma, infecciones respiratorias y pacientes completamente sanos. El conjunto de datos fue recolectado del repositorio de la UCI en https://archive.ics.uci.edu/ml/datasets/Exasens.

Lymphography: La tarea de clasificación de este conjunto de datos es detectar la presencia de linfomas además de su estado actual. El conjunto fue recolectado del repositorio de la UCI en https://archive.ics.uci.edu/ml/datasets/Lymphography.

Lymphography-NF: Este conjunto de datos es una versión de únicamente dos clases del repositorio de KEEL sobre el conjunto de datos original Lymphography. En este conjunto la clase positiva está conformada por las clases "normal" y "fibrosis" mientras que la clase negativa está compuesta por el resto de las clases. Debido a esta reagrupación de clases, el índice IR sufre una disminución, teniendo un valor de 23.67. El conjunto de datos fue obtenido de https://sci2s.ugr.es/keel/dataset.php?cod=1337.

Primary-tumor: Este conjunto de datos tiene como objetivo clasificar a pacientes dentro de 21 clases diferentes, según el tipo de tumor que sufren. Fue recolectado del repositorio de la UCI Machine Learning en https://archive.ics.uci.edu/ml/datasets/primary+tumor.

3.2 Algoritmos de clasificación inteligente de patrones

En esta sección se describen los algoritmos de clasificación inteligente de patrones propuestos para realizar el estudio comparativo con el nuevo modelo presentado en el presente trabajo, los cuales son aplicados a los bancos de datos que se describirán en la sección 3.1. Todos los algoritmos fueron ejecutados en el software de minería de datos WEKA 3.8 [41] usando las configuraciones por defecto presentadas por el software en cada clasificador. Adicionalmente, se utilizó una computadora portátil con un procesador CPU Intel(R) Core(TM) i7-7500U a una velocidad de 2.70Ghz con 16GB de RAM utilizando el sistema operativo Windows 10.

Por otro lado, se seleccionaron los algoritmos presentados a continuación debido a que comprenden a los modelos más relevantes en la tabla de resultados dentro del estado del arte en temas relacionados a la clasificación inteligente de patrones, como se logra observar en [42-44].

Naîve Bayes [25] es un tipo de algoritmo que pertenece a los clasificadores basados en probabilidades. Este algoritmo de clasificación está basado en el Teorema de Bayes, específicamente considerando a todos los atributos independientes desde un enfoque probabilístico.

Otro clasificador utilizado fue el algoritmo **kNN** o ***K-nearest neighbor*** [24], específicamente los modelos 1NN y 3NN. En WEKA el algoritmo de clasificadores es llamado *Instance-Based* (IBk).

La familiar de estos clasificadores se basa en asignar a un patrón de entrada (patrón de clase desconocida) la clase a la que pertenecen los *k* vecinos más cercanos según una función de distancia, usualmente, la distancia euclidiana.

***Multilayer perceptron* (MLP)** [28, 45] es un algoritmo de clasificación muy reconocido dentro de la literatura sobre temas relacionados a Machine Learning.

MLP es una red compuesta de neuronas artificiales (también denominadas *units*) interconectadas entre sí conformando tres diferentes tipos de capas (*layers*), las cuales son: la capa de entrada (input *layer*), la capa oculta (*hidden layer*) y por último la capa de salida (*output layer*). Esta última capa está formada por las neuronas cuyos valores de salida corresponden a la etiqueta de clase de los patrones del conjunto de datos. Por otro lado, este clasificador se encuentra basado en la "Regla Delta Generalizada", la cual adapta los pesos propagando los errores hacia atrás o comúnmente denominada *backpropagation*.

Sequential minimal optimization (**SMO)** [27] es uno de los algoritmos de optimización para máquinas de soporte vectorial (SVM) más importantes y ampliamente utilizado dentro del estado del arte al momento de comparar clasificadores. Este clasificador utiliza el algoritmo de optimización mínima secuencial creado por John Platt para entrenar máquinas de soporte vectorial utilizando funciones *kernel* basadas en funciones lineal, polinomial, de base radial o sigmoide. Este clasificador encuentra un hiperplano que intenta separar eficientemente los subconjuntos de patrones de cada una de las clases.

C4.5 [26] es un árbol de decisión, el cual es una extensión del algoritmo ID3 [46]. Este tipo de clasificador es muy reconocido dentro del estado del arte debido a que es explicable, está basado en la teoría de la información y su estructura jerárquica permite ver cómo se clasifican los patrones de un conjunto de datos.

3.3 Memorias asociativas

Una memoria asociativa $\boldsymbol{M}$ es un sistema de entrada y salida de patrones (ver ecuación **2**), que tiene como objetivo principal aprender a recuperar correctamente patrones completos a partir de patrones de entrada, los cuales pueden estar alterados con diferentes tipos de ruidos (aditivo, sustractivo o mixto) [21, 47].

$$x \rightarrow \boxed{\boldsymbol{M}} \rightarrow y \tag{2}$$

En una memoria asociativa, el patrón de entrada y el patrón de salida son representados por un vector columna denotado como x y y, respectivamente. Cada patrón de entrada se encuentra asociado con un correspondiente patrón de salida, dicha asociación es denotado como (x, y) [21, 47].

Asimismo, una memoria asociativa $\boldsymbol{M}$ es representada por una matriz, cuyo componente ij-ésima es m_{ij}. La matriz $\boldsymbol{M}$ se genera por un conjunto finito de asociaciones previamente conocidas, el cual es nombrado conjunto fundamental. La cardinalidad de este conjunto fundamental es denotada como p. Para un entero positivo μ, la asociación correspondiente se denota de la siguiente forma [21, 47]:

$$\{(x^{\mu}, y^{\mu}) \mid \mu = 1,2, \dots, p\} \tag{3}$$

Existen dos tipos de memorias asociativas. La memoria autoasociativa, la cual cumple con las condiciones siguientes: $x^{\mu} = y^{\mu}\ \forall \mu \in \{1,2, \dots, p\}$. Por otro lado, se declara que la memoria es heteroasociativa si se cumple que $x^{\mu} \neq y^{\mu}\ \exists \mu \in \{1,2, \dots, p\}$ [47].

Las memorias asociativas se componen de dos fases esenciales [47]:

> **Fase de aprendizaje**. Consiste en crear la memoria asociativa (matriz) $\boldsymbol{M}$ que logre almacenar las p asociaciones del conjunto fundamental. La matriz creada tendrá dimensiones $p \times n$, donde n es igual a la dimensión de los patrones de entrada.
> **Fase de recuperación**. Consiste en operar la memoria asociativa (matriz) $\boldsymbol{M}$ con el objetivo de encontrar las condiciones suficientes para obtener el patrón fundamental de salida y^{μ} a partir del patrón fundamental de entrada x^{μ} para cada uno de los elementos del conjunto fundamental.

3.5.1 Lernmatrix de Steinbuch

La *Lernmatrix* de *Steinbuch* [21] es una memoria heteroasociativa, la cual puede funcionar de igual forma como un algoritmo de clasificación de patrones binarios si se escogen correctamente los patrones de salida correspondientes a cada patrón de entrada.

Por lo tanto, si se maneja la Lernmatrix como clasificador, el valor de salida que representa la etiqueta de clase, será sustituida por un vector columna denominado vector *one-hot*, cuyos componentes son asignados cumpliendo la siguiente expresión: $y_k^k = 1$ y $y_j^k = 0,\ j \neq k$; es decir, el valor de la componente k del patrón de salida y^k es *1* mientras todas las demás componentes tienen un valor de *0*. Las fases operacionales son descritas a continuación [48].

Fase de aprendizaje de la Lernmatrix

La fase de aprendizaje consiste en encontrar una forma de generar una matriz $\boldsymbol{M}$ que almacene la información de las p asociaciones del conjunto fundamental. El proceso para determinar cada uno de los componentes m_{ij} se puede describir en dos pasos [48].

1. Cada uno de los componente m_{ij} de la matriz $\boldsymbol{M}$ es inicializada en ceros.
2. Cada componente m_{ij} se actualiza de acuerdo con la regla $m_{ij} + \Delta m_{ij}$, donde:

$$\Delta m_{ij} = \begin{cases} +\varepsilon & si\ y_i^\mu = 1 = x_j^\mu \\ -\varepsilon & si\ y_i^\mu = 1\ y\ x_j^\mu = 0 \\ 0 & en\ cualquier\ otro\ caso \end{cases} \quad (4)$$

Donde cada ε representa cualquier constante positiva previamente seleccionada.

Para ilustrar más detalladamente el proceso de aprendizaje de la Lernmatrix de Steinbuch se presenta a continuación un breve ejemplo.

Fase de recuperación (clasificación) de la Lernmatrix

La fase de recuperación o clasificación en caso de usarse como clasificador consiste en multiplicar la memoria $\boldsymbol{M}$ entrenada anteriormente con un vector de entrada desconocido, con el objetivo de lograr encontrar la clase a la que pertenece dicho vector de entrada. Se espera que con este proceso se obtenga un vector de salida *one-hot* que represente la clase a la que pertenece el patrón desconocido; sin embargo, esto no siempre se consigue [48].

Para realizar la fase de recuperación es necesario calcular la i-*ésima* coordenada del vector de salida (vector que representa la clase del patrón), la cual se obtiene usando la siguiente expresión [21, 48]:

$$y_i^\omega = \begin{cases} 1 & si\ \ \sum_{j=1}^{n} m_{ij} \cdot x_j^\omega = \bigvee_{h=1}^{p} \left[\sum_{j=1}^{n} m_{hj} \cdot x_j^\omega \right] \\ 0 & en\ otro\ caso \end{cases} \quad (5)$$

Se muestra a continuación un ejemplo ilustrativo del proceso operacional de la fase de aprendizaje y recuperación de la *Lernmatrix* de *Steinbuch*. Los patrones y sus asociaciones correspondientes utilizadas para el ejemplo son los siguientes:

$$x^1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} y^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix};\ x^2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} y^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix};\ x^3 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} y^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix};\ x^4 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} y^4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

En este caso, se tienen cuatro clases de patrones de dimensión 7, es decir $n = 7$ y $p = 4$. Como se describió anteriormente, el primero paso para la fase de aprendizaje es crear una matriz de dimensiones $p = 4\ \times\ n = 7$ llena de ceros, por lo tanto, la matriz inicial resulta de la siguiente forma:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Posteriormente, el segundo paso de la fase de aprendizaje es actualizar la matriz creada acorde a la regla descrita en la ecuación **4**; finalmente, al sumar las cuatro asociaciones, se obtiene la siguiente memoria $\boldsymbol{M}$ entrenada:

$$M = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon \end{pmatrix}$$

Como se mencionó anteriormente, la fase de recuperación consiste en obtener a la salida un vector de clase *one-hot* que represente la clase asignada del patrón. Este vector se consigue por medio de la expresión descrita en la ecuación **5**. Por lo tanto, para cada uno de los patrones de entrada se obtienen los siguientes resultados:

$$M \cdot x^1 = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 4\varepsilon \\ 0 \\ 4\varepsilon \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \therefore \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \neq y^1$$

$$M \cdot x^2 = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 4\varepsilon \\ 4\varepsilon \\ 2\varepsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \therefore \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \neq y^2$$

$$M \cdot x^3 = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \varepsilon \\ \varepsilon \\ 7\varepsilon \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \therefore \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = y^3$$

$$M \cdot x^4 = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \\ \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2\varepsilon \\ 4\varepsilon \\ 4\varepsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \therefore \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \neq y^4$$

Se logra observar que, de los cuatro patrones recuperados, solamente el patrón de salida y^3 corresponde a un patrón *one-hot*. Esto ocurre debido a que un patrón de salida y^k es *one-hot* si cumple con qué $y_k^k = 1$ y $y_j^k = 0,\ j \neq k$. Es por ello por lo que los patrones de salida y^1, y^2 y y^4 no son *one-hot*, por el hecho de que no cumplen con la característica antes mencionada. Por lo tanto, únicamente es recuperado (clasificado) correctamente el patrón y^3 de los cuatro patrones a clasificar, debido a que los demás patrones manifiestan ambigüedad en la asignación de clases, clasificando correctamente solo el 25% de los patrones.

3.5.2 Correlograph de Willshaw

El *correlograph* de *Willshaw* [22] es un dispositivo óptico, el cual puede funcionar como una memoria asociativa. Esta memoria asociativa funciona de la siguiente manera.

Fase de aprendizaje

La fase aprendizaje del *Correlograph* está conformada por dos pasos [22].

1. Se crea la memoria asociativa (matriz) $\boldsymbol{M}$ llena de valores igual a cero.
2. Posteriormente se actualiza de acuerdo con la siguiente expresión:

$$m_{ij} = \begin{cases} 1 & si\ y_i^{\mu} = 1 = x_j^{\mu} \\ valor\ anterior & en\ otro\ caso \end{cases} \tag{6}$$

Fase de recuperación

La fase de recuperación consta de presentarle a la memoria asociativa $\boldsymbol{M}$ previamente entrenada un vector de entrada $x^{\omega} \in A^n, A = \{0,1\}$. La forma en la que se le presenta dicho vector de entrada a la memoria asociativa es realizando el producto de la memoria (matriz) $\boldsymbol{M}$ por el vector x^{ω}. Posteriormente se realiza una operación de umbralado, acorde con la expresión que se muestra a continuación [22].

$$y_i^{\omega} = \begin{cases} 1 & si \sum_{j=1}^{n} m_{ji} \cdot x_j^{\omega} \geq u \\ 0 & en\ otro\ caso \end{cases} \tag{7}$$

Asimismo, u es el valor de umbral, el cual mencionan sus creadores que una estimación aproximada de su valor es: $\log_2 n$, donde n es igual a la dimensión de los patrones de entrada [22].

Se muestra a continuación un ejemplo ilustrativo del proceso operacional de la fase de aprendizaje y recuperación del *correlograph* de *Willshaw*. Los patrones y sus asociaciones correspondientes utilizadas para el ejemplo son los siguientes:

$$x^1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}\ y^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix};\ x^2 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} + y^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix};\ x^3 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}\ y^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix};\ x^4 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}\ y^4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Cómo se logra observar se tienen cuatro parejas de patrones de dimensión 7 para los patrones de entrada x^{μ} y de dimensión 4 para los patrones de salida y^{μ}; es decir, $p = 4$, $m = 4$ y $n = 7$. Después de aplicar el proceso de aprendizaje como se describió previamente en la expresión 10, la memoria entrenada resulta de la siguiente forma:

$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

El umbral por utilizar se calculó como $\log_2 n = \log_2 7 = 2.8$. Por otro lado, al realizar la fase recuperación conforme la expresión 11, para cada uno de los patrones de entrada se obtienen los siguientes resultados:

$$M \cdot x^1 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \therefore \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \neq y^1$$

$$M \cdot x^2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 4 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \therefore \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \neq y^2$$

$$M \cdot x^3 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ 7 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \therefore \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \neq y^3$$

$$M \cdot x^4 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \therefore \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \neq y^4$$

Se logra observar que en este caso particular ninguno de los patrones fue recuperado correctamente.

En este capítulo se describieron detalladamente dos pioneros modelos asociativos relevantes en la literatura, así como para el modelo propuesto en el presente trabajo de tesis. Asimismo, se analizaron y describieron los conjuntos de datos bajo estudio, al igual que la medida de desempeño y el método de validación a utilizar en la fase de experimentación descrita en el capítulo 5 del presente documento.

CAPÍTULO 4. PROPUESTA

En este capítulo se presenta y describe detalladamente el nuevo algoritmo de clasificación inteligente de patrones propuesto para el prediagnóstico médico de enfermedades respiratorias, el *Subtractive threshold associative classifier* (STAC). Se explican las ideas principales del algoritmo, así como su funcionamiento; también se aborda la fase de entrenamiento y, por último, se detalla la fase de clasificación del algoritmo STAC.

El nuevo algoritmo STAC está basado principalmente en la memoria asociativa Lernmatrix [21] acompañado de ideas relacionadas a la memoria asociativa Correlograph [22], los cuales fueron descritos en el capítulo 3 de materiales y métodos. El propósito de esta propuesta es mejorar el desempeño de los clasificadores del enfoque asociativo, en diferentes conjuntos de datos pertenecientes al tema del prediagnóstico de enfermedades respiratorias. Una de las bondades del algoritmo STAC es que puede lidiar con conjuntos de datos que presentan desbalance de clases, los cuales son muy comunes dentro del prediagnóstico de enfermedades, debido a que usualmente la cantidad de personas enfermas son mucho menor a la cantidad de personas sanas; es decir, el número de instancias en la clase mayoritaria es mayor que el número de instancias de la clase minoritaria (la clase de enfermos) [3], ayudando de esa forma a mejorar notablemente el diagnóstico médico. El problema del desbalance de clases en los conjuntos de datos es un reto para los algoritmos de clasificación, afectando severamente en su capacidad, lo cual se traduce a un mal diagnóstico. Esta complejidad en los datos es actualmente un tema muy atacado por los investigadores, como se puede ver en [49-53]. Más adelante en las secciones 4.3 y 4.4 se detallarán las fases de aprendizaje y clasificación del nuevo algoritmo STAC, respectivamente. A continuación, se muestra un diagrama general de la metodología utilizada para la propuesta de investigación.

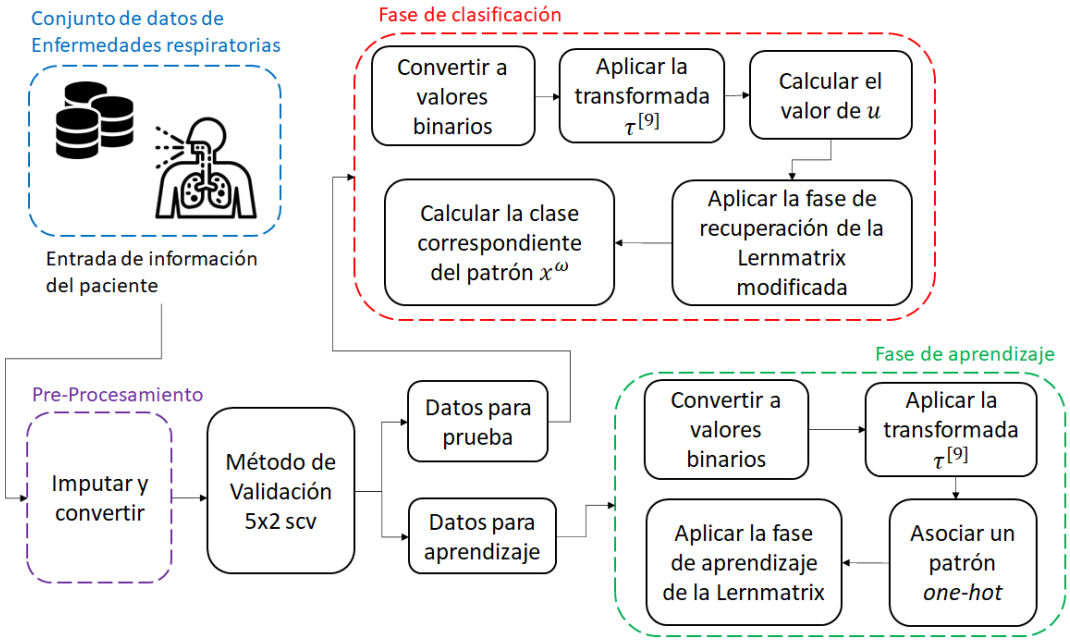


***Figura 1**. Diagrama de la metodología propuesta.*

Por otro lado, el nuevo algoritmo STAC hace uso de un codificador de valores reales a cadenas binarias, así como de una transformada matemática, debido a que el algoritmo propuesto trabaja únicamente con valores binarios. El método Johson-Möbius [54] y la transformada $\tau^{[9]}$ [43] son explicados más adelante.

4.1 Codificador Johnson-Möbius

El codificador Johnson-Möbius es usado para convertir números reales en un conjunto de cadenas binarias. El codificador traslada todos los valores del conjunto de datos con el propósito de eliminar los valores negativos, en este sentido, se realiza una suma del valor mínimo dentro de dicho conjunto; posteriormente, en caso de ser necesario, se fija una cantidad de decimales a tratar y, si se requiere, se truncan los decimales para que queden ajustados con la cantidad de decimales fijado; después, es requerido escalar todos los datos del conjunto con el objetivo de desaparecer dichos valores. Finalmente, para construir la cadena binaria, se toma el número máximo del conjunto como referencia para definir la longitud de la cadena binaria, donde cada número real es representado con tantos unos como indique su valor, antecedidos de una cadena de ceros hasta completar la longitud definida.

Como ejemplo ilustrativo, se tiene el siguiente conjunto de 6 números.

$$1.13, -0.01, 1.35, -0.11, 0.66, 1.41$$

El primer paso es eliminar los valores negativos, para ello se le suma a cada número el valor mínimo, en este caso es 0.11. Por lo tanto, el conjunto original es transformado en un conjunto de puros números reales positivos, como se muestra a continuación:

$$1.24, 0.10, 1.46, 0.0, 0.77, 1.52$$

El segundo paso es definir el número de decimales a usar, en este ejemplo se fijará a 1 decimal. Por lo tanto, el conjunto resulta de la siguiente forma:

$$1.2, 0.1, 1.4, 0.0, 0.7, 1.5$$

Posteriormente, se escalan los valores multiplicando cada número por 10 (debido a que se tiene solamente 1 decimal), para obtener únicamente valores enteros positivos. El conjunto resulta de la siguiente forma:

$$12, 1, 14, 0, 7, 15$$

Finalmente, para este ejemplo, el número máximo es igual a 15. Por lo tanto, al realizar las conversiones y concatenaciones de ceros y unos. En la **Tabla 2** se observa el resultado al aplicar el método Johnson-Möbius para convertir los valores reales del conjunto de datos en cadenas binarias, las cuales serán utilizadas más adelante.

Tabla 2. *Ejemplo ilustrativo del codificador Johnson-Möbius.*

Datos originales	Datos en enteros no negativos	Datos en cadenas binarias
1.13	12	000111111111111
- 0.01	1	000000000000001
1.35	14	011111111111111
- 0.11	0	000000000000000
0.66	7	000000001111111
1.41	15	111111111111111

4.2 Transformación de los patrones

Por otro lado, el nuevo algoritmo STAC aplica un proceso para transformar las cadenas binarias convertidas anteriormente (en la sección 4.1, utilizando el codificador Johnson-Möbius).

Esta transformación utiliza una simple pero poderosa transformada matemática, denominada por los autores, la transformada $\tau^{[9]}$ [43]. La $\tau^{[9]}$ transforma cada componente binario en una dupla de valores binarios, en base a la siguiente expresión:

$$\tau^{[9]}(1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\tau^{[9]}(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Por lo tanto, basándonos en el ejemplo ilustrativo reflejado en la **Tabla 2**, las cadenas binarias resultan de la siguiente forma:

Tabla 3. *Ejemplo ilustrativo de la transformada* $\tau^{[9]}$.

Datos en cadenas binarias	Aplicando la Transformada $\boldsymbol{\tau}^{[9]}$
000111111111111	01010110101010101010101010101010
000000000000001	01010101010101010101010101010110
011111111111111	01101010101010101010101010101010
000000000000000	01010101010101010101010101010101
000000001111111	01010101010101011010101010101010
111111111111111	10101010101010101010101010101010

Finalmente, la información resultante de la aplicación de la transformada $\tau^{[9]}$ es usado más adelante en los procesos de aprendizaje y clasificación del algoritmo STAC.

Antes de comenzar con las fases de aprendizaje y clasificación del algoritmo STAC, en caso de ser necesario, se aplica un pre-procesamiento en el conjunto de datos para tratar los valores perdidos y los valores mezclados.

4.3 Fase de aprendizaje del algoritmo STAC

La fase de aprendizaje de STAC consta de cinco pasos, a continuación, se muestra un diagrama general de su proceso operacional.

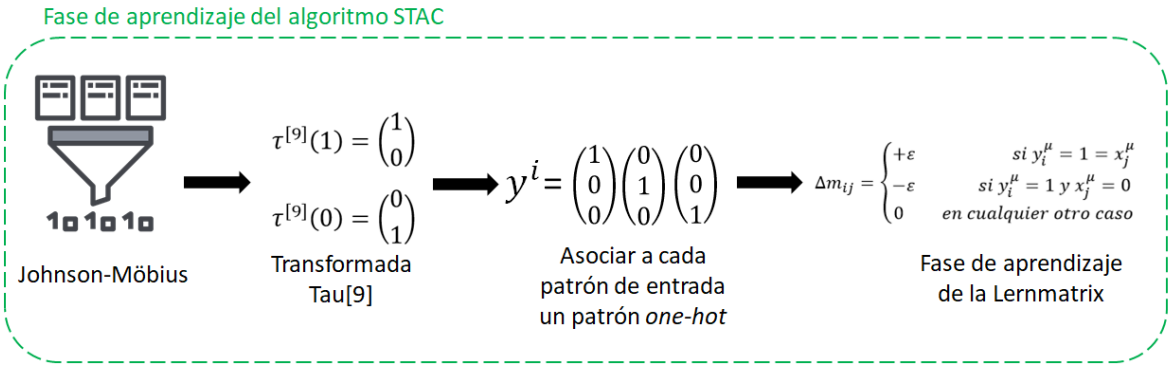


***Figura 2**. Diagrama del proceso de la fase de aprendizaje.*

A profundidad, el primer paso del algoritmo STAC consiste en convertir todos los patrones de entrada usados para el aprendizaje en cadenas binarias aplicando el codificador Johnson-Möbius (descrito anteriormente en este capítulo).

El segundo paso del algoritmo STAC consiste en aplicar la transformada $\tau^{[9]}$ (descrito en la sección 4.3) a todos los componentes de cada uno de los patrones de entrada usados para el aprendizaje y convertidos anteriormente en el paso uno. El tercer paso consiste en asociar a cada patrón de entrada (obtenido del paso anterior) un patrón de salida one-hot.

Y finalmente, en el cuarto paso la fase de aprendizaje de la Lernmatrix original (explicada detalladamente en el capítulo **3**) es aplicado para obtener la memoria $\boldsymbol{M}$. La fase de aprendizaje de la Lernmatrix se describe en dos pasos, las cuales son expresadas a continuación de forma breve:

1. Cada uno de los componente m_{ij} de la matriz $\boldsymbol{M}$ es inicializada en ceros.
2. Cada componente m_{ij} se actualiza de acuerdo con la regla $m_{ij} + \Delta m_{ij}$, donde:

$$\Delta m_{ij} = \begin{cases} +\varepsilon & si\ y_i^\mu = 1 = x_j^\mu \\ -\varepsilon & si\ y_i^\mu = 1\ y\ x_j^\mu = 0 \\ 0 & en\ cualquier\ otro\ caso \end{cases} \tag{8}$$

Donde cada ε representa cualquier constante positiva previamente seleccionada.

A continuación, se presenta un breve ejemplo del proceso de aprendizaje del algoritmo STAC descrito anteriormente.

Ejemplo del proceso de aprendizaje del algoritmo STAC

Se muestra a continuación un ejemplo ilustrativo del proceso operacional de la fase de aprendizaje de STAC. Los patrones utilizados para el ejemplo se detallan a continuación, donde x^1 y x^3 corresponden a la clase A, mientras que el patrón x^2 corresponde a la clase B.

$$x^1 = \begin{pmatrix} 0.28 \\ 0.17 \end{pmatrix};\ x^2 = \begin{pmatrix} 0.06 \\ -0.15 \end{pmatrix};\ x^3 = \begin{pmatrix} 0.11 \\ 0.03 \end{pmatrix}$$

Después de aplicar el paso 1 de la fase de aprendizaje, los patrones resultan de la siguiente forma:

$$x^1 = \begin{pmatrix} 1111 \\ 0111 \end{pmatrix};\ x^2 = \begin{pmatrix} 0011 \\ 0000 \end{pmatrix};\ x^3 = \begin{pmatrix} 0011 \\ 0001 \end{pmatrix}$$

Con el fin de mantener vectores columna, al concatenar las cadenas binarias obtenidas, resultan los siguientes patrones:

$$x^1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix};\ x^2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix};\ x^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Posteriormente, al ejecutar el paso 2 de la fase de aprendizaje del algoritmo STAC, el cual consiste en aplicar la transformada $\tau^{[9]}$ a cada uno de los componentes de todos los patrones. Por lo tanto, se obtienen los siguientes resultados:

$$\Gamma^{[9]}(x^1) = \begin{pmatrix} \tau^{[9]}(1) \\ \tau^{[9]}(1) \\ \tau^{[9]}(1) \\ \tau^{[9]}(1) \\ \tau^{[9]}(0) \\ \tau^{[9]}(1) \\ \tau^{[9]}(1) \\ \tau^{[9]}(1) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix};\ \Gamma^{[9]}(x^2) = \begin{pmatrix} \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(1) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix};\ \Gamma^{[9]}(x^3) = \begin{pmatrix} \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(1) \\ \tau^{[9]}(1) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(0) \\ \tau^{[9]}(1) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}$$

Después, al aplicar el paso 3 de la fase de aprendizaje, las asociaciones quedan de la siguiente forma:

$$x^1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \; y^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}; \; x^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \; y^2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}; \; x^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \; y^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Finalmente, en este caso, se tienen 3 asociaciones de patrones de dimensión 12, es decir $n = 12$ y $p = 3$. Por lo tanto, como se describió en el paso 4 de la fase de aprendizaje de STAC, primero se crea una matriz de dimensiones $p = 3 \times n = 12$ llena de ceros. La matriz inicial resulta de la siguiente forma:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Posteriormente, se actualiza la matriz creada acorde a la regla descrita en la ecuación **8**; finalmente, al sumar las tres asociaciones, se obtiene la siguiente memoria $\boldsymbol{M}$ entrenada:

$$M = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \end{pmatrix}$$

4.4 Fase de clasificación del algoritmo STAC

La fase de clasificación de STAC consta de cinco pasos, a continuación, se muestra un diagrama general de su proceso operacional.

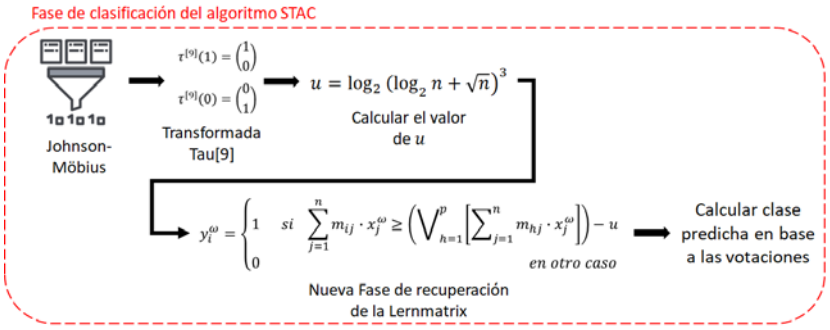


***Figura 3**. Diagrama del proceso de la fase de clasificación (recuperación).*

De forma más detallada, el primer paso de la fase de aprendizaje consiste en convertir el patrón desconocido x^{ω} en valores binarios, usando el codificador Johnson-Möbius (descrito en la sección 4.1). El segundo paso consiste en aplicar la transformada $\tau^{[9]}$ a todos los componentes del patrón desconocido x^{ω} convertido anteriormente en el paso 1 de la fase de clasificación.

En el tercer paso se obtiene el valor de u, el cual es calculado de la siguiente forma:

$$u = \log_2 (\log_2 n + \sqrt{n})^3$$

Donde n es igual a la dimensión de los patrones.

Posteriormente, en el cuarto paso, usando el valor de u calculado en el paso anterior, se ejecuta una modificación de la fase de recuperación de la Lernmatrix original, la cual combina la idea de utilizar un umbral para mejorar la recuperación, como hace uso la memoria asociativa Correlograph. El vector recuperado se consigue por medio de la siguiente expresión:

$$y_i^{\omega} = \begin{cases} 1 & si \quad \sum_{j=1}^{n} m_{ij} \cdot x_j^{\omega} \geq \left(\bigvee_{h=1}^{p} \left[\sum_{j=1}^{n} m_{hj} \cdot x_j^{\omega} \right] \right) - u \\ 0 & en\ otro\ caso \end{cases} \tag{9}$$

Finalmente, el quinto paso consiste en un sistema de votaciones ponderadas según las posiciones de cada clase correspondientes al patrón y^{ω} recuperado en el paso 4, para obtener la clase predicha del patrón de entrada desconocido x^{ω}. Lo anterior se expresa de la siguiente forma:

$$E_i = \frac{H_i}{K_i}$$

Donde H representa la suma de los componentes del patrón y^{ω} cuyas posiciones corresponden a la clase K_i. Asimismo, la clase K_i es asignada al patrón x^{ω} únicamente si se cumple la condición de que $E_i = \bigvee_{k=1}^{c} E_k$. c representa la cantidad de clases que conforman al conjunto de datos.

Ejemplo del proceso de clasificación del algoritmo STAC

Se muestra a continuación un breve ejemplo ilustrativo del proceso operacional de la fase de clasificación de STAC. Tomando en cuenta los resultados obtenidos en el ejemplo de la fase de aprendizaje. Se tiene una memoria $\boldsymbol{M}$ entrenada, la cual se expresa como:

$$M = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \end{pmatrix}$$

Previo a los pasos 1 y 2 de la fase de clasificación, se tiene un patrón desconocido x^{ω}:

$$x^{\omega} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

Posteriormente, se obtiene el valor de u, el cual es calculado de la siguiente forma:

$$u = \log_2 (\log_2 12 + \sqrt{12})^3 = 9$$

Después, aplicamos el paso 4 del algoritmo STAC con el propósito de recuperar su patrón, utilizando la expresión **9** descrita anteriormente. Por lo tanto, se obtiene el siguiente resultado:

$$M \cdot x^{\omega} = \begin{pmatrix} \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon \\ -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & -\varepsilon & \varepsilon & \varepsilon & -\varepsilon \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -4\varepsilon \\ 8\varepsilon \\ 4\varepsilon \end{pmatrix}$$

En este caso, el número mayor dentro del vector recuperado es igual a 8; por lo tanto, al aplicar la operación descrita en la ecuación **9**, $valor\ mayor\ del\ vector\ recuperado - u$, se obtiene que $8 - 9 = -1$. Entonces cada componente que sea mayor o igual a -1 tendrá un valor equivalente a 1. El nuevo patrón se convertiría como se muestra a continuación:

$$y^{\omega} = \begin{pmatrix} -4\varepsilon \\ 8\varepsilon \\ 4\varepsilon \end{pmatrix} \rightarrow \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

En consecuencia, como se indica en el paso 5 de la fase de clasificación, se realiza una votación con proporciones para conocer a que clase pertenece el patrón de entrada desconocido x^{ω}.

Para ello, se suman los componentes en las posiciones en cada clase correspondiente sobre el nuevo patrón y^{ω} convertido, para luego calcular su proporción. Por ejemplo, en este caso, la clase A se encuentran en la posición 1 y 3, y la clase B en la posición 2. Por lo tanto, en la clase A la suma de sus componentes da un resultado de 1, de modo que tiene una proporción de $\frac{1}{2} = 0.5$. Por otro lado, en la clase B, la suma de sus componentes da un resultado de 1, pero su proporción es de $\frac{1}{1} = 1$. Esto nos indica que como la proporción de la clase B es mayor a la de las demás clases (en este caso la clase A), se le asigna al patrón desconocido x^{ω} la clase B.

Para concluir el capítulo, el algoritmo STAC es un clasificador novedoso, el cual combina técnicas de dos memorias asociativas pioneras, además de ideas recientes. Por otra parte, dicho algoritmo es transparente en su funcionalidad, debido a que se permite conocer exactamente porque un patrón fue clasificado a una clase determinada. Esta característica del clasificador es una clara ventaja contra algunos otros clasificadores utilizados dentro del estado del arte.

CAPÍTULO 5. RESULTADOS Y DISCUSIÓN

Este capítulo se presentan el método de validación y la medida de desempeño utilizada para evaluar a los algoritmos en el desarrollo del presente trabajo. Por otro lado, se reportan los resultados experimentales obtenidos para el prediagnóstico médico de las enfermedades respiratorias más comunes, utilizando el nuevo algoritmo de clasificación propuesto, **STAC**, así como algunos de los algoritmos de clasificación más relevantes dentro del estado del arte.

5.1 Método de validación utilizado

En esta sección se describe el método de validación utilizado en la fase de experimentación que se describe en el presente capítulo.

Para obtener resultados confiables al momento de medir el desempeño en los diferentes clasificadores en la fase de experimentación es necesario anteriormente haber implementado un método de validación, el cual divide el conjunto de datos original en dos conjuntos: un conjunto de prueba *P* y un conjunto de entrenamiento *E*.

Existen muchas formas de definir estos conjuntos de datos, el más utilizado y recomendado por diversos autores es el método *k-fold cross-validation* [55, 56]. Sin embargo, debido a que los conjuntos de datos seleccionados en el presente trabajo de investigación presentan en su mayoría desbalance de clases, se decidió usar el método 5 x *stratified 2-fold cross-validation* (5x2 scv) [38, 57], el cual es ampliamente recomendado para conjuntos de datos desbalanceados, puesto que conservan aproximadamente el mismo porcentaje de patrones de cada clase para cada uno de los pliegues (*fold*). De esa forma, los conjuntos de prueba son lo más representativos posibles del conjunto de datos original, evitando el sesgo entre clases.

A continuación, se muestra una figura que describe un breve ejemplo del funcionamiento del método antes mencionado.

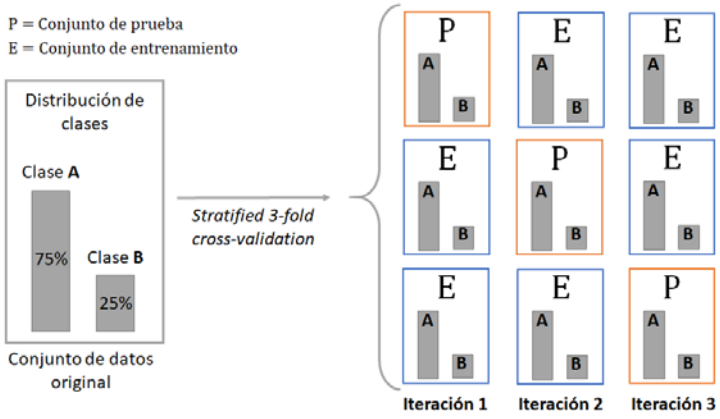


Figura 4. *Ejemplo del método stratified k-fold cross-validation con k=3.*

5.2 Medida de desempeño utilizada

En esta sección se presenta la medida de desempeño aplicada a los diferentes algoritmos de clasificación implementados con el fin de compararlos adecuadamente en la fase experimental descrito en el presente capítulo.

Dentro de la literatura relacionada al área de clasificación inteligente de patrones, existen diversas formas de medir el desempeño de un clasificador aplicado. Estas medidas se originan de los resultados obtenidos después de un experimento realizado, los cuales son representados en una matriz llamada "Matriz de confusión".

En el caso de un problema de clasificación de dos clases, la matriz de confusión contiene cuatro valores, como se ilustra en la **Figura 5**.

		Clases predichas	
		Positivos	**Negativos**
Clases reales	**Positivos**	TP	FN
	Negativos	FP	TN

Figura 5. *Ilustración de una matriz de confusión para un problema de dos clases.*

Los resultados obtenidos por la matriz de confusión son los siguientes:

- TP (*True Positive*): Este valor representa la cantidad de patrones que fueron clasificados como positivos y que pertenecen realmente a la clase positiva.
- TN (*True Negative*): Este valor representa la cantidad de patrones pertenecientes a la clase negativa y que fueron clasificados como negativos.
- FP (*False Positive*): Este valor representa la cantidad de patrones que fueron clasificados como positivos, pero pertenecen realmente a la clase negativa.
- FN (*False Negative*): Este valor representa la cantidad de patrones pertenecientes a la clase positiva, pero fueron clasificados como negativos.

Se puede observar que los valores TP (*True Positive*) y TN (*True Negative*) equivalen a los patrones que fueron clasificados correctamente, mientras que los valores de FP (*False Positive*) y FN (*False Negative*) equivalen a los patrones que fueron clasificados incorrectamente.

Los valores resultantes de una matriz de confusión (como se muestra en la **Figura 5**), son utilizados para obtener diferentes medidas de desempeño. Una de las métricas más comunes para medir el desempeño de los clasificadores es el valor de *Accuracy*, el cual se obtiene calculando el porcentaje de patrones correctamente clasificados dentro de todos los patrones existentes del conjunto de datos original. Lo antes mencionado se describe en la ecuación **10**.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{10}$$

Sin embargo, esta medida de desempeño no es adecuada para conjuntos de datos desbalanceados, debido a que existe un gran sesgo hacia el número de patrones correctamente clasificados de las clases mayoritarias, provocando resultados que no representan verdaderamente la capacidad de los clasificadores. Por lo tanto, es necesario aplicar una medida de desempeño recomendada para este tipo de problemas. Para ello existe la medida de desempeño Balanced Accuracy (BA) [58], la cual es construida a partir de otras dos medidas de desempeño, *Sensitivity* y *Specificity*. BA se calcula como se muestra en la ecuación **11** [58].

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

$$BA = \frac{Sensitivity + Specificity}{2}$$

A continuación, se desarrolla un ejemplo para ilustrar la diferencia significativa entre las medidas de desempeño antes mencionadas para conjuntos de datos desbalanceados. Ver **Figura 6**.

		Clases predichas	
		Positivos	Negativos
Clases reales	Positivos	1436	13
	Negativos	16	19

Figura 6. *Ejemplo de una matriz de confusión para un problema de dos clases.*

El valor de *Accuracy* de la matriz de confusión de la **Figura 6** es la siguiente:

$$Accuracy = \frac{1436 + 19}{1436 + 13 + 16 + 19} = \frac{1455}{1484} = 0.980 \tag{12}$$

Sin embargo, el valor de BA de la matriz de confusión de la **Figura 6** es la siguiente:

$$Sensitivity = \frac{1436}{1436 + 13} = 0.991$$

$$Specificity = \frac{19}{19 + 16} = 0.542 \tag{13}$$

$$BA = \frac{0.991 + 0.542}{2} = 0.766$$

Como se puede observar en la ecuación **12**, la medida de desempeño *Accuracy* obtuvo un valor de 98%, lo cual podría ser considerado como un buen desempeño, debido a que el resultado obtenido refleja que se logró clasificar correctamente 98% de los patrones del conjunto de datos; es decir, se obtuvo únicamente un 2% de error. Sin embargo, al aplicar la medida de desempeño *Balanced Accuracy* (ecuación **13**) se obtiene un valor de 76.6%, el cual manifiesta un mayor error de clasificación que el resultado obtenido por la medida *Accuracy*, dado que el conjunto de datos se encuentra severamente desbalanceado y la medida BA disminuye el sesgo entre la clase minoritaria y la clase mayoritaria, obteniendo de esa forma un resultado que refleje la verdadera capacidad del clasificador [59].

Este fenómeno sucede debido a que el sesgo de las distribuciones de clases es notablemente severo, provocando resultados que no reflejan un desempeño real. De modo que la medida de desempeño *Accuracy* se convierte en un valor poco confiable sobre el rendimiento en los algoritmos de clasificación aplicados a conjuntos de datos desbalanceados.

Por otro lado, dentro de los conjuntos de datos seleccionados para la fase de experimentación existe la presencia de conjuntos con más de dos clases. Por lo tanto, la medida de desempeño *Balanced Accuracy* (BA) para *k* clases se calcula de la siguiente manera [58]:

$$BA = \frac{1}{k}\sum_{i=1}^{k}\frac{T_i}{N_i} \tag{14}$$

Donde, T_i representa el número de patrones clasificados correctamente de cada clase i, y N_i representa el total de patrones pertenecientes a cada clase i, en un total de k clases.

Debido a que los conjuntos de datos seleccionados presentan en su mayoría desbalance de clases, se optó por utilizar la medida de desempeño antes descrita, *Balanced Accuracy* (BA).

5.3 Pruebas estadísticas

Los distintos algoritmos de clasificación encontrados en el enfoque del aprendizaje automático y la selección de un algoritmo final como ganador, es una práctica común en el campo de la investigación y la aplicación de modelos de aprendizaje automático. Los modelos implementados para la fase de experimentación se evalúan utilizando un determinado método de validación y comparando directamente los valores obtenidos mediante el cálculo de una medida de desempeño. Aunque se trata de un enfoque sencillo e intuitivo, es complejo determinar la diferencia entre una capacidad real del algoritmo o simplemente una casualidad estadística.

Por lo tanto, para abordar ese problema y determinar qué algoritmo de clasificación obtuvieron los mejores resultados, es necesario utilizar pruebas de hipótesis. Las pruebas de hipótesis permiten encontrar si existe o no la presencia de una diferencia significativa en el desempeño de los diferentes algoritmos de clasificación utilizados. Para el presente trabajo de tesis se seleccionó la prueba no paramétrica de Friedman, debido a que es apropiada para comparar múltiples muestras [60].

La prueba de Friedman [61, 62] consiste en clasificar las muestras según su valor de rango, siendo el menor rango la muestra con el mejor desempeño y el mayor rango la muestra con el peor desempeño obtenido, promediando rangos para muestras idénticas.

En la prueba de Friedman la hipótesis nula plantea que los desempeños obtenidos son similares (no existen diferentes entre ellos). Por lo tanto, la prueba rechaza la hipótesis nula si su valor de probabilidad *p-value* es inferior o igual al nivel de significación establecido. En esta investigación se considera un nivel de significación de $\alpha = 0.05$, para un 95% de confianza. El rechazo de la hipótesis nula indica la existencia de diferencias significativas entre los distintos algoritmos de clasificación.

Por otro lado, si la prueba de Friedman rechaza la hipótesis nula, se sugiere aplicar una prueba *post-hoc* para determinar en qué algoritmos existen diferencias significativas [63]. Para esta investigación se seleccionó la prueba de Holm [64] debido a que es recomendada para el análisis de múltiples conjuntos de datos [60, 63].

La prueba de Holm ajusta el valor de significancia α. El procedimiento de la prueba consiste en ordenar los valores de probabilidad *p-value* de forma ascendente. Posteriormente, para rechazar todas las hipótesis nulas asociadas a los *p-value* se compara dicho valor con el resultado de la división entre el nivel de significación y el número total de hipótesis donde el *p-value* no ha sido comparado; es decir, si $P_n < \frac{\alpha}{l-n}$, la prueba rechaza la hipótesis nula. En esta investigación se utilizó el software KEEL [38] para el cálculo de la prueba de Friedman y la prueba Holm.

5.4 Resultados experimentales de clasificación

En tabla **1** se muestra el desempeño obtenido por el algoritmo propuesto así como su comparación con los clasificadores más relevantes del estado del arte usando diferentes conjuntos de datos referentes a las enfermedades respiratorias más comunes [5]. Los resultados que alcanzaron el mayor desempeño se encuentran resaltados en negritas.

El algoritmo propuesto fue implementado y ejecutado usando el software Matlab 2021. Por otro lado, los demás algoritmos fueron ejecutados en el software WEKA en la versión 3.8, utilizando los parámetros predeterminados ofrecidos por el mismo software. Para el algoritmo kNN se utilizó la función de distancia euclidiana. Asimismo, para el algoritmo SVM se utilizó el algoritmo de optimización SMO.

Se puede observar en la **Tabla 4** los resultados obtenidos por los diferentes clasificadores aplicados a los doce conjuntos de datos propuestos para la presente investigación. De forma clara se observa que el algoritmo propuesto STACT logró un alto rendimiento, obteniendo el mejor valor de BA en seis de los doce conjuntos de datos utilizados. Como por ejemplo en los conjuntos de datos: *PostOperative*, *Lung-Cancer*, *ACPs Lung Cancer*, *Lymphography*, *Lymphography-NF* y *Primary-tumor*.

Tabla 4. *Resultados obtenidos por los diferentes clasificadores de acuerdo con la medida BA.*

Dataset	Naïve Bayes	3-NN	MLP	SVM	C4.5	STAC
PostOperative	0.321	0.311	0.295	0.328	0.327	**0.352**
Thyroid	0.726	0.548	0.784	0.496	**0.983**	0.724
Newt-thyroid1	**0.988**	0.913	0.965	0.745	0.926	0.960
Newt-thyroid2	**0.989**	0.909	0.966	0.757	0.904	0.969
Thoracic-Surgery	**0.578**	0.508	0.523	0.500	0.511	0.508
Lung-Cancer	0.569	0.513	0.513	0.506	0.492	**0.594**
Survey Lung-Cancer	0.716	0.711	**0.779**	0.774	0.666	0.761
ACPs Lung Cancer	0.634	0.610	0.635	0.681	0.250	**0.949**
Exasens-COPD	0.875	0.852	0.885	0.820	**0.910**	0.902
Lymphography	0.578	0.434	0.491	0.641	0.582	**0.867**
Lymphography-NF	0.747	0.498	0.793	0.698	0.598	**0.965**
Primary-tumor	0.271	0.230	0.234	0.252	0.233	**0.340**
Veces en las que fue el mejor	3	0	1	0	2	**6**

El resultado con el menor desempeño fue obtenido por el conjunto de datos *Primary-tumor*; sin embargo, el modelo propuesto destaca siendo el mejor valor de BA obtenido de entre todos los demás modelos, con un valor de 0.340. Otros conjuntos de datos complicados para esta tarea de clasificación son: *Thoracic-Surgery*, *Lung-Cancer* y *PostOperative*, de los cuales los desempeños más altos fueron 0.578, 0.594 y 0.352, respectivamente. Donde únicamente uno de estos conjuntos de datos (Thoracic-Surgery) no fue obtenido por el modelo propuesto STAC.

A causa del Teorema de No Free Lunch, el modelo propuesto STAC no fue el clasificador con el mejor desempeño en todos los conjuntos de datos. Esto debido a que dicho teorema indica que no existe ningún clasificador que sea capaz de ser el mejor en todo tipo de problemas [19].

Sin embargo, a favor de nuestra propuesta, se puede señalar que, en la mayoría de los casos, los valores de desempeños conseguidos por el clasificador propuesto STAC se encuentran cercanos a los desempeños más altos obtenidos por los demás clasificadores, tal es el caso del conjunto de datos Survey Lung-Cancer, donde el algoritmo STAC obtuvo un resultado de 0.761, el cual no se encuentran tan alejado del mejor desempeño, con un valor de 0.779 alcanzado por el clasificador MLP. Otros casos similares ocurren en los conjuntos de datos Newt-thyroid1, Newt-thyroid2 y Exasens, donde el modelo propuesto resulto con valores de desempeños iguales a 0.960, 0.969 y 0.902, respectivamente, muy similares a los mejores desempeños obtenidos por los demás clasificadores. Para realizar un análisis comparativo con mayor fiabilidad en los resultados, se utilizó la prueba de Friedman [62] (el cual fue descrito en la sección 5.3 del presente capítulo), para demostrar la existencia de diferencias significativas en los desempeños observados.

En la **Tabla 5** se observa el *ranking* obtenido por la prueba de Friedman de acuerdo con los diferentes algoritmos de clasificación presentados. El modelo propuesto STAC ocupa el primer lugar en el *ranking*, con un valor de 2.0417, con respecto a los 5 algoritmos restantes, convirtiéndolo en el mejor modelo para la tarea de clasificación descrita en el presente documento de tesis. Por otro lado, el algoritmo que ocupa el último lugar en la tabla del *ranking* de la prueba de Friedman es el algoritmo 3-NN, con un valor de 5.

Tabla 5. *Resultados de la prueba de Friedman.*

Algoritmo	Ranking[1]
STAC	2.0417
Naïve Bayes	2.7500
MLP	3.0417
C4.5	4.0000
SVM	4.1667
3-NN	5.0000

[1] ordenados del mejor al peor.

Después de realizar la prueba de Friedman, la hipótesis nula es rechazada con un valor de confianza del 95% y un valor de probabilidad de $p = 0.001231$, el cual se encuentra en gran medida por debajo del nivel de significación establecido para esta investigación, el cual es de $\alpha = 0.05$. Por lo tanto, se demuestra la existencia de diferencias significativas entre los distintos algoritmos de clasificación. Debido a los resultados de la prueba de Friedman, se aplicó una prueba de *post-hoc*, la prueba Holm [64]. Los resultados se pueden observar en la **Tabla 6**. La prueba rechaza la hipótesis con un valor de p ajustado menor o igual a 0.025. Por lo tanto, se observa que existen diferencias significativas entre los desempeños obtenidos por el algoritmo propuesto STAC y los clasificadores: 3NN, SVM y C4.5.

Tabla 6. *Comparación post-hoc obtenida por la prueba Holm.*

i.	Algoritmo	z	p	Holm
5	3NN	3.8733	0.0001	0.0100
4	SVM	2.7822	0.0053	0.0125
3	C4.5	2.5640	0.0103	0.0166
2	MLP	1.3093	0.1904	0.0250
1	Naïve Bayes	0.9274	0.3537	0.0500

Después de realizar los experimentos descritos en el presente capítulo se observa que el modelo propuesto STAC destaco con competitivos resultados; debido a las diferencias significativas entre los desempeños conseguidos por el algoritmo, obteniéndolas en tres de los cinco algoritmos usados en el estado del arte bajo la misma tarea de clasificación. Por lo tanto, los resultados obtenidos respaldan la afirmación de que la propuesta del nuevo modelo STAC expuesto en el presente trabajo de tesis es adecuado para el prediagnóstico de las enfermedades respiratorias más comunes.

CAPÍTULO 6. CONCLUSIONES Y TRABAJO FUTURO

En este capítulo se presentan las conclusiones generales de este trabajo de tesis, así como los posibles trabajos futuros que podrían surgir a partir de esta investigación.

6.1 Conclusiones

En el presente trabajo de tesis, se propuso y presentó un nuevo algoritmo asociativo de clasificación inteligente de patrones, STAC (Subtractive Threshold Associative Classifier), diseñado para el prediagnóstico de enfermedades respiratorias. El algoritmo STAC es capaz de lidiar sobresalientemente con el desbalance de clases, una complejidad en los datos que aparece con mucha frecuencia en datos relacionados a la medicina y el prediagnostico de enfermedades.

Asimismo, otra ventaja del clasificador STAC, es que es un modelo interpretable; es decir, el algoritmo STAC es transparente en su proceso de clasificación, ya que se conoce exactamente la razón por la que un patrón es clasificado a una determinada clase.

Los resultados experimentales realizados en el capítulo 5, señalan la sobresaliente capacidad del algoritmo propuesto STAC, debido a que superan a varios de los algoritmos de clasificación más utilizados en el estado del arte referente al prediagnóstico de enfermedades respiratorias; sobresaliendo exactamente en seis de los 12 conjuntos de datos utilizados en la fase experimental.

Además, de acuerdo con la prueba de Friedman, el mejor clasificador en los experimentos ejecutados fue el algoritmo STAC, indicando la presencia de diferencias significativas, con un valor de probabilidad de $p = 0.001230$; asimismo, la prueba de Holm *post-hoc* refleja que existe de igual forma la presencia de diferencias significativas en el desempeño obtenido por el algoritmo propuesto y los demás clasificadores.

En ese contexto, lo mencionado anteriormente comprueba que el nuevo algoritmo STAC es adecuado para el prediagnóstico de enfermedades respiratorias, con resultados competitivos e incluso superiores a los clasificadores más utilizados en el estado del arte, tales como: Naïve Bayes, kNN, MLP, SVM, y C4.5.

6.2 Trabajo futuro

En los trabajos futuros se tendrá la intención de aplicar el nuevo algoritmo STAC sobre conjuntos de datos con diferentes enfoques, con el objetivo de evaluar su desempeño y comportamiento en este tipo de aplicaciones no médicas; asimismo, se propone comparar el algoritmo STAC con más clasificadores del estado del arte, incluyendo los clasificadores novedosos que surjan en la literatura.

Por otro lado, se propone utilizar técnicas de selección o extracción de características, combinado con técnicas de algoritmos evolutivos para reducir la dimensionalidad de los datos, y de ese modo, disminuir el consumo computacional del algoritmo. Además, se propone el uso del algoritmo STAC para trabajar y clasificar imágenes médicas, aplicando previamente técnicas de pre-procesamiento de imágenes.

Por último, se propone considerar problemas de clasificación multi-etiqueta.

Referencias

[1] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," *Expert Systems with Applications,* vol. 67, pp. 239-251, 2017.

[2] S. Woloshin, N. Patel, and A. S. Kesselheim, "False negative tests for SARS-CoV-2 infection—challenges and implications," *New England Journal of Medicine,* vol. 383, no. 6, p. e38, 2020.

[3] A. Lindberg, "Developing theory through integrating human and machine pattern recognition," *Journal of the Association for Information Systems,* vol. 21, no. 1, p. 7, 2020.

[4] National Cancer Institute. (2022, Marzo). *Respiratory disease*. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/respiratory-disease

[5] Forum of International Respiratory Societies, The global impact of respiratory disease, Third Edition ed.: European Respiratory Society, 2021. [Online]. Available: firsnet.org/images/publications/FIRS_Master_09202021.pdf.

[6] N. Sengupta, M. Sahidullah, and G. Saha, "Lung sound classification using cepstral-based statistical features," *Computers in biology medicine,* vol. 75, pp. 118-129, 2016.

[7] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *Pattern Analysis and Applications,* pp. 1-14, 2021.

[8] World Health Organization. (Marzo, 2022). *The top 10 causes of death*. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[9] World Health Organization. (Marzo, 2022). *WHO coronavirus (COVID-19) dashboard*. Available: https://covid19.who.int/

[10] Johns Hopkins Coronavirus Resource Center. (Marzo, 2022). *COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*. Available: https://coronavirus.jhu.edu/map.html

[11] Instituto Nacional de Estadística y Geografía (INEGI), "Estadística de defunciones registradas de Enero a Junio de 2021 (Preliminar)," 2022, Available: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2022/dr/dr2021.pdf.

[12] K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Computerized medical imaging graphics,* vol. 31, no. 4-5, pp. 198-211, 2007.

[13] J. E. Luján-García, C. Yáñez-Márquez, Y. Villuendas-Rey, and O. Camacho-Nieto, "A transfer learning method for pneumonia classification and visualization," *Applied Sciences,* vol. 10, no. 8, p. 2908, 2020.

[14] J. M. Marimón and J. M. Navarro-Marí, "Métodos de diagnóstico rápido de las infecciones respiratorias," *Enfermedades Infecciosas y Microbiología Clínica,* vol. 35, no. 2, pp. 108-115, 2017.

[15] J. V. Estapé *et al.*, "Métodos moleculares de diagnóstico de infecciones respiratorias.¿ Ha cambiado el esquema diagnóstico?," *Enfermedades Infecciosas y Microbiología Clínica,* vol. 34, pp. 40-46, 2016.

[16] L. Pham, I. McLoughlin, H. Phan, M. Tran, T. Nguyen, and R. Palaniappan, "Robust deep learning framework for predicting respiratory anomalies and diseases," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 164-167: IEEE.

[17] J. E. Luján-García, M. A. Moreno-Ibarra, Y. Villuendas-Rey, and C. Yáñez-Márquez, "Fast COVID-19 and Pneumonia Classification Using Chest X-ray Images," *Mathematics,* vol. 8, no. 9, p. 1423, 2020.
[18] D. Spathis and P. Vlamos, "Diagnosing asthma and chronic obstructive pulmonary disease with machine learning," *Health informatics journal,* vol. 25, no. 3, pp. 811-827, 2019.
[19] 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.
[20] C. Yáñez-Márquez, "Toward the bleaching of the black boxes: minimalist machine learning," *IT Professional,* vol. 22, no. 4, pp. 51-56, 2020.
[21] K. Steinbuch, "Die lernmatrix," *Kybernetik,* vol. 1, no. 1, pp. 36-45, 1961.
[22] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Non-holographic associative memory," *Nature,* vol. 222, no. 5197, pp. 960-962, 1969.
[23] World Health Organization. (Marzo, 2022). *Air pollution. 2021*. Available: https://www.who.int/health-topics/air-pollution
[24] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Systems with Applications,* vol. 164, p. 113981, 2021.
[25] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," presented at the Eleventh Conference on Uncertainty in Artificial Intelligence, 2013.
[26] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.
[27] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.
[28] B. Widrow and M. A. Lehr, "30 years of adaptive neural networks: perceptron, madaline, and backpropagation," *Proceedings of the IEEE,* vol. 78, no. 9, pp. 1415-1442, 1990.
[29] I. Cardoso *et al.*, "Analysis of machine learning algorithms for diagnosis of diffuse lung diseases," *Methods of information in medicine,* vol. 57, no. 05/06, pp. 272-279, 2018.
[30] J. Finkelstein and I. C. Jeong, "Machine learning approaches to personalize early prediction of asthma exacerbations," *Annals of the New York Academy of Sciences,* vol. 1387, no. 1, pp. 153-165, 2017.
[31] J. L. Amaral, A. J. Lopes, J. Veiga, A. C. Faria, and P. L. Melo, "High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements," *Computer methods programs in biomedicine,* vol. 144, pp. 113-125, 2017.
[32] P. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1-4: IEEE.
[33] S. Gonem, W. Janssens, N. Das, and M. Topalovic, "Applications of artificial intelligence and machine learning in respiratory medicine," *Thorax,* vol. 75, no. 8, pp. 695-701, 2020.
[34] Y. Xiong, X. Ba, A. Hou, K. Zhang, L. Chen, and T. Li, "Automatic detection of mycobacterium tuberculosis using artificial intelligence," *Journal of thoracic disease,* vol. 10, no. 3, p. 1936, 2018.
[35] A. Christe *et al.*, "Computer-aided diagnosis of pulmonary fibrosis using deep learning and CT images," *Investigative radiology,* vol. 54, no. 10, p. 627, 2019.
[36] G. Soldati, M. Demi, A. Smargiassi, R. Inchingolo, and L. Demi, "The role of ultrasound lung artifacts in the diagnosis of respiratory diseases," *Expert review of respiratory medicine,* vol. 13, no. 2, pp. 163-172, 2019.
[37] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," *BMC bioinformatics,* vol. 15, no. 1, pp. 1-8, 2014.

[38] J. Alcalá-Fdez *et al.*, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," *Multiple-Valued Logic Soft Computing,* vol. 17, 2011.

[39] D. Dua and C. Graff, *UCI Machine Learning Repository*. University of California, Irvine, School of Information.

[40] A. Fernández, S. García, M. J. del Jesus, and F. Herrera, "A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets," *Fuzzy Sets Systems,* vol. 159, no. 18, pp. 2378-2398, 2008.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, no. 1, pp. 10-18, 2009.

[42] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast tumor classification using an ensemble machine learning method," *Imaging,* vol. 6, no. 6, p. 39, 2020.

[43] J.-L. Velazquez-Rodriguez, Y. Villuendas-Rey, O. Camacho-Nieto, and C. Yáñez-Márquez, "A novel and simple mathematical transform improves the perfomance of lernmatrix in pattern classification," *Mathematics,* vol. 8, no. 5, p. 732, 2020.

[44] C. A. Rolón-González, R. Castañón-Méndez, A. Alarcón-Paredes, I. López-Yáñez, and C. Yáñez-Márquez, "Improving the Performance of an Associative Classifier in the Context of Class-Imbalanced Classification," *Electronics,* vol. 10, no. 9, p. 1095, 2021.

[45] L. F. S. Hoffmann, F. C. P. Bizarria, and J. W. P. Bizarria, "Detection of liner surface defects in solid rocket motors using multilayer perceptron neural networks," *Polymer Testing,* vol. 88, p. 106559, 2020.

[46] B. Hssina, A. Merbouha, H. Ezzikouri, and M. Erritali, "A comparative study of decision tree ID3 and C4. 5," *International Journal of Advanced Computer Science Applications,* vol. 4, no. 2, pp. 13-19, 2014.

[47] C. Yáñez-Márquez, I. López-Yáñez, M. Aldape-Pérez, O. Camacho-Nieto, A. J. Argüelles-Cruz, and Y. Villuendas-Rey, "Theoretical foundations for the alpha-beta associative memories: 10 years of derived extensions, models, and applications," *Neural Processing Letterss,* vol. 48 no. 2, pp. 811-847, 2018.

[48] F. A. Sánchez Garfias, J. L. Díaz de León Santiago, and C. Yáñez Márquez, "Analisis experimental de las condiciones suficientes para recuperacion perfecta en la Lernmatrix," *Serie Azul, IT-172,* 2003.

[49] M. M. Ahsan and Z. Siddique, "Machine learning-based heart disease diagnosis: A systematic literature review," *Artificial Intelligence in Medicine,* p. 102289, 2022.

[50] J. Brownlee, *Imbalanced classification with python: Better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery, 2020.

[51] H. Ibrahim, S. Anwar, and M. Ahmad, "Classification of imbalanced data using support vector machine and rough set theory: A review," in *Journal of Physics: Conference Series*, 2021, vol. 1878, no. 1, p. 012054: IOP Publishing.

[52] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys,* vol. 52, no. 4, pp. 1-36, 2019.

[53] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access,* vol. 7, pp. 93010-93022, 2019.

[54] K. S. Papadomanolakis, A. P. Kakarountas, N. Sklavos, and C. E. Goutis, "A Fast Johnson-Mobius Encoding Scheme for Fault Secure Binary Counters," *In Proceedings of the Design, Automation and Test in Europe,* pp. 4-8, 2002.

[55] R. T. Nakatsu, "An evaluation of four resampling methods used in machine learning classification," *IEEE Intelligent Systems,* vol. 36, no. 3, pp. 51-57, 2020.

[56] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Third Edition ed. (Elsevier Inc). Elsevier Inc, 2011.

[57] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation,* vol. 10, no. 7, pp. 1895-1923, 1998.

[58] V. García, R. A. Mollineda, and J. S. Sánchez, "Theoretical analysis of a performance measure for imbalanced data," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 617-620: IEEE.

[59] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information sciences,* vol. 250, pp. 113-141, 2013.

[60] S. Garcia and F. Herrera, "An Extension on Statistical Comparisons of Classifiers over Multiple Data Sets for all Pairwise Comparisons," *Journal of machine learning research,* vol. 9, no. 12, 2008.

[61] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics,* vol. 11, no. 1, pp. 86-92, 1940.

[62] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the american statistical association,* vol. 32, no. 200, pp. 675-701, 1937.

[63] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research,* vol. 7, pp. 1-30, 2006.

[64] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics,* pp. 65-70, 1979.