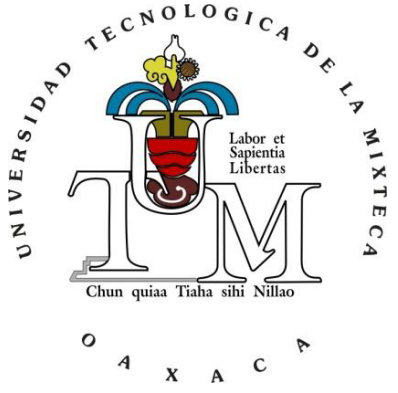


Universidad Tecnológica de la Mixteca

Diseño de una red neuronal para su aplicación en el monitoreo en tiempo real en una cinética microbiana usando espectroscopia Raman

Tesis:

Para obtener el título de

Ingeniero en Física Aplicada

Presenta:

Juan Ramón Velasco Reyes

Director:

Dr. José Aníbal Arias Aguilar

Co–director:

Dr. Maxvell Gustavo Jiménez Escamilla

Huajuapan de León, Oaxaca, México, Octubre, 2024

Dedicatoria

A mi madre, Elvira Reyes García, y a mi padre, Ramón Velasco Méndez, por creer en mí, por sacrificarse por mi educación y por su apoyo constante. Su dedicación me ha permitido alcanzar mis metas.

A mi hermano, José Ramón Velasco Reyes, por su compañía en los momentos difíciles y quien me inspira a ser un ejemplo a seguir.

A mi gran amigo, Luis Manuel Leyva Sanches quien era una persona con muchos sueños y propósitos, y vio en mi a un amigo, compañero y maestro. Su determinación me inspira en cada momento de mi vida. Un profundo y sincero agradecimiento en donde quiera que se encuentre.

A mis tres mejores amigos, Víctor Jerónimo Porras, Sergio Adrián Cruz Hernández y Jonathan Said Unda López, por hacer mi estancia en la universidad, tan divertida y memorable. Gracias por los momentos compartidos y por su amistad incondicional.

A mi novia y mejor amiga, Jimena Alejandra León Álvarez por su apoyo incondicional. Su presencia en mi vida ha sido una fuente constante de inspiración y motivación.

A todos ustedes, les dedico este logro.

Agradecimientos

A las personas de Labcitec, por darme la oportunidad de colaborar con este proyecto y proporcionarme los datos necesarios para la realización de este trabajo.

A mi director, el Dr. José Aníbal Arias Aguilar, por todo el apoyo en temas complicados y por proporcionarme material de estudio que me permitió culminar el proyecto.

A mi codirector, el Dr. Maxvell Gustavo Jiménez Escamilla, por todo el apoyo en los temas de biología que fueron claves en este trabajo.

A mis revisores de tesis, la Dra. Norma Francenia Santos Sánchez, la Dra. Yesica Espinosa Cerón, y el Dr. Ignacio Arroyo Fernández, por las observaciones y correcciones realizadas sobre mi trabajo de tesis.

Su colaboración ha sido esencial para el éxito de este proyecto.

Resumen

La espectroscopia Raman es una técnica de análisis que ha sido implementada para el monitoreo de procesos biológicos. A partir de modelos de regresión permite visualizar en tiempo real el cambio en la concentración de diferentes compuestos. Por otra parte, los modelos de redes neuronales de memoria a corto y largo plazo (LSTM) permiten realizar predicciones de series temporales a partir de uno o más valores analizados. A partir de esto se plantea la posibilidad de implementar este tipo de modelos al monitoreo de procesos biológicos en conjunto con los modelos de regresión, con el objetivo de realizar predicciones sobre el cambio en la concentración de los compuestos de interés. En este proyecto se configuró un modelo LSTM con 2 capas y 80 unidades neuronales que permite realizar hasta 12 horas de predicciones a partir de 12 horas de información previa sobre la concentración de la fuente de carbono, la biomasa y del ácido orgánico de interés. Utilizando el error cuadrático medio (MSE) como función de costo se logró obtener un error de entrenamiento, validación y prueba de 5.01E-04, 5.05E-04 y 5.05E-04 respectivamente. Para observar el desempeño del modelo LSTM se realizaron 32 predicciones en intervalos de una hora, utilizando los datos de una cinética microbiana monitoreada mediante espectroscopia Raman. Al comparar los datos de las predicciones con los datos del modelo de regresión se obtuvo un valor de error absoluto medio (MAE) mínimo de 1.194 (2.9 %) y un valor MAE máximo de 5.8999 (18.7 %); al compararlos con los datos de laboratorio se obtuvo un valor MAE mínimo de 0.5235 (1.5 %) y un valor MAE máximo de 14.4597 (19.3 %); y por último se realizó la comparación con un modelo logístico generado a partir de los datos de laboratorio en cada una de las predicciones, obteniendo un valor MAE mínimo de 4.2642 (6.2 %) y un valor MAE máximo de 55.8067 (88.9 %). Se observó que el desempeño del modelo LSTM depende ampliamente del modelo de regresión que se esté utilizando. Además, debido a la alta sensibilidad que tienen los microorganismos a los parámetros del ambiente es conveniente modificar el modelo LSTM para que analice datos de otros sensores importantes para el desempeño de la producción de los microorganismos durante el cultivo.

Índice General

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El monitoreo de procesos químicos y biológicos es de suma importancia, pues estos deben de llevarse a cabo bajo condiciones controladas. Se ha desarrollado una gama de diferentes sensores que proporcionan información acerca del ambiente en el que se encuentra el proceso en cuestión. Sensores como los de temperatura, de oxígeno y de pH proporcionan información que permite tomar decisiones basándose en datos obtenidos del propio proceso. Sin embargo, para obtener información más detallada se hace uso de otras técnicas de monitoreo [1].

En los procesos biológicos no se puede obtener información acerca de sus componentes directamente, pues para realizar estos análisis es necesario manipular el ambiente en el que se está llevando a cabo este proceso y en algunos casos resulta difícil el acceso e inclusive peligroso. La espectroscopia Raman utilizada como un método de análisis en línea permite tener cierta transparencia del proceso en el momento, supervisarlo y controlarlo el tiempo necesario. Se puede obtener información sobre los nutrientes, los metabolitos, la densidad molecular, entre otros. De esta manera es posible tomar mejores decisiones a partir de datos que pueden ser procesables, evitando problemas de seguridad y de calidad [1-3].

El análisis químico se puede dividir en dos partes principales: preprocesamiento de datos y la modelación de datos. Existen diferentes métodos de preprocesamiento para realizar la corrección de los datos obtenidos con la espectroscopia Raman. Asimismo, en la parte de modelación de datos se utilizan diferentes algoritmos, como por ejemplo, la técnica de mínimos cuadrados parciales basada en técnicas multivariables para cuantificar sustancias químicas [4].

Las redes neuronales se han desarrollado e implementado en el área de la microbiología como una alternativa a los modelos de regresión convencionales comúnmente utilizados. Gracias a su gran capacidad de aprendizaje y adaptabilidad en el procesamiento de datos, las redes neuronales han sido aplicadas para describir el perfil de crecimiento de bacterias comparándolas con modelos estadísticos [5].

Parte I

Planteamiento del Problema

Capítulo 2

Descripción del problema

Desde hace ya varios años existe una continua necesidad por desarrollar métodos de monitoreo en tiempo real para la caracterización y el monitoreo de bioprocesos. Estos métodos deberían ser rápidos, no invasivos, poco destructivos, precisos y económicos. La espectroscopia Raman es una herramienta que se ha ido desarrollando a lo largo de algunos años, y la cual permite analizar sistemas biológicos gracias a las numerosas ventajas que posee. Al requerir una pequeña cantidad de muestra, este método permite realizar el análisis de estos sistemas sin llegar a ocupar una gran cantidad de material, y al tener una mínima interferencia con el agua facilita el monitoreo de reacciones en solución acuosa [6]. Shaw et al. [7] realizaron un monitoreo en línea de la biotransformación por levadura de glucosa en etanol en el cual utilizan la configuración experimental mostrada en la figura 2.1.

La información de los modos vibracionales proporcionadas por la espectroscopia Raman pueden ser utilizadas como una huella dactilar espectroscópica, ya que cada muestra presenta un espectro característico, esta información puede resultar útil para determinar la composición química de una muestra o alguna propiedad de esta. Hoy en día, la espectroscopia Raman es aplicada en muchos campos de estudios que emplean la información de los espectros proporcionados por esta técnica. A partir de un procesamiento de los datos pueden llegar a determinar ciertas características de la muestra de estudio [4].

El poder trabajar utilizando fibra óptica para la iluminación de la muestra facilita la movilidad del equipo y permite el monitoreo de muestras líquidas al poder sumergir la fibra óptica en los contenedores donde se esté llevando a cabo el proceso químico. En el sector industrial se aplica

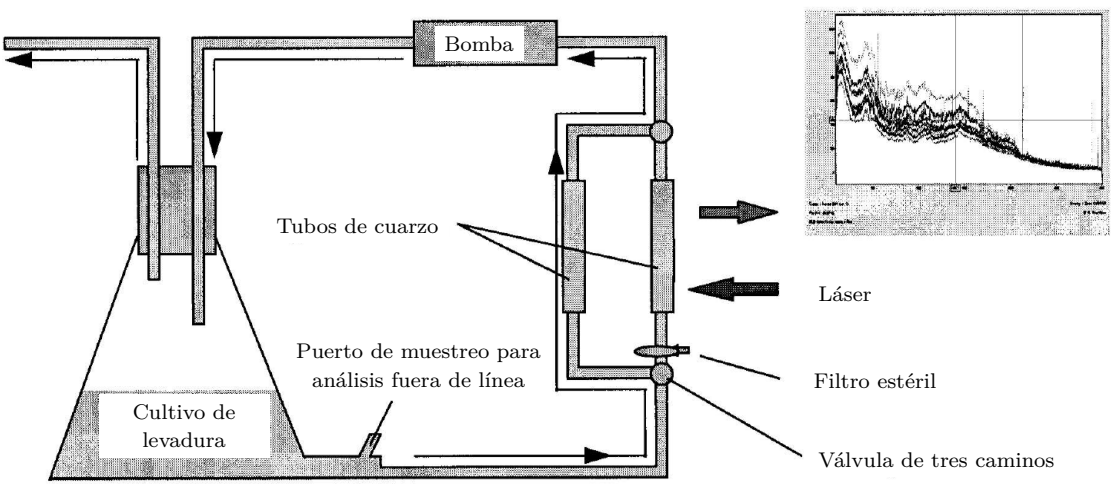


Figura 2.1: Configuración experimental diseñada por Shaw et al. [7] para el monitoreo en línea de la biotransformación por levadura de glucosa en etanol.

la espectroscopia Raman para el análisis cualitativo y cuantitativo de especies moleculares en disoluciones acuosas. De esta manera es posible monitorear la concentración de los componentes de una determinada reacción que esté sucediendo en ese determinado momento dentro del reactor o contenedor. Para relacionar el espectro con la concentración de un determinado analito, la quimiometría multivariada es una herramienta que permite crear un modelo de regresión. Uno de los métodos más empleados es la regresión de mínimos cuadrados parciales (PLSR por sus siglas en inglés) que relaciona los espectros con las concentraciones del analito que se desea monitorear. Para que el modelo pueda proporcionar una buena predicción de la concentración del analito es necesario trabajar con un conjunto de datos de entrenamiento, de los cuales se conoce la concentración real que se desea predecir con el modelo. Esta concentración deberá ser obtenida a través de algún método de determinación de concentraciones, como bien podría ser la cromatografía líquida de alta eficiencia (HPLC por sus siglas en inglés) [6]. Para comprender mejor el proceso de la creación del modelo, se muestra un esquema de los pasos que se llevan a cabo durante el procesamiento de los datos (ver figura 2.2).

Una vez generado el modelo, este debe probarse empleando un conjunto de datos de validación. En este conjunto de datos se conoce la concentración del analito para el cual se desea crear el modelo, de esta manera se verifica la exactitud con la que el modelo predice la concentración de este analito. Una vez creado, el modelo puede ser utilizado en el monitoreo de la reacción que se desea analizar[1, 8].

2.1 Hipótesis

El modelo de red neuronal diseñado, tendrá la capacidad de predecir el tiempo en que el cultivo microbiano que se está monitoreando, produzca una concentración determinada de los compuestos de interés a partir de las concentraciones proporcionadas por el espectrómetro Raman, que monitorea en tiempo real la cinética microbiana.

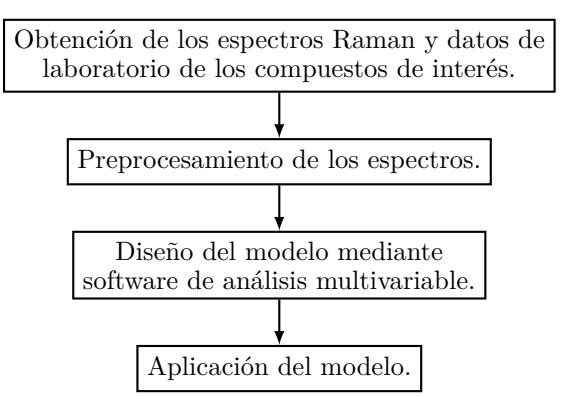


Figura 2.2: Esquema del proceso para la obtención de un modelo de predicción a partir de datos de los espectros Raman de un sistema biológico o químico.

2.2 Objetivos

Objetivo general

Crear un modelo de red neuronal profunda que permita predecir el tiempo que le tomará a un cultivo microbiano producir una determinada concentración de compuestos, y que correspondan con las cinéticas del monitoreo mediante espectroscopia Raman, obtenidos a partir de modelos de regresión generados por medio de software de análisis multivariable.

Objetivos específicos

1. Preparar los conjuntos de entrenamiento, validación y prueba que se implementaran en el entrenamiento de la red.
2. Diseñar y entrenar un modelo de red neuronal.
3. Validar el diseño del modelo de red neuronal comparando los resultados obtenidos por la red.

2.3 Metas

1. Recolectar una cantidad considerable de datos de una cinética microbiana.
2. Seleccionar el modelo microbiano que describa la cinética microbiana.
3. Elegir los conjuntos de entrenamiento, validación y prueba que serán utilizados para entrenar la red.
4. Escoger la arquitectura o arquitecturas de red neuronal que permita obtener los resultados deseados.
5. Finalizar el entrenamiento de la red neuronal.
6. Reducir el error de entrenamiento y validación lo más posible.
7. Comprobar el funcionamiento de la red neuronal con un nuevo conjunto de datos no utilizados en el entrenamiento de la red.
8. Validar la red neuronal aplicándola al monitoreo de una cinética microbiana en tiempo real.

2.4 Justificación

Los avances tecnológicos en aprendizaje profundo permiten el análisis de datos, ya sea en forma de imágenes, señales de audio, espectrogramas, entre otros. A partir de esto se pueden tomar decisiones respecto al funcionamiento de un equipo, ya sea una máquina, un clasificador, un decodificador, etc. El aprendizaje profundo aplicado en distintos procesos que requieren un análisis de grandes cantidades de datos puede ayudar a reducir el tiempo que normalmente lleva completar estos procesos con la ventaja de no requerir de alguna intervención humana. En cuanto a la biología y la medicina, el aprendizaje profundo permite la identificación y diagnóstico de anomalías en el organismo, ayuda al monitoreo de drogas y la reconstrucción de imágenes médicas para diagnósticos [9].

El aprendizaje profundo tiene distintas aplicaciones dentro del proceso de monitoreo mediante espectroscopia Raman. El análisis quimiométrico está conformado por el preprocesamiento y el modelado de los datos. Dentro del preprocesamiento de datos existen diferentes algoritmos de los cuales varios están basados en el aprendizaje automático, un ejemplo de estos es la corrección automática de picos en los espectros Raman antes pasar al modelado de los mismos [10]. Dentro del modelado de datos el algoritmo PLSR es utilizado para la cuantificación o identificación de diferentes compuestos químicos o inclusive células cancerígenas [4].

Dentro de los algoritmos de aprendizaje profundo que se han inventado recientemente resaltan las redes neuronales recurrentes (RNN por sus siglas en inglés). Este tipo de arquitectura de red neuronal permite realizar predicciones sobre el futuro del valor de una secuencia, basándose en ciertos aspectos de los valores de la secuencia que fue adoptando a lo largo del tiempo [11]. Los modelos RNN han sido utilizados en clasificar diferentes tipos de sangre utilizando espectroscopia Raman como punto de partida [12] e incluso para la clasificación de bacterias patógenas transmitidas a través de la comida [13].

Como se acaba de mencionar, la implementación del aprendizaje profundo en la espectroscopia mediante dispersión Raman tiene un gran número de aplicaciones en una amplia variedad de áreas. En cuanto a la industria, la espectroscopia Raman se ha implementado en el monitoreo de procesos biológicos y químicos que han aumentado la eficiencia en los procesos de diferentes productos [3]. La implementación de redes neuronales en la espectroscopia Raman para el monitoreo de este tipo de procesos podría aumentar aún más la eficiencia, que es lo que se busca hacer en este tipo de empresas.

Parte II

Marco Teórico

Capítulo 3

Espectroscopia Raman

En el año 1928, el físico C. V. Raman, de origen indio, observó que parte de la radiación dispersada por ciertas partículas tenía una longitud de onda distinta a la radiación incidente [6], demostrando así el fenómeno de la dispersión inelástica de los fotones, postulado por primera vez en 1923 por Adolf Smekal [2].

En el experimento que realizó Raman, un telescopio enfocaba la luz del sol sobre una muestra que contenía un líquido purificado, o bien podría ser un gas libre de contaminantes, como podría ser el vapor de agua. Una segunda lente se colocó en el sistema para recoger la radiación dispersada por la muestra y mediante una serie de filtros ópticos se pudo observar la existencia de radiación dispersada con una frecuencia distinta a la radiación incidente, que en principio es en lo que se basa la espectroscopia Raman [2].

3.1 Fundamentos

Los fotones que componen la luz pueden, o no, interactuar con la materia, estos pueden ser absorbidos o dispersados. Para que un fotón sea absorbido, y la molécula promovida a un estado excitado de mayor energía, es necesario que la energía del fotón coincida con la brecha de energía entre el estado fundamental de la molécula y un estado excitado. Sin embargo, también puede ocurrir que los fotones interactúen con la molécula y posteriormente se dispersen [2].

La técnica de dispersión es comúnmente utilizada para determinar características y aspectos importantes de las moléculas, siendo la dispersión Raman una de las principales técnicas de dispersión utilizada para la identificación de moléculas presentes en una muestra [2, 6].

Dispersión Raman

Cuando los fotones tienen una energía distinta a la brecha energética entre dos niveles de energía de una molécula, ocurre que estos fotones interactúan con la molécula y son dispersados. Estos fotones dispersados pueden ser observados al recolectar la radiación dispersada a un determinado ángulo respecto a la luz incidente, normalmente de 90° [2, 6].

La radiación es comúnmente caracterizada por la longitud de onda (λ). Cuando una molécula interacciona con un campo electromagnético, la transferencia de energía puede ocurrir cuando se cumple la condición de la frecuencia de Bohr que relaciona la energía con la frecuencia (ν). En la espectroscopia vibracional es comúnmente utilizado el número de onda ($\bar{\nu}$), el cual tienen

una relación lineal con la energía de la radiación[2, 14].

$$\lambda = \frac{c}{\nu} \tag{3.1}$$

$$\nu = \frac{\Delta E}{h} \tag{3.2}$$

$$\bar{\nu} = \frac{\nu}{c} \tag{3.3}$$

Considerando las ecuaciones (3.1–3.3) y observando que la energía es proporcional al recíproco de la longitud de onda, en el espectro electromagnético el intervalo de mayor energía se encuentra alrededor de los $10^{-11}m$ de longitud de onda correspondiente a los rayos Gamma y rayos X, mientras que el intervalo de menor energía está alrededor de los $10m$ correspondiente a las microondas y a las ondas de radio, tal como se puede apreciar en la figura 3.1 [2, 14]. La espectroscopia infrarroja utiliza un intervalo de frecuencias de tal manera que permite que

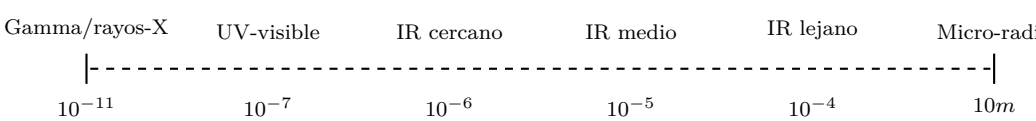


Figura 3.1: Espectro electromagnético de acuerdo a la longitud de onda que lo representa [2].

los fotones al interactuar con la molécula puedan coincidir con la brecha energética que existe entre dos niveles de energía y varios de ellos sean absorbidos por la molécula. En contraste, la espectroscopia Raman utiliza únicamente una frecuencia de radiación para irradiar la muestra y esta sea dispersada por la interacción con las moléculas con una energía diferente a la de la onda incidente y, por lo tanto, con una longitud de onda distinta. En la dispersión Raman, la luz interacciona con las moléculas distorsionando la nube de electrones alrededor de ella, creando por un corto periodo de tiempo lo que se conoce como un "estado virtual" de energía, sin embargo, debido a que este estado de energía es inestable, el fotón es inmediatamente irradiado [2, 6, 14].

Si la nube de electrones es la única involucrada en el proceso de dispersión, debido a que los electrones son relativamente ligeros, los fotones que serán dispersados lo harán con un nulo o leve cambio en su frecuencia, prácticamente imperceptible. A este proceso se le considera dispersión elástica o dispersión de Rayleigh. Este tipo de dispersión es el tipo de dispersión dominante que ocurre cuando se irradia algún material o muestra, no proporciona información relevante. Por otro lado, cuando se induce algún movimiento nuclear durante el proceso de dispersión, parte de la energía puede ser transferida de la molécula al fotón o del fotón a la molécula, cuando se da este caso, los fotones dispersados por la molécula llevarán una energía distinta a los fotones incidentes, lo cual implica un cambio en su frecuencia, que a pesar de que son cambios pequeños, pueden ser detectados y proporcionar información acerca de la molécula con la que tuvo alguna interacción. Este proceso es lo que se conoce como dispersión inelástica o dispersión Raman [2].

Los estados virtuales no son estados reales de energía de la molécula, estos estados son creados debido a la polarización en la nube de electrones a causa de los fotones que interactúan con esta. La energía de estos estados está determinada por la frecuencia de la luz incidente [2].

Dispersión de Stokes y anti-Stokes

Cuando ocurre el proceso de dispersión inelástica, la molécula que originalmente se encontraba en un estado vibracional de energía (m) después de abandonar el estado virtual, puede permanecer en un nivel mayor de energía que el original (n), a este proceso se le conoce como dispersión Stokes. Sin embargo, muchas moléculas se encontrarán en un estado de excitación (n) al momento de ser irradiadas y puede ocurrir que durante la dispersión su estado de excitación final sea inferior (m), a este proceso se le conoce como dispersión anti-Stokes, estos procesos pueden proporcionar información acerca de las transiciones vibracionales, rotacionales y otras de baja frecuencia que ocurren en las moléculas [2, 6]. En el diagrama de la figura 3.2 se muestran gráficamente estos procesos de dispersión. La intensidad relativa de estos dos

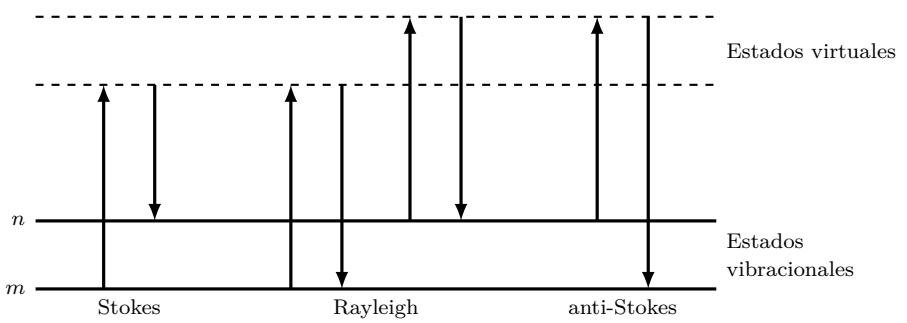


Figura 3.2: En este diagrama se representan los procesos de dispersión de Rayleigh y Raman. Las líneas punteadas representan los estados virtuales de energía. las flechas indican la absorción (flechas que suben) y la radiación (flechas que bajan) de energía de la molécula. Los estados de energía de la molécula m y n representan los estados de menor y mayor energía respectivamente [2].

procesos dependerá de la cantidad de moléculas que se encuentren en cada uno de estos dos estados de energía, sin embargo, debido a que es más probable que una molécula se encuentre en su estado base de energía, el cual es el estado de menor energía, se puede deducir que la dispersión anti-Stokes será más débil en comparación con la dispersión Stokes [2].

3.2 Instrumentación

En los bioprocesos suelen ocurrir diferentes cambios en sus propiedades intensivas y extensivas debido a influencias físicas y químicas. Los métodos tradicionales de análisis por lo general son demasiado lentos impidiendo hacer correcciones a pequeños cambios que sucedan en el bioproceso. El monitoreo mediante espectrometría Raman es una solución a este problema, permitiendo detectar cambios que ocurren durante el proceso en tiempo real y tomar acciones

para corregirlos sin que afecte el resultado del producto final. Algunos espectrómetros modernos están capacitados con un alto poder de computación y algoritmos avanzados basados en quimiometria que les permiten tener una alta capacidad de análisis y control de procesos complejos; estos instrumentos están compuestos por una fuente láser, un sistema para iluminar la muestra y un espectrómetro apropiado para el monitoreo [3, 6].

Fuentes de iluminación

El descubrimiento del láser en 1960 revolucionó varias áreas de la espectroscopia. En la espectroscopia Raman se utilizan rayos láser como fuente de iluminación debido a su alta intensidad, la cual es necesaria para producir dispersión Raman suficientemente intensa para poder realizar mediciones. Se debe ser cuidadoso al seleccionar la fuente de iluminación, debido a que la intensidad de la dispersión Raman varía con la cuarta potencia de la frecuencia, sin embargo, fuentes de longitud de onda corta pueden llegar a ocasionar fluorescencia intensa o inclusive fotodescomposición en la muestra, lo cual es una desventaja importante en los bioprocesos, impidiendo un correcto monitoreo. En la tabla 3.1 se muestran algunas fuentes de láser más utilizadas para la espectroscopia mediante dispersión Raman. Ciertas muestras con color o algunos disolventes son capaces de absorber la radiación Raman incidente o dispersada, ocasionando una pérdida de información que puede llegar a ser significativa para el monitoreo de los bioprocesos [6, 15].

Tabla 3.1: Fuentes más comunes de láser usados en espectroscopia Raman [6].

Tipo de láser	**Longitud de onda [nm]**
Ion de argón	488.0 ó 514.5
Ion de criptón	413.1, 530.9, 647.1
Helio-neón	632.8
Diodo	660-880
Nd-YAG	1064

Sistema de iluminación con fibra óptica

La espectroscopia Raman permite una mejor manipulación de la muestra al ser iluminada, una de las ventajas es el hecho de usar radiación visible o infrarroja cercana, esto permite la implementación de fibra óptica para transmitir la radiación de excitación hasta una distancia aproximada de 100 m. Estas fibras transmiten la radiación de excitación directamente a la muestra, la fibra óptica permite la implementación del monitoreo en muestras líquidas, ya que puede ser sumergida dentro de estas. Por otro lado, una segunda fibra recolecta la dispersión Raman y la transmite al espectrómetro donde se analizará la información recolectada. Las fibras ópticas que se utilizan en la espectroscopia Raman, están compuestas por una o varias

fibras en el centro que se encargaran de transmitir la luz del láser hacia la muestra, esta fibra es comúnmente denominada como fibra de excitación; alrededor de estas fibras se encuentran aquellas fibras denominadas fibras de recolección (tal como se muestra en la figura 3.3), estas se encargan de transmitir la dispersión Raman que logran captar, directamente hacia el espectrómetro, en donde se analizará la señal recibida [2, 6].

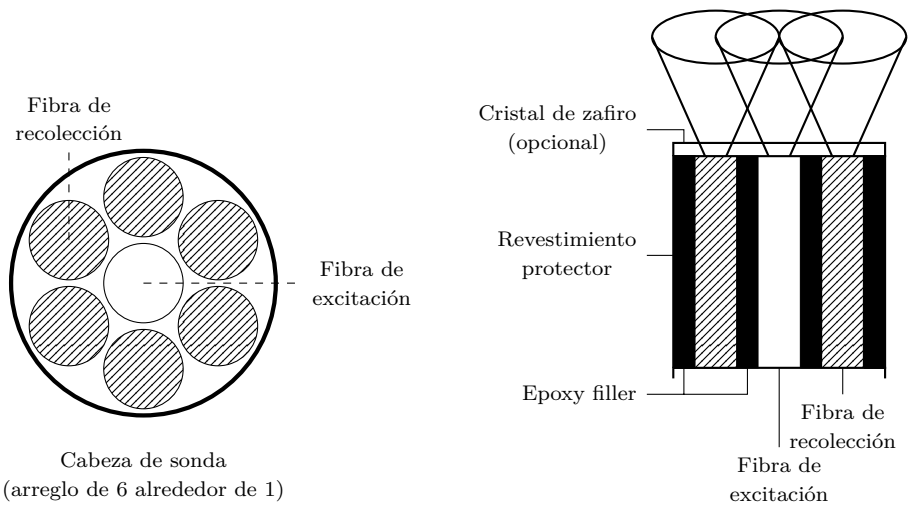


Figura 3.3: a) Corte frontal de una fibra óptica utilizada para espectroscopia Raman, se puede apreciar una fibra de excitación rodeada de seis fibras de recolección. b) Corte lateral de la misma fibra óptica, se puede observar las separaciones y recubrimientos que poseen entre fibras, el cristal que se ubica en la salida de la fibra no altera la medición de la dispersión Raman recolectada [2].

Espectrómetros Raman

Los instrumentos utilizados para la espectroscopia Raman hacen uso de distintos dispositivos de alta calidad para separar las líneas pertenecientes a la dispersión Raman, las cuales son relativamente débiles en comparación con la dispersión Rayleigh. Comúnmente estos dispositivos ocupan dos o inclusive tres monocromadores, con la finalidad de separar, de cualquier otra luz, las frecuencias correspondientes a la dispersión Raman e incrementarla para poder observar mejor los picos Raman individualmente. La tecnología de filtros se ha ido desarrollando, obteniendo los llamados filtros de ranura o de muesca que cumplen muy bien estas funciones en los espectrómetros [2, 6].

Para la detección de la intensidad de la dispersión Raman, la mayoría de los espectrómetros Raman comerciales están equipados con sistemas que cuentan los fotones. Uno de los detectores más utilizados son los dispositivos de acoplamiento de carga (CCD, por sus siglas en inglés). Un detector CCD es un semiconductor basado en silicio conformado por una matriz de elementos fotosensibles, los cuales generan fotoelectrones y los almacenan como una pequeña

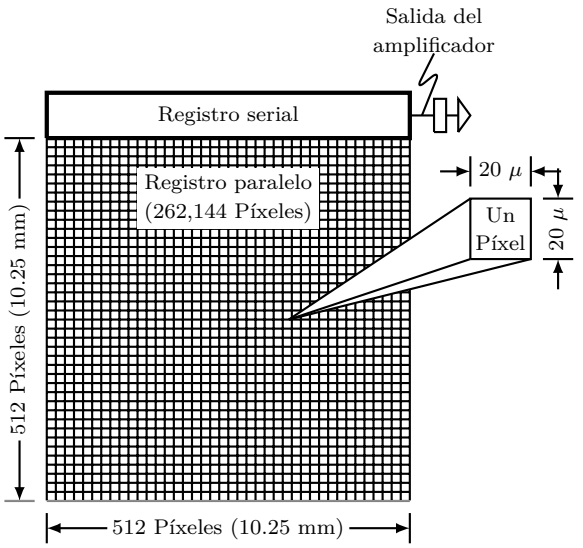


Figura 3.4: Esquema de un detector CCD [14].

carga, en la figura 3.4 se puede observar un esquema de la parte interna de un detector CCD. Los dispositivos de acoplamiento de carga pueden ser bidimensionales, o en algunos casos, lineales [6, 14].

Otros instrumentos de espectroscopia Raman que se utilizan son aquellos conocidos como instrumentos Raman de transformada de Fourier (FT-Raman), los cuales están equipados con un interferómetro de Michelson, y con una fuente de Nd-YAG de onda continúa comúnmente de 1064 nm (1.064 μm) que elimina virtualmente la fluorescencia o la fotodescomposición de las muestras lo cual permite poder estudiar compuestos fluorescentes o con colorantes. Otra

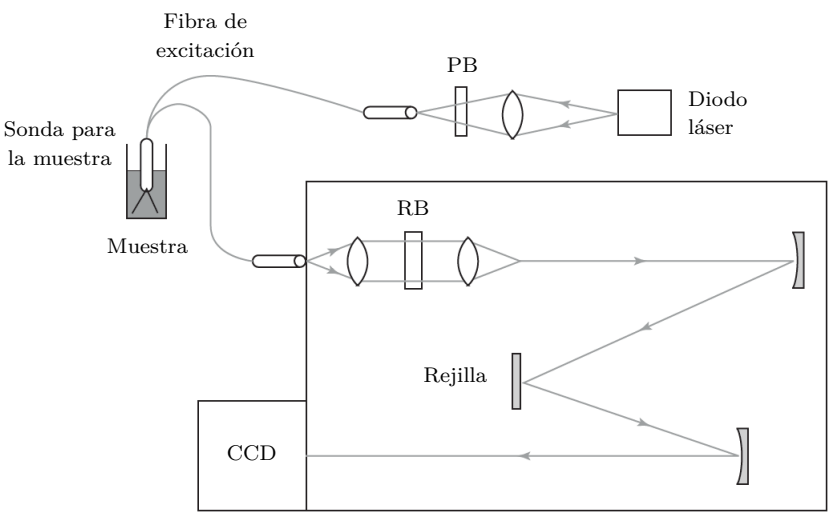


Figura 3.5: Espectrómetro Raman de fibra óptica con espectrógrafo y detector **CCD**. El divisor de haz (**PB**) se utiliza para aislar una parte de la luz del láser. El filtro de muesca (**RB**) reduce al mínimo la radiación de Rayleigh [6].

ventaja de estos instrumentos de transformada de Fourier es que permite obtener espectros y mediciones de alta resolución con una gran precisión de la frecuencia en comparación con otros instrumentos [6].

En la figura 3.5 se muestra el arreglo de un espectrómetro Raman de fibra óptica que utiliza un filtro de muesca para eliminar la luz parasitaria de la dispersión Raman, la cual es recolectada por un detector CCD [6]. La implementación de los dispositivos mencionados anteriormente proporcionan una lectura de la dispersión Raman de alta calidad.

3.3 Aplicaciones de la espectroscopia Raman

La espectroscopia Raman es una técnica que permite identificar algunos compuestos a partir de su espectro Raman característico y es útil en el análisis de sistemas biológicos debido a su capacidad para trabajar en medios acuosos sin sufrir interferencias. Desde su introducción en la industria gracias a los instrumentos FT-Raman, la implementación de fibra óptica y el desarrollo de nuevos detectores. Esta técnica ha sido aplicada en una amplia variedad de áreas, incluyendo polímeros, semiconductores, arte, arqueología, biotecnología, ciencias forenses, análisis de procesos y materiales [2, 6].

En la industria farmacéutica y biológica, la espectroscopia Raman es valiosa para el estudio de materiales *in situ*, estructuras físicas, moléculas biológicas y sistemas acuosos en biorreactores. El uso combinado de pequeñas muestras, microscopios y fibra óptica ha aumentado la popularidad de la espectroscopia Raman en esta industria [2].

Además, la espectroscopia Raman es útil en el monitoreo de fermentaciones para producir compuestos como etanol a partir de glucosa y para monitorear reacciones que se llevan a cabo en la industria. La implementación de fibra óptica, diferentes tipos de filtros e instrumentos FT-Raman hacen que sea más fácil de utilizar y han mejorado su aplicabilidad [2, 3].

Capítulo 4

Análisis Multivariable

En diferentes métodos de laboratorio o monitoreo en línea una gran cantidad de variables son medidas simultáneamente, dando como resultado datos multivariados. En el caso de la espectroscopia Raman se analizan diferentes longitudes de onda dependiendo de los compuestos que se estén monitoreando. Existe una gran variedad de métodos de análisis multivariable que permiten analizar la importancia que tiene cada una de las variables involucradas para realizar un buen monitoreo o análisis de los compuestos, en los cuales destacan los siguientes [16]:

- Análisis de componentes principales (PCA).
- Métodos de análisis de conglomerados (CA).
- El método del vecino mas próximo a K (KNN).
- Regresión lineal múltiple (MLR).
- Regresión de componentes principales (PCR).
- Regresión por mínimos cuadrados parciales (PLS).

La implementación de la regresión PLS o PCR, junto a la espectroscopia Raman han permitido el desarrollo de equipos de monitoreo para su aplicación en distintos procesos de fermentación, así como en la detección de sustancias ilícitas [3, 17].

4.1 PLS-R (Regresión por Mínimos Cuadrados Parciales)

La regresión PLS es una generalización recientemente desarrollada de la regresión lineal múltiple (MLR), en la cual se utilizan combinaciones lineales de las variables predictoras X y además permite modelizar simultáneamente varias variables de respuesta Y. Este enfoque fue originado alrededor de 1975 por Herman Wold para la modelización de datos en cadenas de matrices (bloques), y en 1980, Svante Wold y Harald Martens modificaron el modelo más simple de dos bloques para adaptarlo a conjuntos de datos complicados de ciencia y tecnología [16, 17].

Para el desarrollo de un modelo de regresión PLS se parte de un conjunto de datos de entrenamiento el cual cuenta con un total de N observaciones con K X-variables que se denotan por $\mathbf{x}_k$ $(k = 1, \dots, K)$, y M Y-variables $\mathbf{y}_m$ $(m = 1, 2, \dots, M)$ formando así dos matrices $\mathbf{X}$ y $\mathbf{Y}$ con dimensiones $(N \times K)$ y $(N \times M)$ respectivamente [17].

El modelo PLS busca una correlación lineal entre las Y-variables y los predictores $\mathbf{T}$ de $\mathbf{X}$ denotados por $\mathbf{t}_a$ $(a = 1, 2, \dots, A)$ los cuales son una estimación lineal de las variables originales

$\mathbf{x}_k$ con coeficientes w^*_{ka} $(a = 1, 2, \ldots, A)$ tal como se muestra en las ecuaciones 4.1 y 4.2 [17].

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{4.1}$$

$$t_{ia} = \sum_k W^*_{ka} X_{ik} \tag{4.2}$$

La relación lineal de las Y-variables con los predictores al multiplicarlos por los coeficientes c_{am} se expresan en las ecuaciones 4.3 y 4.4, en donde f_{im} expresa las desviaciones entre las observaciones y las respuestas del modelo [17].

$$\mathbf{Y} = \mathbf{T}\mathbf{C}' + \mathbf{F} \tag{4.3}$$

$$y_{im} = \sum_a c_{ma} t_{ia} + f_{im} \tag{4.4}$$

Al definir los "coeficientes del modelo de regresión PLS" b_{mk} se puede expresar la relación lineal de las Y-variables con las variables originales como en las ecuaciones 4.5 y 4.6 en donde la matriz $\mathbf{B}$ está definida como se muestra en las ecuaciones 4.7 y 4.8 [17].

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^*\mathbf{C}' + \mathbf{F} = \mathbf{X}\mathbf{B} + \mathbf{F} \tag{4.5}$$

$$y_{im} = \sum_a c_{ma} \sum_k w^*_{ka} x_{ik} + f_{im} = \sum_k b_{mk} x_{ik} + f_{im} \tag{4.6}$$

$$\mathbf{B} = \mathbf{W}^*\mathbf{C}' \tag{4.7}$$

$$b_{mk} = \sum_a c_{ma} w^*_{ka} \tag{4.8}$$

La regresión PLS se puede ver como la generación de nuevas variables a partir de una combinación lineal de las antiguas, estas nuevas variables funcionan como predictores de Y [17].

Capítulo 5

Redes neuronales

Una red neuronal es una máquina diseñada para modelar la forma en que el cerebro realiza una determinada tarea. Usualmente, son implementadas utilizando componentes electrónicos o simuladas mediante algún software en un equipo de cómputo. Estas redes neuronales están conformadas por un conjunto de varias neuronas interconectadas de una forma masiva, cada neurona actúa como una "unidad de procesamiento" logrando, de esta forma, el mejor rendimiento posible de la red [18]. Haykin [18] ofrece la siguiente definición de una red neuronal:

> Una red neuronal es un procesador distribuido masivamente en paralelo compuesto por unidades de procesamiento simples que tiene una propensión natural a almacenar conocimiento experiencial y ponerlo a disposición para su uso. Se parece al cerebro en dos aspectos:
>
> 1. El conocimiento es adquirido por la red de su entorno a través de un proceso de aprendizaje.
> 2. Los parámetros de conexión de las interneuronas, conocidas como pesos sinápticos, se utilizan para almacenar el conocimiento adquirido.
>
> Haykin [18].

Una neurona es una unidad de procesamiento de información y es el componente fundamental para una red neuronal. La figura 5.1 muestra el modelo de una red neuronal en donde se pueden observar tres elementos básicos que conforman una neurona [18]:

1. Un conjunto de conexiones que comunican los datos de entrada con la neurona y cada uno posee su propio peso sináptico (ω_{kj}).
2. Un contador que se encarga de sumar los datos de entrada (x_j) ponderados por sus respectivos pesos sinápticos, creando así una combinación lineal de estos.
3. Una función de activación ($\varphi(\nu_k)$) que es la encargada de limitar la amplitud de salida de una neurona.

Las funciones de activación definen la salida de una neurona de acuerdo al valor obtenido por los datos de entrada. Existen dos tipos de funciones de activación básicas, la función escalón y la función sigmoide (ecuaciones 5.1 y 5.2 respectivamente).

$$\varphi(\nu_k) = \begin{cases} 1 & si \quad \nu \geq 0 \\ 0 & si \quad \nu < 0 \end{cases} \tag{5.1}$$

$$\varphi(\nu_k) = \frac{1}{1 + \mathrm{e}^{-\alpha\nu_k}} \tag{5.2}$$

El *sesgo* (b_k) tiene la función de intensificar o disminuir la entrada de la función de activación

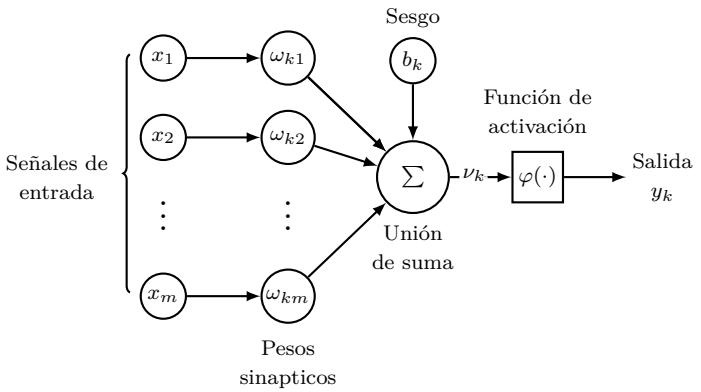


Figura 5.1: Modelo no lineal de una neurona. La neurona es etiquetada como k y los subíndices de los pesos sinápticos (kj) hacen referencia a la neurona en cuestión y al número de peso que representan respectivamente [18].

(ν_k) dependiendo del valor que este contenga.

En términos matemáticos, se puede describir las operaciones realizadas por la neurona k de la figura 5.1 con las siguientes ecuaciones:

$$u_k = \sum_{j=1}^{m} \omega_{kj} x_j \tag{5.3}$$

$$\nu_k = u_k + b_k \tag{5.4}$$

$$y_k = \varphi(u_k + b_k) \tag{5.5}$$

donde x_j son las señales de entrada a la neurona; ω_{kj} son los pesos sinápticos respectivos de cada entrada; u_k es la combinación lineal de salida dada por las señales de entrada; b_k es el sesgo; ν_k es la salida dada por la ponderación de las señales de entrada y el sesgo; $\varphi(\cdot)$ es la función de activación; y y_k es la señal de salida de la neurona [18, 19].

Con el fin de simplificar las operaciones, se define la siguiente entrada con su respectivo peso sináptico:

$$x_0 = +1 \tag{5.6}$$

$$\omega_{k0} = b_k. \tag{5.7}$$

Tal que las ecuaciones 5.4 y 5.5 pueden ser reescritas de la siguiente manera:

$$\nu_k = \sum_{j=0}^{m} \omega_{kj} x_j \tag{5.8}$$

$$y_k = \varphi(\nu_k). \tag{5.9}$$

Lo cual se puede observar gráficamente en la figura 5.2.

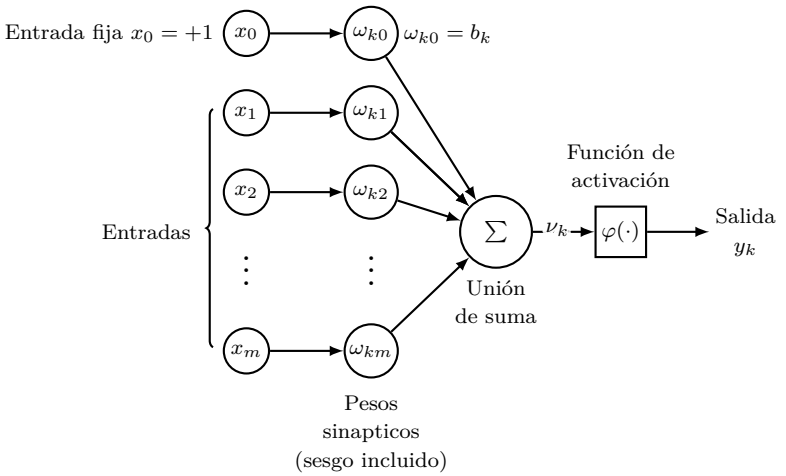


Figura 5.2: Modelo no lineal de una neurona. Este modelo, similar al de la figura 5.1, agrega una entrada x_0 y un peso sináptico ω_{k0}, el cual contiene el sesgo b_k de la neurona [18].

De acuerdo con Haykin [18], se puede simplificar el modelo utilizando la idea de gráficos de flujo siempre y cuando las señales fluyan en la dirección que indican las flechas en los enlaces; las señales de los nodos sea igual a la suma algebraica de todas las señales que ingresan a ese nodo; y que la señal de salida de un nodo se transmita en cada enlace de salida de dicho nodo, siendo independientes de las funciones de transferencia que actúen en cada nodo. Con base en lo mencionado anteriormente, Haykin [18] propone la siguiente definición matemática de una red neuronal:

> Una red neuronal es un grafo dirigido que consta de nodos con enlaces sinápticos y de activación interconectados, y se caracteriza por cuatro propiedades:
>
> 1. Cada neurona está representada por un conjunto de enlaces sinápticos lineales, un sesgo aplicado externamente y un enlace de activación posiblemente no lineal. El sesgo está representado por un enlace sináptico conectado a una entrada fija en +1.
> 2. Los enlaces sinápticos de una neurona ponderan sus respectivas señales de entrada.
> 3. La suma ponderada de las señales de entrada define el campo local inducido de la neurona en cuestión.
> 4. El enlace de activación aplasta el campo local inducido de la neurona para producir una salida.
>
> Haykin [18].

5.1 Arquitecturas de redes neuronales

La palabra arquitectura hace referencia a la estructura general de la red, cuantas neuronas debe tener y como deben estar conectadas entre sí, la forma en que se estructuran las neuronas

dentro de la red neuronal está íntimamente relacionada con el algoritmo de aprendizaje que se utilizara para entrenar a la red [11, 18]. Se identifican tres clases fundamentales de arquitectura de redes:

1. Redes de una sola capa: En las redes neuronales las neuronas se organizan en forma de capas, este tipo de redes se crea únicamente con una sola capa de neuronas a las cuales les llegan los datos de entrada y de estas se obtienen los datos de salida de la red neuronal (figura 5.3).

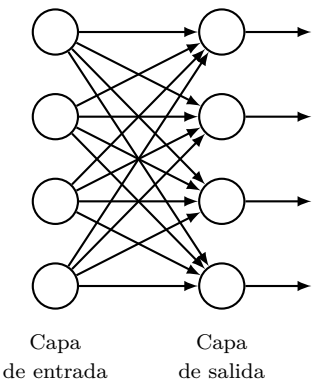


Figura 5.3: Red neuronal con una sola capa de neuronas [18].

2. Redes multicapa: Contiene más de una capa de neuronas, estas capas extras, denominadas "capas ocultas", tienen la finalidad de intervenir de una manera útil entre la entrada y salida de la red (figura 5.4).

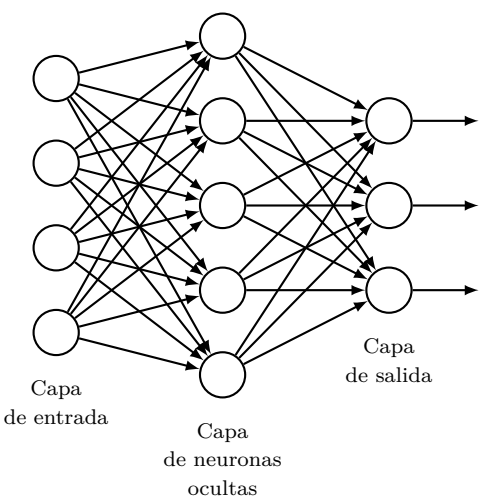


Figura 5.4: Red neuronal con una capa oculta de neuronas [18].

3. Redes recurrentes: Posee al menos un circuito de retroalimentación, en donde la salida de una neurona se usa como un dato de entrada para la misma neurona, alguna otra neurona que se encuentre en su misma capa, o en alguna de las capas anteriores (en el

caso de las redes multicapa). Este tipo de redes analiza valores obtenidos en un periodo de tiempo previo (figura 5.5).

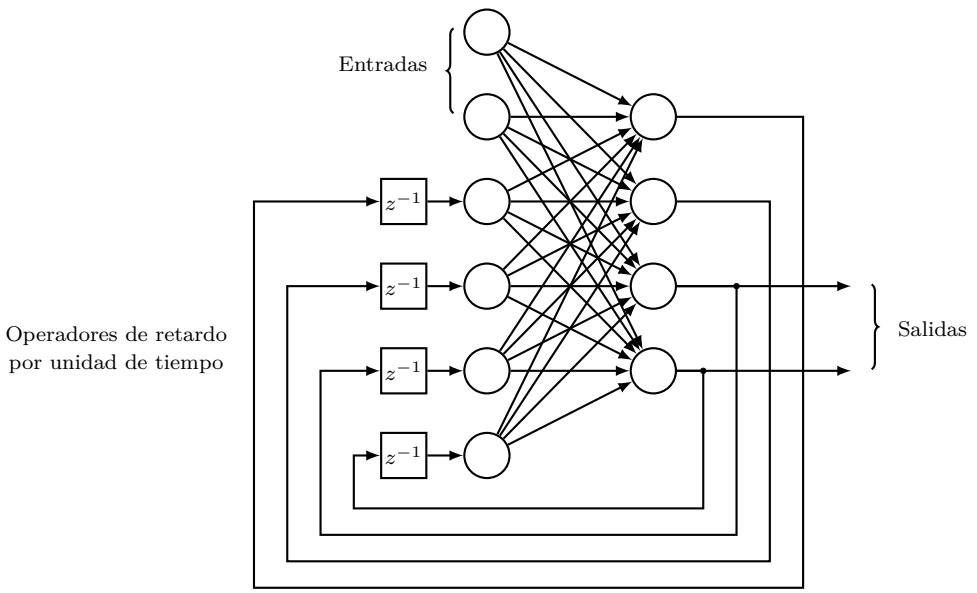


Figura 5.5: Red recurrente con neuronas ocultas [18].

Perceptrones multicapa

Los perceptrones multicapa son un tipo de modelo de red neuronal que usan formas paramétricas para las funciones base cuyos valores de los parámetros se adaptan en la fase de entrenamiento. En un perceptron multicapa, cada neurona de la red posee una función de activación no lineal diferenciable; como su nombre lo indica, contiene una o más capas ocultas; y exhibe un alto grado de conectividad, cuyo alcance depende de los pesos sinápticos de la red [18, 19].

RNN (Recurrent Neural Network)

Las RNNs son un tipo de red neuronal especializadas en el procesamiento de una secuencia de valores $x^{(1)}, ..., x^{(\tau)}$. Estas redes neuronales son capaces de procesar series de datos de diferentes tamaños. Esto lo logran compartiendo parámetros en diferentes partes dentro de la red, de tal forma que la red neuronal sea capaz de comprender el contexto de la serie de valores que se le han ingresado. La aplicación de este tipo de estructuras tiene un profundo impacto en la capacidad de aprendizaje de las redes neuronales, ademas de un aumento significativo de su desempeño [11, 18].

Existen diferentes patrones de diseños de RNNs, algunos de ellos son los siguientes:

- RNNs que producen una salida en cada paso de tiempo y que tienen conexiones de retroalimentación entre algunas de sus capas ocultas (figura 5.6).

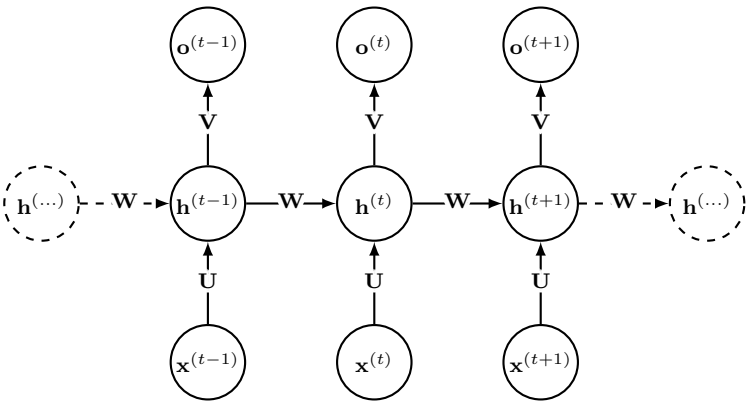


Figura 5.6: Representación gráfica del flujo de información de una RNN en el tiempo. Las conexiones de retroalimentación de esta RNN se dan entre sus capas ocultas **h**. La RNN genera una salida **o** en cada instante de tiempo para la secuencia de valores de entrada **x** [11].

- RNNs que producen una salida en cada paso de tiempo y esta a su vez funciona como entrada de alguna capa oculta en el siguiente paso de tiempo (figura 5.7).

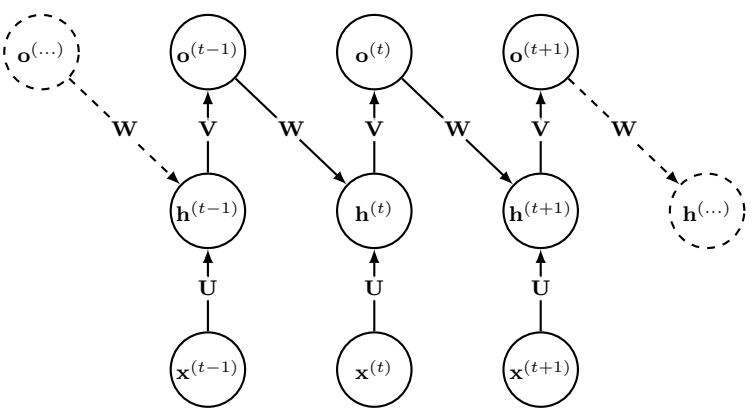


Figura 5.7: Representación gráfica del flujo de información de una RNN en el tiempo. Las conexiones de retroalimentación de esta RNN se dan entre su capa de salida **o** y sus capas ocultas **h**. La RNN genera una salida **o** en cada instante de tiempo para la secuencia de valores de entrada **x** [11].

- RNNs con conexiones entre sus capas ocultas, leen la secuencia de datos entera antes de producir una única salida (figura 5.8).

Para cada neurona dentro de la configuración de la RNN mostrada en la figura 5.6 en cada paso de tiempo desde $t = 1$ hasta $t = \tau$ se aplican las siguientes ecuaciones

$$\nu^{(t)} = \mathbf{b} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{x}^{(t)} \tag{5.10}$$

$$\mathbf{h}^{(t)} = \varphi(\nu^{(t)}) \tag{5.11}$$

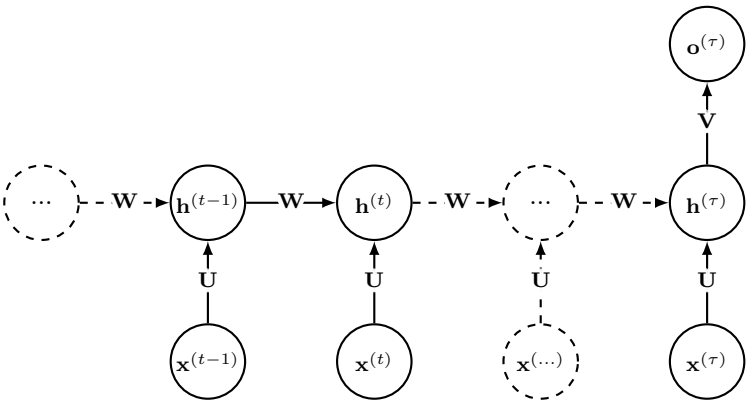


Figura 5.8: Representación gráfica del flujo de información de una RNN en el tiempo. Las conexiones de retroalimentación de esta RNN se dan entre sus capas ocultas **h**. La RNN genera una única salida **o** en el tiempo τ para la secuencia de valores de entrada **x** [11].

$$\mathbf{o}^{(t)} = \mathbf{c} + \mathbf{V}\mathbf{h}^{(t)} \tag{5.12}$$

donde $\varphi(\nu)$ representa una función de activación; **b** y **c** representa los sesgos que se aplican en sus respectivas neuronas; **U**, **V** y **W** son vectores que contienen los valores de los pesos que se aplican en cada enlace entre sus respectivas neuronas [11]. Se puede deducir similaermente las ecuaciones aplicadas en las configuraciones de las RNNs que se muestran en las figuras 5.7 y 5.8.

LSTM (Long Short-Term Memory)

Las redes de memoria a corto y largo plazo (LSTM por sus siglas en ingles) son un tipo de modelo de RNN. Este tipo de redes pueden modificar los pesos de algunas de sus unidades, lo cual permite a la red acumular información con el paso del tiempo la cual puede ser interpretada como algún tipo de característica o categoría de la secuencia de entrada que esta red este analizando. Una vez la información haya sido utilizada, la red puede permitirse olvidar la información acumulada mediante mecanismos que devuelven los pesos modificados a su estado inicial. Esta característica de las LSTMs brinda una solución al problema del desvanecimiento o explosión del gradiente que se encuentra muy presente en las RNN [11].

Las LSTMs se han desempeñado extremadamente bien en diversas aplicaciones, como por ejemplo, en el reconocimiento y generación de escritura a mano, reconocimiento de voz, como maquinas de traducción, análisis de datos, entre muchas otras. Como modelos de predicción de series en tiempo real y series de tiempo caóticas y estocásticas, las LSTMs se han desempeñado considerablemente bien, en comparación con otros modelos tanto paramétricos como no paramétricos [11, 20, 21].

La estructura de una red LSTM esta conformada por unidades (ver figura 5.9) conectadas

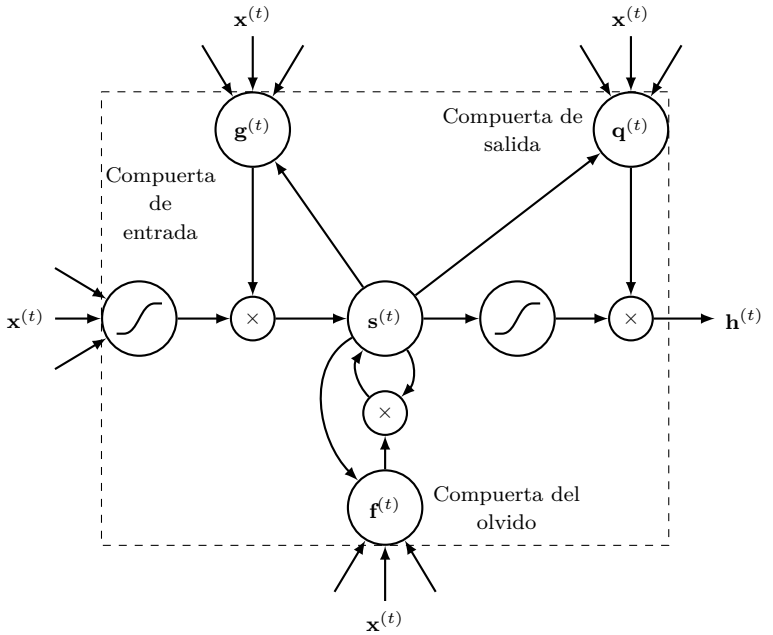


Figura 5.9: Diagrama de una red neuronal recurrente LSTM. Aquí se representa la interacción entre cada uno de los nodos que la conforman, siendo $\mathbf{g}^{(t)}$ la compuerta de entrada; $\mathbf{q}^{(t)}$ la compuerta de salida; $\mathbf{f}^{(t)}$ la compuerta del olvido; $\mathbf{s}^{(t)}$ la unidad de estado de la red; $\mathbf{h}^{(t)}$ el vector de salidas de la red neuronal al tiempo t y; $\mathbf{x}^{(t)}$ los datos de entrada de la red neuronal en el tiempo t. Los nodos ($\times$) multiplican sus entradas [11, 22].

entre si, cuyas salidas son almacenadas conforme pasa el tiempo ($\mathbf{o}^{t-1}$ es un vector que contiene las salida del resto de unidades de la red previas al tiempo t), cada unidad esta conformada por tres nodos que son conocidos como compuertas. La compuerta de entrada ($\mathbf{g}^{(t)}$) se encarga de procesar la información recibida en el tiempo t como normalmente lo haría una neurona artificial regular. La compuerta de salida ($\mathbf{q}^{(t)}$) se encarga de determinar que tanta información acumulada se ocupara en el instante de tiempo t. La compuerta del olvido ($\mathbf{f}^{(t)}$) es la que determina si se continuara almacenando la información procesada o será eliminada. Dentro de cada nodo de la red neuronal se aplican las siguientes ecuaciones:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f o_j^{(t-1)} + \sum_j V_{i,j}^f s_j^{(t-1)}\right) \tag{5.13}$$

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g o_j^{(t-1)} + \sum_j V_{i,j}^g s_j^{(t-1)}\right) \tag{5.14}$$

$$q_i^{(t)} = \sigma\left(b_i^q + \sum_j U_{i,j}^q x_j^{(t)} + \sum_j W_{i,j}^q o_j^{(t-1)} + \sum_j V_{i,j}^q s_j^{(t-1)}\right) \tag{5.15}$$

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)} + \sum_j V_{i,j} s_j^{(t-1)}\right) \tag{5.16}$$

$$o_i^{(t)} = \tanh(s_i^{(t)}) q_i^{(t)} \tag{5.17}$$

en donde σ es una función sigmoidal que toma valores entre 0 y 1, que junto con la función $\tanh(\cdot)$ son utilizadas como funciones de activación en sus respectivas neuronas; $\mathbf{b}$, $\mathbf{U}$, $\mathbf{W}$ y $\mathbf{V}$ son los bias y los vectores de peso que se aplican en sus respectivas neuronas (se pueden identificar por los superíndices f, g y q) [11].

5.2 Aprendizaje supervisado

Los procesos de aprendizaje de las redes neuronales se pueden caracterizar en dos grandes grupos, el aprendizaje supervisado y el no supervisado. El aprendizaje supervisado es una forma en que la red neuronal puede obtener conocimiento, el cual Haykin [18] define como:

> El conocimiento se refiere a información almacenada o modelos utilizados por una persona o máquina para interpretar, predecir y responder adecuadamente al mundo exterior.
>
> Haykin [18].

En el aprendizaje supervisado se tiene conocimiento sobre el entorno y este conocimiento será presentado a la red neuronal como un conjunto de ejemplos de entrada y salida, representando estas últimas los valores óptimos que deberá emular la red neuronal. Estos ejemplos deberán ser representativos al entorno que se quiere estudiar. En principio la red no posee conocimiento sobre el entorno y con ejemplos se ajustarán los valores para poder obtener las salidas óptimas deseadas, con ayuda de una señal de "error" que se le proporcionará en cada iteración de entrenamiento con la finalidad de reducirlo lo más posible. De esta forma el conocimiento se transfiere a la red a través del entrenamiento [18].

El algoritmo de *backpropagation* es un método desarrollado a mediados de la década de 1980 que representó un hito en las redes neuronales, proporcionando un método computacionalmente eficiente para el entrenamiento de perceptrones multicapa. El entrenamiento se desarrolla en dos fases, la *fase de avance* y la *fase de retroceso* [11, 18].

En la *fase de avance*, después de fijar los pesos sinápticos de la red, una señal de entrada es propagada a través de la red, capa por capa, hasta la salida. Los cambios que ocurren en esta fase se dan únicamente en los potenciales de activación y las salidas de las neuronas en la red.

En la *fase de retroceso*, se produce una señal de error que compara la salida de la red con una respuesta deseada y es transmitida a través de toda la red, capa por capa, y de forma regresiva. En esta fase se realizan cambios a los valores de los pesos sinápticos en las interconexiones entre las neuronas que conforman la red para reducir el error detectado.

Capítulo 6

Modelos de crecimiento

El estudio de cultivos de bacterias es una disciplina que ha sido ampliamente desarrollada, la evolución de la población y la calidad de los productos se ven afectadas por las condiciones ambientales en que se encuentran [23, 24]. Se pueden considerar diferentes fases en el crecimiento de estos cultivos, tanto positivas como negativas, y nulas. Para el crecimiento de un cultivo bacteriano, Monod [23] define la "concentración celular" como el número de células individuales por unidad de volumen de un cultivo, y la "densidad bacteriana" como el peso seco de células por unidad de volumen de un cultivo; además, proporciona la siguiente sucesión de fases en un cultivo bacteriano, caracterizadas por variaciones en la tasa de crecimiento:

1. Fase de retraso: tasa de crecimiento nula;
2. Fase de aceleración: aumento de la tasa de crecimiento;
3. Fase exponencial: tasa de crecimiento constante;
4. Fase de retardo: disminución de la tasa de crecimiento;
5. Fase estacionaria: tasa de crecimiento nula;
6. Fase de declive: tasa de crecimiento negativa.

En la fase exponencial se puede identificar el valor máximo que alcanza la tasa de crecimiento del cultivo (μ_{max}), la cual se mantienen constante a lo largo de esta fase. El tiempo que le toma al cultivo llegar, desde su concentración inicial de población (N_0), a tener esta tasa de crecimiento constante, se conoce como "tiempo de retardo" (λ) que es definida simplemente como la duración de la fase de retraso. En cierto periodo de tiempo la tasa de crecimiento comienza a disminuir hasta llegar a ser nula, siendo este el momento en el que el cultivo entra en su fase estacionaria con una concentración máxima de su población (N_{max}). En la figura 6.1 se pueden apreciar tres fases representativas del crecimiento de un monocultivo de la bacteria *E. coli* [23-25].

La evolución microbiana se puede expresar de manera general de la siguiente forma:

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = \mu_m \mu_Q(Q)\mu_P(P)\mu_S(S)N(t) \quad \text{con} \quad N(t=0) = N_0 \tag{6.1}$$

con las apropiadas ecuaciones diferenciales y condiciones iniciales, donde μ_m representa la tasa de crecimiento en la fase exponencial; $\mu_Q(Q)$ es una función que describe la fase de retraso; $\mu_S(S)$ describe el agotamiento del sustrato durante la evolución microbiana; $\mu_P(P)$ explica la inhibición del crecimiento debido a productos tóxicos; y $N(t)$ representa el nivel de población de los microorganismos con N_0 siendo la población al inicio del cultivo microbiano [24].

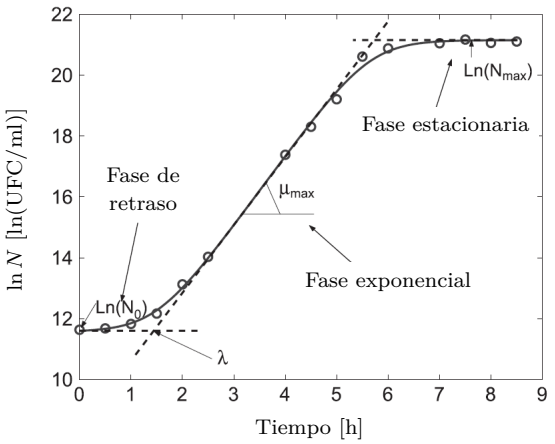


Figura 6.1: Modelo de crecimiento de Baranay y Roberts de una curva tipica de monocultivo: crecimiento de *E. coli* K12 a 35°C [24].

De acuerdo con Jiménez-Escamilla et al. [26] los productos metabólicos y el sustrato se comportan de acuerdo a las siguientes ecuaciones:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{1}{Y_{x/S}}\frac{\mathrm{d}N(t)}{\mathrm{d}t} \tag{6.2}$$

$$\frac{\mathrm{d}P}{\mathrm{d}t} = Y_{P/x}\frac{\mathrm{d}N(t)}{\mathrm{d}t} \tag{6.3}$$

donde $Y_{x/S}$ y $Y_{P/x}$ son sus respectivos rendimientos. Mientras que Kim et al. [27] menciona que la razón del consumo de glucosa en una fermentación de *K. oxytoca* puede ser descrita de una forma mas general a través de la siguiente ecuación:

$$-\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{1}{Y_N}\frac{\mathrm{d}N(t)}{\mathrm{d}t} + \sum_i \frac{1}{Y_{P_i}}\frac{\mathrm{d}P_i}{\mathrm{d}t} \tag{6.4}$$

donde Y_N y Y_{P_i} son los rendimientos estequiométricos de la biomasa y los productos respectivos. Estas tres ecuaciones dan indicios de que la obtención de productos se da de forma inversa al consumo del sustrato e incluso se puede seguir produciendo durante la fase estacionaria siempre y cuando la fermentación se esté llevando a cabo en un medio de cultivo continuo.

6.1 Modelo logístico

La elección de las funciones para describir el crecimiento de los microorganismos es de vital importancia para generar un buen modelo. Las curvas de crecimiento se utilizan en una amplia gama de disciplinas, siendo la biología una de ellas. Uno de los más utilizados es el modelo de crecimiento de tipo logístico, el cual es ampliamente utilizado en fenómenos que presentan un tipo de crecimiento sigmoidal, por ejemplo, el crecimiento poblacional de microorganismos

[25, 28].

Este modelo incorpora en la ecuación 6.1 una función de inhibición de tipo logística. De esta manera el modelo es capaz de describir la fase estacionaria de la tasa de crecimiento de un cultivo bacteriano [24]. El modelo logístico describe la tasa de crecimiento a partir de la siguiente ecuación:

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = \mu_m \left(1 - \frac{N(t)}{N_m}\right) N(t) \tag{6.5}$$

donde, $N(t)$ es la función que representa la concentración de los microorganismos al tiempo t; N_m representa el valor máximo de concentración que puede alcanzar la población de microorganismos; y μ_m la tasa máxima de crecimiento que se puede alcanzar. Una vez que $N(t)$ alcanza su valor máximo, la taza de crecimiento entra en su fase estacionaria.

$$1 - N(t)/N_m \tag{6.6}$$

La función de inhibición (ecuación 6.6) es una función monótona decreciente con valores entre uno y cero [24], tal como se muestra en la figura 6.2.

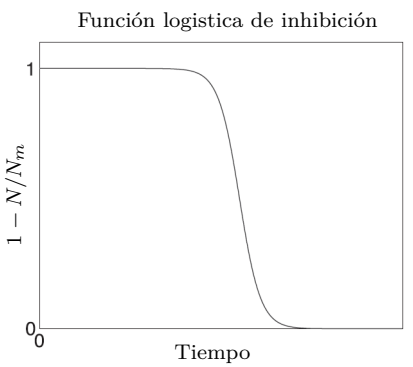


Figura 6.2: Curva de una función logística de inhibición [24].

6.2 Modelo de Gompertz

En 1825, Benjamin Gompertz, un matemático británico propuso un modelo como una forma de explicar la curva de mortalidad en los seres humanos. Este modelo, así como también sus diversas modificaciones, se han aplicado para describir crecimiento bacteriano, consumo de sustrato e incluso la taza de producción de biogas, entre muchas otras aplicaciones [28].

La forma más común de la ecuación de Gompertz está dada como

$$N(t) = N_m \cdot \exp\left[-\exp(\alpha - \beta t)\right] \tag{6.7}$$

donde $N(t)$ es la función que representa la concentración de los microorganismos al tiempo

t; N_m representa el valor máximo de concentración que puede alcanzar la población de microorganismos; α y β están dados de la siguiente forma para un problema de crecimiento bacteriano

$$\alpha = \frac{\mu_m \cdot \mathrm{e}}{N_m}\lambda + 1 \tag{6.8}$$

$$\beta = \frac{\mu_m \cdot \mathrm{e}}{N_m} \tag{6.9}$$

en donde λ representa el tiempo de retardo del crecimiento poblacional del cultivo de microorganismos; e es la base de logaritmo natural [25, 28].

Tal como se muestra en la figura 6.3 los modelos de crecimiento ayudan a describir el comportamiento de ciertos fenómenos, y dependiendo de la naturaleza del fenómeno estudiado, algunos modelos se ajustaran mejor que otros.

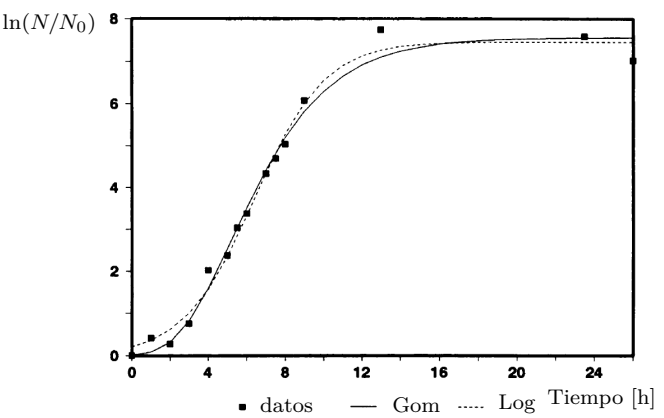


Figura 6.3: Curva de crecimiento de *L. plantarum* a 40°C ajustada utilizando el modelo de Gompertz (Gom) y el modelo logístico (Log) [25].

6.3 Comparación entre modelos

Se han desarrollado diversos modelos para describir curvas de crecimiento de tipo sigmoidal (en la tabla 6.1 se muestran algunos de estos modelos presentes en la literatura) buscando reducir los datos medidos a un cierto número de parámetros. Chu [28] realiza una comparación en donde muestra que para el caso de la adsorción de cromo (IV) por médula de fibra de coco la ecuación de Gompertz se ajusta mejor a los datos que la ecuación logística (figura 6.4a), y por otro lado, para el caso de la adsorción de azul de metileno por carbón activado, tanto la ecuación logística como la ecuación de Gompertz no logran realizar un ajuste realmente eficiente (figura 6.4b) [25, 28].

Debido a la asimetría de los datos en la figura 6.4b, la ecuación de Gompertz no realiza un buen ajuste. Chu [28] realiza un par de modificaciones en esta ecuación, obteniendo dos ecuaciones

Tabla 6.1: Modelos de crecimiento modificadas en términos de la tasa de crecimiento [25].

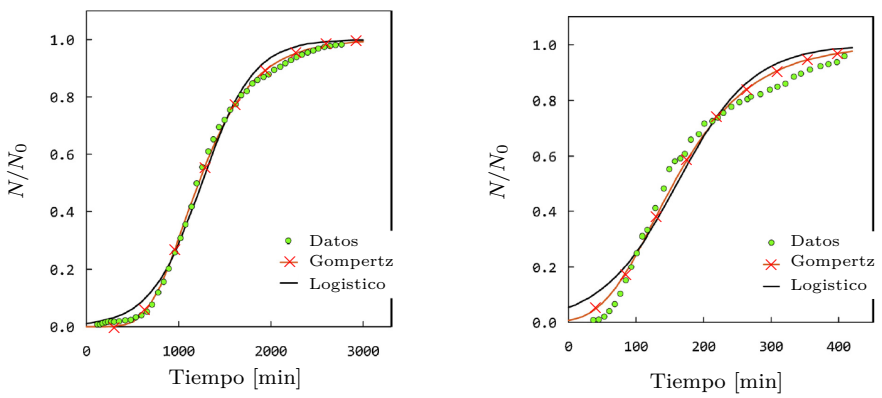
Modelo	Ecuación modificada
Logístico	$N(t)=\frac{N_m}{\left\{1+\exp\left[\frac{4\cdot\mu_m}{N_m}\cdot(\lambda-t)+2\right]\right\}}$
Gompertz	$N(t)=N_m\cdot\exp\left\{-\exp\left[\frac{\mu_m\cdot e}{N_m}\cdot(\lambda-t)+1\right]\right\}$
Richards	$N(t)=N_m\cdot\left\{1+\nu\cdot\exp(1+\nu)\cdot\exp\left[\frac{\mu_m}{N_m}\cdot(1+\nu)\cdot(1+\nu^{-1})\cdot(\lambda-t)\right]\right\}^{-\nu^{-1}}$
Stannard	$N(t)=N_m\cdot\left\{1+\nu\cdot\exp(1+\nu)\cdot\exp\left[\frac{\mu_m}{N_m}\cdot(1+\nu)\cdot(1+\nu^{-1})\cdot(\lambda-t)\right]\right\}^{-\nu^{-1}}$
Schnute	$N(t)=\left(\mu_m\cdot\frac{(1-\alpha)}{N_m}\right)\cdot\left[\frac{1-\alpha\cdot\exp(N_m\cdot\lambda+1-\alpha-\alpha-N_m\cdot t)}{1-\alpha}\right]^{1/\alpha}$

más que solucionan este problema. La ecuación *Power law Gompertz* (ecuación 6.10) y la ecuación *Log-Gompertz* (ecuación 6.11) se ajustan mejor que la ecuación de Gompertz no modificada y que la ecuación logística para el caso de la adsorción de azul de metileno por carbón activado, tal como se muestra en la figura 6.5 [28].

$$N(t) = N_m \exp\left\{-\exp\left[\frac{\mu_m e}{N_m}(\lambda - t^n) + 1\right]\right\} \tag{6.10}$$

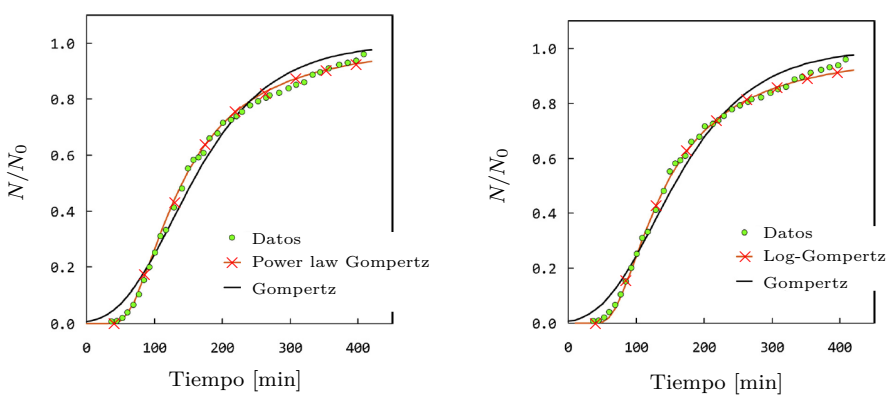
$$N(t) = N_m \exp\left\{-\exp\left[\frac{\mu_m e}{N_m}(\lambda - \ln t) + 1\right]\right\} \tag{6.11}$$

Por estos motivos se debe tener especial cuidado al momento de seleccionar el modelo con el que se trabajará.



(a) (b)

Figura 6.4: Ajuste mediante la ecuación logística y de Gompertz de los datos de la (a) adsorción de cromo (IV) por medula de fibra de coco y la (b) adsorción de azul de metileno por carbón activado [28].



(a) (b)

Figura 6.5: Comparación entre el ajuste mediante la ecuación de Gompertz y sus ecuaciones modificadas, la ecuación (a) Power law Gompertz y la ecuación (b) Log-Gompertz; de los datos de la adsorción de azul de metileno por carbón [28].

Parte III

Metodología

Capítulo 7

Diseño y desarrollo

7.1 Obtención de datos.

La empresa Labcitec S. A. de C. V. proporcionó acceso a un conjunto de 33 archivos con el registro del monitoreo de cultivos microbianos de un mismo organismo, específicamente, la bacteria *E. coli*. Los datos fueron recolectados en un periodo de dos años, llevándose a cabo en distintos meses, a diferentes intervalos de tiempo.

Todos los medios de cultivo tenían características similares como el control de pH, temperatura, agitación, entre otros, de tal forma que las bacterias en cada cultivo se desarrollaran bajo las mismas condiciones. Los cultivos fueron monitoreados mediante espectroscopia Raman y HPLC como método de análisis de laboratorio.

Los archivos consistían de tres secciones. La primera sección contiene el registro de los espectros Raman obtenidos a partir del monitoreo en línea. Estos espectros fueron recolectados en intervalos de alrededor de setenta segundos a lo largo de la evolución del cultivo. La segunda sección contiene el registro de los resultados de las muestras de laboratorio que fueron tomadas aproximadamente cada dos horas. Esta sección recopila los datos de la concentración de los organismos presentes en el cultivo, así como del sustrato y de diferentes productos. La tercera sección relaciona los valores de laboratorio con la suma de tres espectros Raman correspondientes a un tiempo posterior en que se tomó la muestra. Estos archivos fueron utilizados para desarrollar los modelos de regresión de acuerdo al diagrama de la figura 2.2.

En el apéndice A se ilustra el cambio de los espectros Raman obtenidos del monitoreo de un cultivo. Los cambios de intensidad en determinados números de onda indican el cambio de la concentración de algún compuesto.

7.2 Modelos de regresión.

Para obtener los valores de concentración de la biomasa, la glucosa (sustrato), y el ácido orgánico (producto) se utilizó la tercera sección de cada uno de los archivos. Con estos datos se generaron los modelos de regresión que relacionan los espectros Raman con un valor de concentración correspondiente para cada uno de los compuestos.

Estos modelos de regresión fueron generados mediante el uso de un software de análisis multivariable. En el mercado existe una gran cantidad de opciones disponibles, sin embargo, para este proyecto se optó por la utilización del software SIMCA P en su version 14.1 cuyo acceso fue proporcionado por la misma empresa.

Selección del modelo.

Este software proporciona diferentes tipos de modelos. Se utilizó un modelo PLS. Tiene sentido pensar en una correlación lineal dada la naturaleza del problema pues de acuerdo con la literatura la intensidad de la dispersión es directamente proporcional a la concentración de moléculas presentes en la muestra [2]. De acuerdo con lo anterior, el modelo PLS es una buena opción que permite asignar valores de concentración a cada compuesto de acuerdo al espectro de dispersión Raman que se presente.

Datos de entrenamiento.

De los 33 archivos disponibles se seleccionaron únicamente 21 de ellos como el conjunto de entrenamiento para los modelos de regresión. Algunos contenían únicamente información correspondiente a la fase de retraso del cultivo o simplemente ruido en la sección de los espectros. Otros contenían errores en los valores de laboratorio, y algunos otros no correspondían a un medio de cultivo en continuo, lo cual claramente no concordaba con las características del resto de cultivos. Por estas razones, 12 de los archivos fueron descartados debido a no contar con información suficientemente relevante o demasiados errores en los datos.

Preprocesamiento de los espectros.

Como parte del preprocesamiento se eliminaron espectros "basura" que estaban en el conjunto de datos. Estos espectros por alguna razón ajena al fenómeno de dispersión presentaban inconsistencias con el resto del conjunto. Algunos de ellos ni siquiera contenían información y todas las intensidades marcaban valores cercanos a cero o simplemente contenían ruido, posiblemente por alguna alteración del entorno.

El software permite aplicar diferentes filtros espectrales en la parte del preprocesamiento de los espectros Raman. Se aplicó un filtro Savitzky-Golay como método de suavizado, este filtro realiza el calculo a partir de submodelos cuadráticos móviles, cada uno de 45 puntos de datos de longitud, excluyendo los bordes. Posteriormente, siguiendo las recomendaciones de la literatura [29], se aplico el filtro de primera derivada que realiza los cálculos a partir de submodelos móviles, cada uno de 45 puntos de datos de longitud, con una distancia entre cada punto de datos igual a 1, excluyendo los bordes. Por último se aplicó un filtro SNV (Standard Normal Variate) como método de normalización.

Datos atípicos.

El software proporciona diferentes herramientas que ayudan a identificar aquellos datos atípicos que puedan causar que el modelo no realice buenas predicciones. Las gráficas de puntuación

muestran como se sitúan entre si el espacio X (condiciones X) y los valores de respuesta. Con ayuda del gráfico de dispersión de puntuaciones (figura 7.1) se determinaron aquellas observaciones consideradas como valores atípicos mediante una elipse de confianza basada en la T^2 de Hotelling, por defecto a un nivel de significación de 0.05.

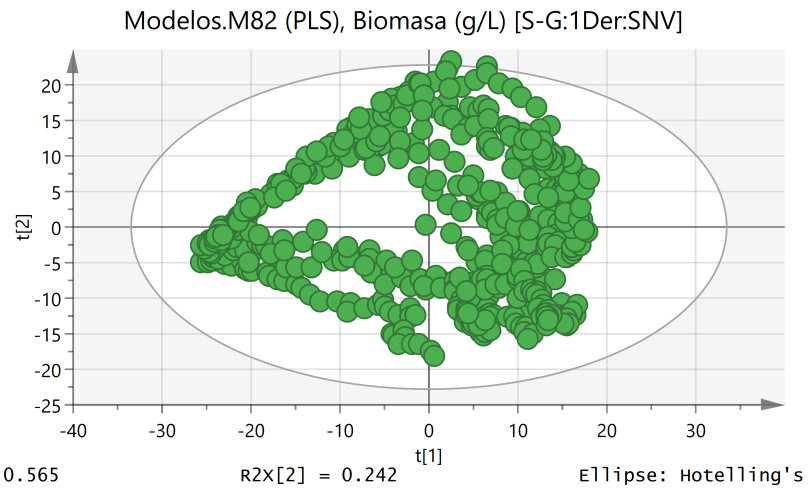


Figura 7.1: Gráfica de dispersión de puntuaciones del modelo M82 generado para predecir los valores de concentración de la biomasa.

Otra herramienta que se utilizó es el gráfico de rango T^2 de Hotelling (figura 7.2) que muestra la distancia desde el origen en el plano del modelo (espacio de puntuación) para cada una de las observaciones. Además, esta gráfica presenta dos límites marcados, los valores por encima del primer limite (nivel 0.05) se pueden considerar como valores sospechosos, mientras que los que están por encima del segundo limite (nivel 0.01) pueden realmente ser considerados como valores atípicos peligrosos.

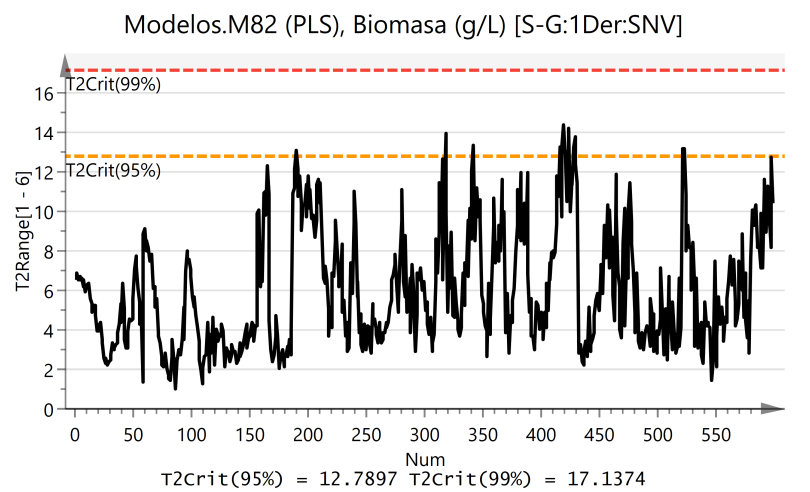
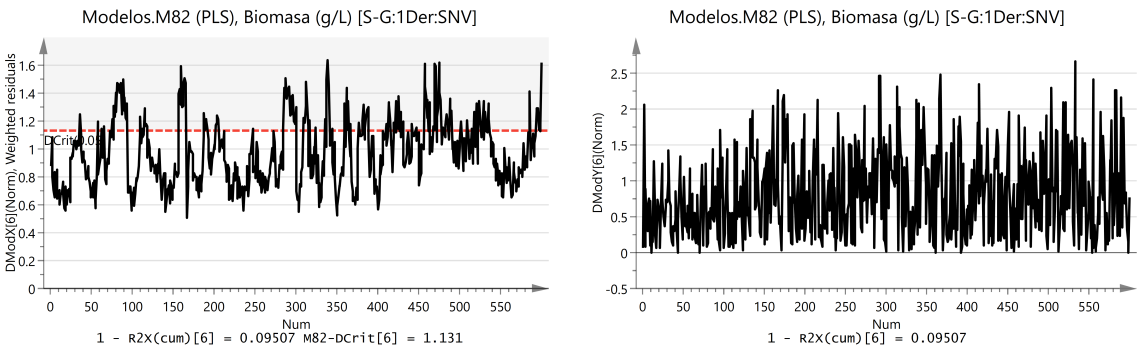


Figura 7.2: Gráfica de rango T^2 de Hotelling del modelo M82.

Así mismo se hizo uso de las gráficas de distancia al modelo (figura 7.3), las cuales proporcio-

nan una estimación de a qué distancia del plano del modelo, en el espacio X o Y, se encuentra cada observación. Los valores grandes o por encima de un cierto limite indican la presencia de datos atípicos en el espacio X o Y.



(a) (b)

Figura 7.3: Gráficas de distancia al modelo (a) en es espacio X y (b) en el espacio Y para el modelo M82.

Estas observaciones marcadas como datos atípicos pudieron generarse por errores que van desde algún desajuste del sensor, algún dato erróneo de laboratorio o alguna medición que no corresponde al espectro que se le fue asignado. Para poder obtener un modelo lo suficientemente confiable para aplicarlo al monitoreo de un proceso biológico es necesario generar diferentes modelos eliminando aquellos valores atípicos o cambiando el número de componentes del modelo buscando reducir lo más posible dos de las métricas proporcionadas por el software, el error cuadrático medio de estimación (RMSEE por sus siglas en inglés), y el error cuadrático medio de validación cruzada (RMSEcv por sus siglas en inglés).

Se generaron 3 modelos de regresión diferentes. El primer modelo se enfocó en la concentración de microorganismos presentes en el medio de cultivo (biomasa); el segundo modelo, en la concentración del sustrato (glucosa); y el tercer modelo en la concentración de los productos (ácido orgánico) generados por estas bacterias.

7.3 Validación de los modelos de regresión.

Debido a la alta sensibilidad de sensor, es comprensible esperar una alta variabilidad de los valores de concentración calculados por el modelo de regresión a lo largo del tiempo. Es por este motivo que se consideró válido un modelo que a pesar de que sus mediciones no sean exactas, cumplan por lo menos con un cierto margen de error, lo mínimo posible.

Para poder validar estos tres modelos se utilizó un archivo que no fue implementado en el entrenamiento del modelo, además de otros tres archivos que si fueron utilizados para el entrenamiento. Se utilizaron estos archivos ya que se hizo uso de los espectros de la primera sección correspondiente a los espectros Raman recolectados en cada cultivo buscando reducir

el error cuadrático medio de predicción (RMSEP por sus siglas en inglés), además de optimizar el coeficiente de determinación R^2 con respecto a la predicción.

Para obtener el coeficiente de determinación R^2 relacionar los valores de concentración de laboratorio con los espectros a los que fueron relacionados en la tercera sección. Esto se logró mediante un script de Python que relacionaba los espectros utilizados con el valor correspondiente de concentración de los tres componentes, obteniendo en total 1958 observaciones con valores de concentración de biomasa, ácido orgánico y/o glucosa. Estos datos fueron considerados como el set de validación para los modelos de regresión.

7.4 Modelo de predicción.

Una vez generados los modelos de regresión para cada compuesto, se aplicaron a cada uno de los conjuntos de espectros Raman para cada uno de los archivos utilizados para el entrenamiento de estos modelos, obteniendo las cinéticas de cada uno de sus componentes. El conjunto de datos de todas estas cinéticas fue utilizado para el diseño y entrenamiento del modelo de predicción.

Selección del modelo.

Estas cinéticas se pueden ver como series temporales, en las cuales las redes neuronales recurrentes LSTM han sido aplicadas de manera eficiente [20, 30]. El problema que se presenta es realizar una predicción del valor de concentración en un tiempo posterior a partir de una ventana de datos dada.

Se implementó un modelo de red neuronal LSTM de disparo único, con una ventana de datos y un tiempo de predicción de 12 horas cada uno.

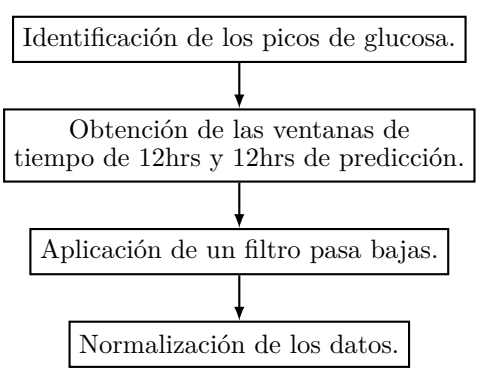


Figura 7.4: Esquema del preprocesamiento de las cinéticas para el entrenamiento del modelo de predicción.

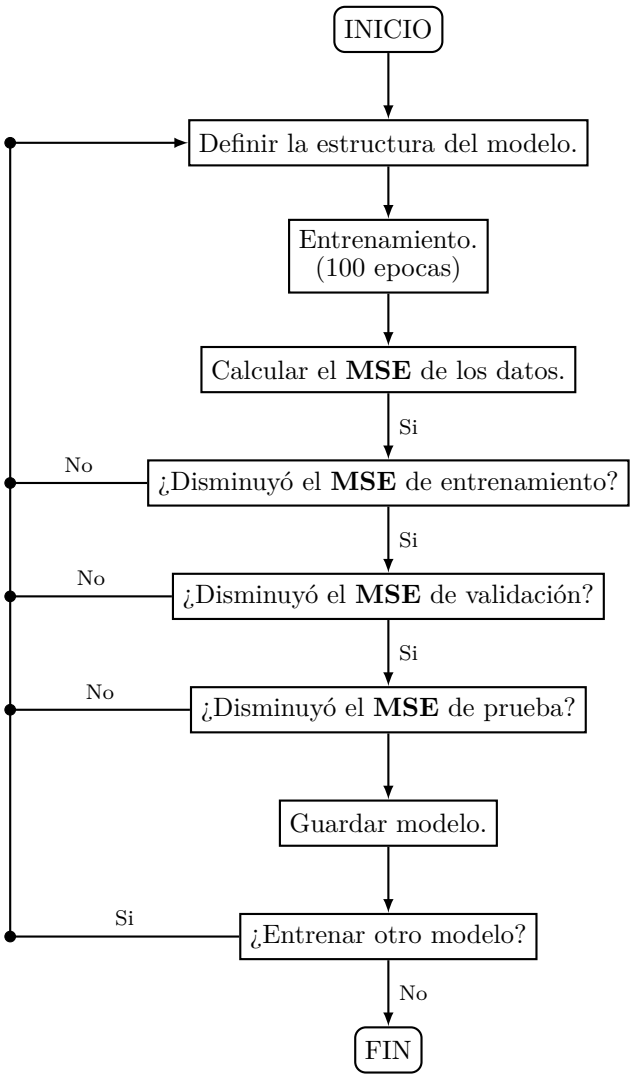


Figura 7.5: Esquema de la búsqueda de malla para el entrenamiento del modelo.

Datos de entrenamiento.

Se hizo uso de 19 archivos de datos, de los cuales se obtuvieron las ventanas de datos que se implementaron para el entrenamiento de la red neuronal. La obtención de estas ventanas se llevó a cabo de acuerdo al diagrama mostrado en la figura 7.4 para cada uno de los 19 archivos.

El filtro pasa bajas implementado fue diseñado para eliminar la alta variabilidad de la señal de los datos de los modelos de regresión.

Entrenamiento del modelo.

Una vez definido el modelo de red neuronal LSTM se realizó una búsqueda de malla para encontrar los mejores parámetros de capas y el número de neuronas para el modelo. El conjunto de datos que sería utilizado, se separó en un 70 % para el conjunto de entrenamiento, 20 % para el conjunto de validación y el restante 10 % para el conjunto de prueba. El proceso para determinar el mejor modelo es mostrado por el diagrama de flujo de la figura 7.5.

Validación del modelo.

Para realizar la validación del modelo se hizo uso de tres cinéticas que no fueron implementadas en el entrenamiento de ninguno de los modelos (regresión y predicción). Esto se realizó con el fin de observar el desempeño del modelo para realizar predicciones con datos no antes vistos.

7.5 Simulación de un monitoreo.

Los dispositivos de monitoreo en tiempo real no cuentan con la posibilidad de acceder a los datos de concentración en tiempo real para su manipulación, lo cual representó un problema para realizar la implementación del modelo durante el monitoreo. Por esta razón se optó por simular un monitoreo en tiempo real.

Para realizar esta simulación se utilizó uno de los archivos que contenían varias muestras de laboratorio, con el fin de obtener la mayor cantidad de datos. Este archivo pertenece al segundo conjunto de validación que fue utilizado en el modelo de predicción.

Cada una de las predicciones fueron realizadas en intervalos de 1 hora durante todo el proceso. Para observar el desempeño del modelo de predicción en la simulación, se obtuvieron el error absoluto medio (**MAE**), el error absoluto medio porcentual (**MAPE**) y el error cuadrático medio (**RMSE**) de los datos de predicción calculados por el modelo con respecto a los datos correspondientes del modelo de regresión, los datos de laboratorio y los datos de un modelo logístico generado a partir de los datos de laboratorio. Los errores fueron calculados para los datos hasta 3, 6, 9 y 12 horas de los datos de predicción.

Parte IV

Resultados y Conclusiones

Capítulo 8

Modelos de regresión.

8.1 Disposición de los datos.

En la primera sección de los archivos se puede encontrar el registro de los espectros Raman recolectados durante el monitoreo del cultivo microbiano. En esta sección los datos se encuentran distribuidos como en la tabla 8.1. En la primer columna se registra una etiqueta única para cada espectro, la cual contiene tanto información del cultivo al que pertenece, como de la fecha y hora en que fue recolectado. La segunda columna muestra la fecha y hora con un formato mas legible. Por último, las 2200 columnas restantes contienen la información de la intensidad correspondiente a cada número de onda, el intervalo del número de onda va desde los $200cm^{-1}$ hasta los $2400cm^{-1}$.

La segunda sección corresponde al reporte de laboratorio, este reporte se presenta como una

Tabla 8.1: Disposición de los datos de los espectros Raman del cultivo bacteriano.

Etiqueta	**Fecha:Hora**	**Intensidad del número de onda ($\bar{\nu}$) en el espectro Raman**				
		200	**201**	...	**2399**	**2400**
ABC000_00000000-000000	DD/MM/AAAA HH:MM:SS	7424.12	7343.21	...	305.561	244.541
ABC000_00000000-000001	DD/MM/AAAA HH:MM:SS	10158.9	10036.7	...	450.689	422.463
ABC000_00000000-000002	DD/MM/AAAA HH:MM:SS	11594.7	11448.7	...	700.269	722.019
⋮	⋮	⋮	⋮	⋱	⋮	⋮
ABC000_00000000-002221	DD/MM/AAAA HH:MM:SS	7463.7	7405.58	...	490.325	397.028
ABC000_00000000-002222	DD/MM/AAAA HH:MM:SS	7490.46	7403.72	...	593.689	640.831
ABC000_00000000-002223	DD/MM/AAAA HH:MM:SS	7481.02	7414.97	...	927.462	960.302

Tabla 8.2: Disposición de los datos en el reporte de laboratorio.

Fecha:Hora	**Horas**	**ABC000 - VAR1(g/L)**	...	**ABC000 - VAR7(g/L)**
DD/MM/AAAA HH:MM	**0**	26.53	...	0.01
DD/MM/AAAA HH:MM	**12**	27.8	...	0.01
DD/MM/AAAA HH:MM	**16**	28.4	...	0.02
⋮	⋮	⋮	⋱	⋮
DD/MM/AAAA HH:MM	**72**	12.97	...	27.94
DD/MM/AAAA HH:MM	**74**	11.3	...	21.03
DD/MM/AAAA HH:MM	**77**	8.13	...	19.36

tabla de datos (tabla 8.2). La primer columna de esta tabla contiene la fecha y hora en que se realizó la toma de muestras del cultivo. Desde la primera muestra recolectada comienza un conteo de tiempo del cual se registran las horas transcurridas para cada muestra en la segunda columna. El resto de columnas, etiquetadas con el código del cultivo del cual fueron recolectadas y la sustancia que fue analizada, contiene la información de la concentración de la sustancia correspondiente.

La tercera sección realiza una relación entre la concentración de las sustancias del reporte de laboratorio y la suma de tres espectros Raman consecutivos recolectados después de la fecha y hora indicada. Tal como se muestra en la tabla 8.3 la fecha y hora se registran en la primer columna. Las siguientes seis columnas consecutivas registran los valores de concentración de las respectivas sustancias analizadas por el laboratorio. Por último, las siguientes 2200 columnas registran los valores correspondientes a la suma de los valores de intensidad de cada uno de los

Tabla 8.3: Disposición de los datos de espectros relacionados a los valores de concentración correspondiente utilizados para el conjunto de entrenamiento.

Fecha:Hora	**VAR1(g/L)**	...	**VAR6(g/L)**	**Intensidad del número de onda ($\bar{\nu}$) en el espectro Raman**		
				200	...	**2400**
DD/MM/AAAA HH:MM	27.8	...	0.01	20019.94	...	751.1332
DD/MM/AAAA HH:MM	28.4	...	0.02	37942.55	...	2509.541
DD/MM/AAAA HH:MM	27.27	...	0.25	40135.8	...	1819.788
⋮	⋮	⋱	⋮	⋮	⋱	⋮
DD/MM/AAAA HH:MM	14.63	...	25.38	21653.9	...	1678.537
DD/MM/AAAA HH:MM	12.97	...	27.94	21962.18	...	1857.575
DD/MM/AAAA HH:MM	11.3	...	21.03	22226.65	...	1469.552

Tabla 8.4: Disposición de los datos de espectros relacionados a los valores de concentración correspondiente utilizados para el conjunto de validación.

Etiqueta	**Fecha:Hora**	...	**VAR3(g/L)**	**Intensidad del número de onda ($\bar{\nu}$) en el espectro Raman**	
				...	**2400**
ABC000_00000000-000000	**DD/MM/AAAA HH:MM**	...	0.01	...	751.1332
ABC000_00000000-000001	**DD/MM/AAAA HH:MM**	...	0.02	...	665.541
ABC000_00000000-000002	**DD/MM/AAAA HH:MM**	...	0.25	...	664.788
⋮	⋮	⋮	⋱	⋮	
ABC000_00000000-002221	**DD/MM/AAAA HH:MM**	...	25.38	...	599.537
ABC000_00000000-002222	**DD/MM/AAAA HH:MM**	...	27.94	...	512.575
ABC000_00000000-002223	**DD/MM/AAAA HH:MM**	...	21.03	...	640.552

números de onda correspondientes. Esta sección fue utilizada en el conjunto de entrenamiento de los modelos de regresión.

En la tabla 8.4 se muestra como fueron ordenados los datos de la glucosa, el ácido orgánico y la biomasa con los espectros Raman correspondientes de la primera sección de los archivos, para formar el conjunto de validación de los modelos de regresión. La primer columna contiene la etiqueta correspondiente del espectro Raman, la segunda muestra la fecha y hora en que fue recolectada. Las siguientes tres columnas contienen los datos de laboratorio y el resto los valores de intensidad del espectro para cada número de onda.

8.2 Parámetros.

Después de generar varios modelos se obtuvieron tres modelos PLS, el modelo M21, M58 y M82 que determinan la concentración de la glucosa, el ácido orgánico y la biomasa, respectivamente. Cada uno de los modelos está compuesto por un determinado número de componentes (A) que son generados a partir de una cantidad de observaciones (N) de las cuales algunas son descartadas por ser observaciones atípicas. Los parámetros de cada uno de estos modelos se presentan en la tabla 8.5.

Tabla 8.5: Parámetros de los mejores modelos de regresión generados.

Modelo	**Tipo**	**A**	**N**	**Etiqueta**
M21	PLS	4	906	Glucosa (g/L)
M58	PLS	4	402	Á. Orgánico (g/L)
M82	PLS	6	600	Biomasa (g/L)

A: Componentes; N: Observaciones.

8.3 Validación.

Para la determinación de un modelo aceptable se buscó optimizar sus parámetros mediante la búsqueda de observaciones atípicas utilizando las herramientas proporcionadas por el software de análisis multiplicable SIMCA P, como la gráfica de dispersión de puntuaciones (figura 7.1), la gráfica de rango T^2 de Hotelling (figura 7.2) y las gráficas de distancia al modelo en sus respectivos espacios (figura 7.3).

Tabla 8.6: Métricas de los mejores modelos de regresión generados.

Modelo	**RMSEE**	**RMSEcv**	**RMSEP**	$\mathbf{R^2}$
M21	2.7103	2.7333	4.6499	0.8847
M58	2.4106	2.4954	5.1505	0.9074
M82	2.2461	2.4502	4.8514	0.7487

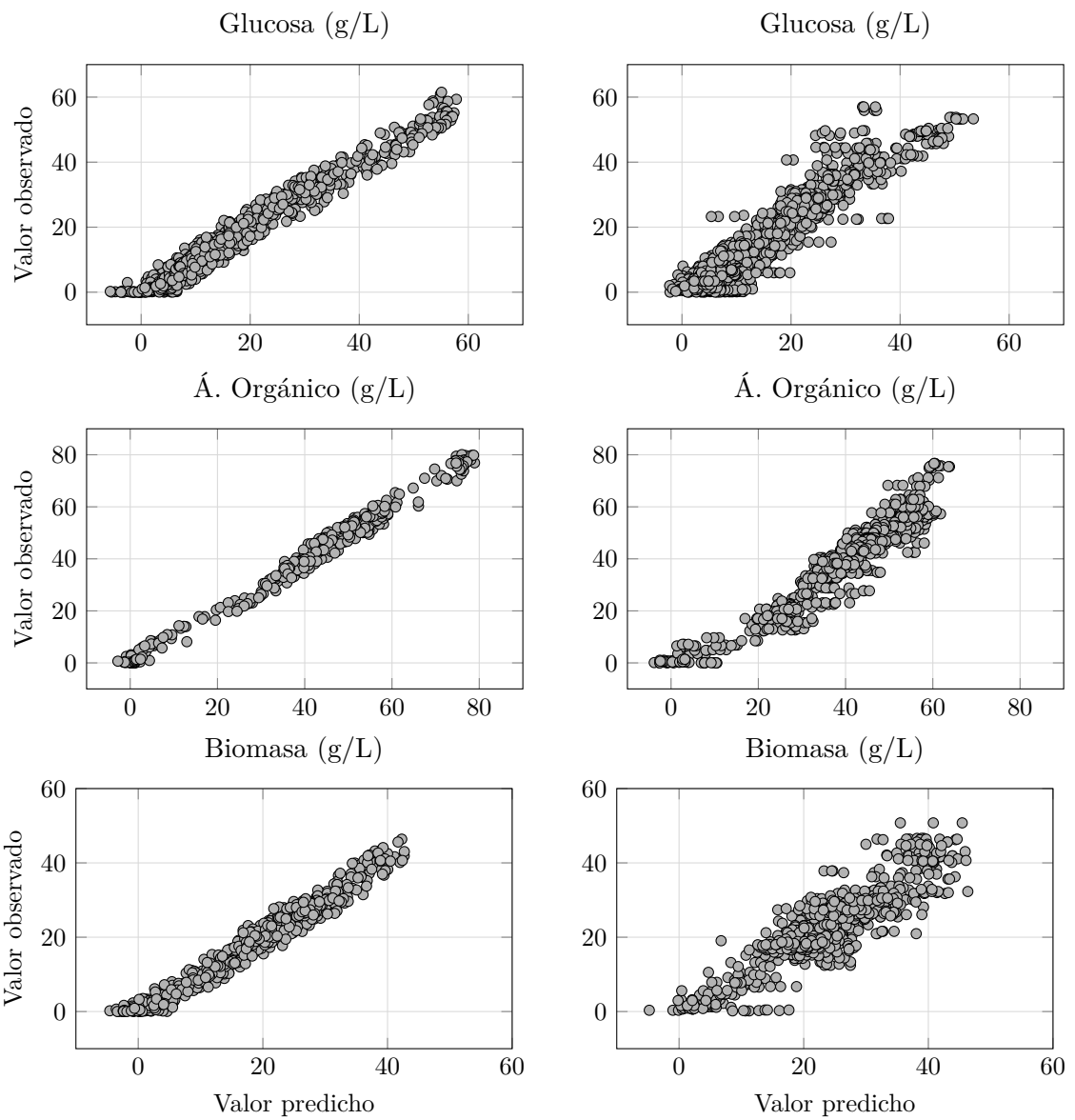


Figura 8.1: Gráficas de dispersión de los datos observados contra los datos predichos por los modelos M21, M58 y M82 para la glucosa, el ácido orgánico y la biomasa, respectivamente. Los datos de la izquierda son calculados respecto al set de entrenamiento mientras que los datos de la derecha fueron calculados respecto al set de validación.

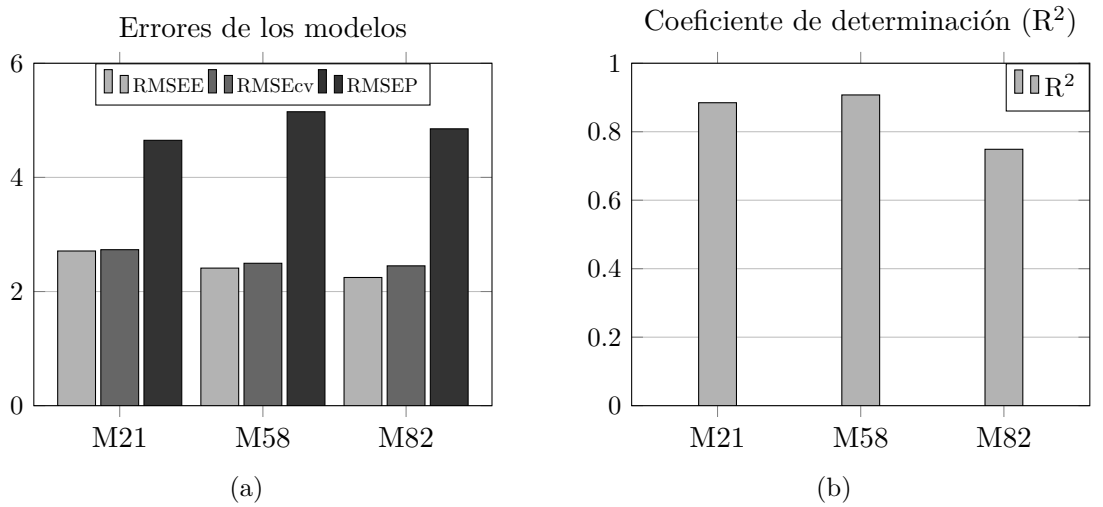


Figura 8.2: (a) Gráfica de barras de las métricas RMSEE, RMSEcv y RMSEP, y (b) gráfica de barras del coeficiente de determinación R^2 de los modelos de regresión.

Para la validación de cada uno de los modelos de regresión se buscó reducir los valores de las métricas, RMSEE, RMSEcv, RMSEP, y el coeficiente de determinación R^2. Los valores de estas métricas se presentan en la tabla 8.6.

En la figura 8.1 se pueden observar los valores observados (datos de laboratorio) contra los valores predichos por los modelos sobre el conjunto de entrenamiento y el de validación. Además, con ayuda de las gráficas de barras de la figura 8.2 se puede observar el desempeño de cada uno de estos modelos al comparar las métricas RMSEE, RMSEcv y RMSEP además del coeficiente de correlación (R^2).

Tabla 8.7: Disposición de los datos de espectros relacionados a los valores de concentración correspondiente utilizados para el conjunto de validación. Los modelos M21, M58 y M82 calculan los valores de concentración de la glucosa, el ácido orgánico y la biomasa respectivamente en gramos por litro.

Obs ID (Primary)	**Obs ID (Tiempo)**	**M21**	**M58**	**M82**
ABC000...0000	**DD/M...:MM**	48.2247	-4.73619	0.329042
ABC000...0001	**DD/M...:MM**	49.0166	-4.11807	-1.30494
ABC000...0002	**DD/M...:MM**	49.3932	-3.41667	-3.07327
⋮	⋮	⋮	⋮	⋮
ABC000...2222	**DD/M...:MM**	-1.20876	43.1841	23.9698
ABC000...2223	**DD/M...:MM**	2.67823	44.3037	24.5867

Al aplicar los modelos a los datos de los espectros Raman de la primera sección de los archivos (tabla 8.1) se obtuvo un conjunto de datos como los mostrados en la tabla 8.7. En la cual la primera columna registra las etiquetas de cada uno de los espectros correspondientes. En la segunda columna se registra la fecha y la hora en se obtuvo el espectro. Las siguientes

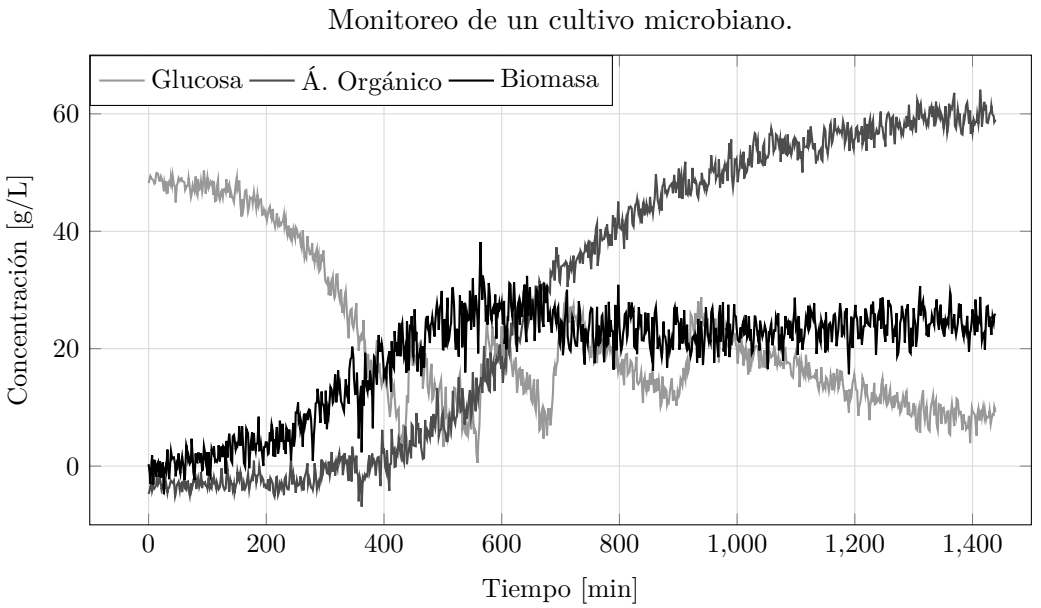


Figura 8.3: Cinética microbiana de un cultivo microbiano monitoreado mediante espectroscopia Raman. Los valores de concentración son obtenidos al aplicar el modelo correspondiente al espectro Raman obtenido durante el monitoreo.

columnas corresponden a los valores de concentración (en gramos por litro) calculados por cada uno de los modelos correspondientes.

Estos modelos relacionan la intensidad de las longitudes de onda en los espectros con un valor de concentración correspondiente. Durante el monitoreo de un cultivo microbiano se obtiene un conjunto de espectros Raman obtenidos en diferentes instantes de tiempo. Al aplicar estos modelos a cada uno de los espectros se obtiene un conjunto de valores correspondientes a la concentración de cada uno de los compuestos en el tiempo correspondiente en que se obtuvo el espectro Raman. Al generar la gráfica de estos valores con respecto al tiempo correspondiente del muestreo de los espectros se puede obtener la cinética del cultivo microbiano (figura 8.3).

Capítulo 9

Modelos de predicción.

9.1 Entrenamiento.

Para la configuración del modelo de predicción se optó por implementar un modelo LSTM de "disparo único", en donde el modelo realiza una predicción de secuencia completa en un solo paso a partir de una ventana de datos con un tamaño definido. Para entrenar el modelo de la mejor forma posible fue necesario obtener las ventanas de entradas y salidas, y eliminar aquellas que terminaran con un cambio brusco en los datos de la glucosa, lo que podría causar que el modelo no realice correctamente las predicciones.

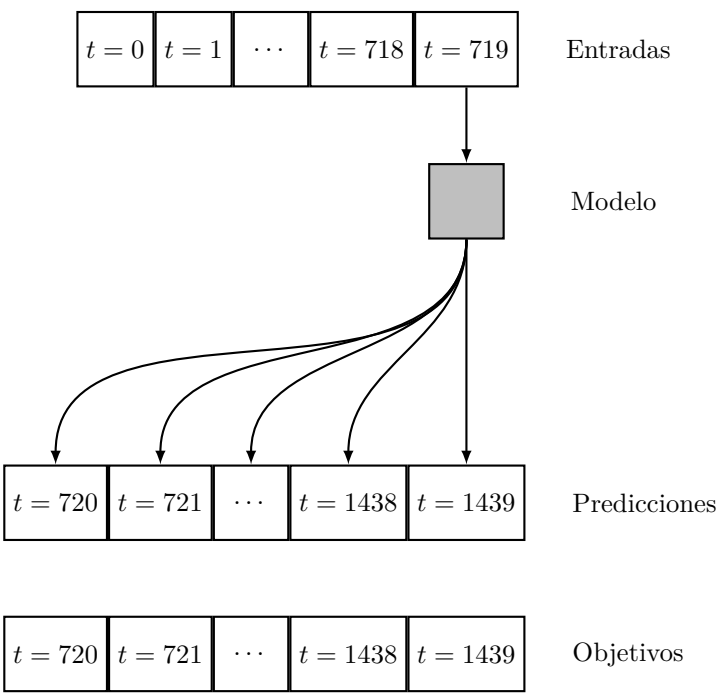


Figura 9.1: Diagrama de la configuración del modelo de predicción de "disparo único".

La ventana de datos de entrada se configuró para analizar un total de 12 horas, la cual contiene información de la concentración de la glucosa, del ácido orgánico y de la biomasa minuto a minuto desde el comienzo del monitoreo. Esta matriz de concentración es formada por 3 vectores de 720 datos cada uno y se obtiene después de realizar el preprocesamiento de los datos. Primero se aplica un filtro pasa bajas para eliminar la dispersión de los datos y suavizar las curvas de concentración. Posteriormente los valores negativos que pudieran aparecer son intercambiados por ceros. Por último esta matriz es normalizada utilizando un valor máximo de 100 gramos por litro.

Por otro lado la ventana de datos de salida aunque también se configuró para contener hasta 12 horas de datos de predicción, únicamente contiene información acerca de la concentración del ácido orgánico. Para obtener estas ventanas de datos, fue necesario aplicar el preprocesamiento de los datos a un total de 1440 datos, lo cual son 24 horas de información del ácido orgánico, partiendo desde el primer dato correspondiente al ácido orgánico en la matriz de concentración de la ventana de entrada, y tomando los últimos 720 datos, lo cual corresponde las siguientes 12 horas de información posteriores al ultimo dato de la ventana de entrada.

Para determinar los mejores parámetros como el número de capas, el número de neuronas por capa y el tamaño de lote de entrenamiento, se realizó un total de 7 entrenamientos en los cuales se llevó a cabo una búsqueda en malla de los parámetros óptimos para el modelo LSTM utilizando una tarjeta gráfica NVIDIA GTX 1050 Ti y las librerías *TensorFlow* y *scikit-learn* de Python.

$$MSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{9.1}$$

Se seleccionó el error cuadrático medio (MSE por sus siglas en ingles) como función de costo, la cuál esta dada como se muestra en la ecuación 9.1 en donde $\hat{y}$ representa el valor predicho por el modelo, mientras que y es el valor objetivo que pertenece a la ventana de predicción definida para el entrenamiento. Además, se utilizó el método de parada anticipada (early stopping) para evitar un sobre ajuste del modelo. Como resultado de estos entrenamientos se obtuvieron 7 modelos cuyos parámetros se muestran en la tabla 9.1.

Con una configuración inicial de 100 épocas de entrenamiento se recopiló el MSE de cada uno de los conjuntos al aplicarles el modelo una vez entrenado y se recopiló cada uno de estos valores en la tabla 9.2. Tal como se observa en la gráfica de barras de la figura 9.2 el mejor modelo de los generados en los 7 entrenamientos realizados es el modelo M25 que obtuvo un error de entrenamiento de $5.01E-04$, un error de validación de $5.05E-04$ y un error de prueba de $5.05E-04$.

Tabla 9.1: Parámetros de los mejores modelos LSTM generados.

Modelo	**capas**	**unidades**	**tamaño de lote**
M5	1	80	35
M19	2	80	35
M24	2	80	22
M25	2	80	24
M33	2	82	35
M37	2	92	35
M41	3	64	35

Tabla 9.2: Errores de entrenamiento (MSE_train), validación (MSE_val) y prueba (MSE_test) de los modelos LSTM.

Modelo	MSE_train	MSE_val	MSE_test
M5	1.26E-03	1.28E-03	1.30E-03
M19	6.39E-04	6.21E-04	6.52E-04
M24	8.02E-04	7.61E-04	8.25E-04
M25	5.01E-04	5.05E-04	5.05E-04
M33	8.33E-04	8.12E-04	8.42E-04
M37	7.35E-04	7.12E-04	7.81E-04
M41	6.63E-04	6.43E-04	7.02E-04

Al realizar la validación del modelo M25 con otras tres cinéticas nunca antes vistas por el modelo se obtuvo un error de predicción de $7.13E-04$ para la primera cinética, de $1.12E-02$ para la segunda y de $9.35E-04$ para la tercera. El error mas grande obtenido se dio debido a una cinética en la cual hubo un extenso tiempo durante el cual no hubo actividad microbiana.

Como se puede observar, la diferencia entre los errores de los conjuntos de prueba y las cinéticas no son demasiado significativos, por lo cual se seleccionó el modelo M25 para realizar la simulación del monitoreo en línea de un cultivo microbiano.

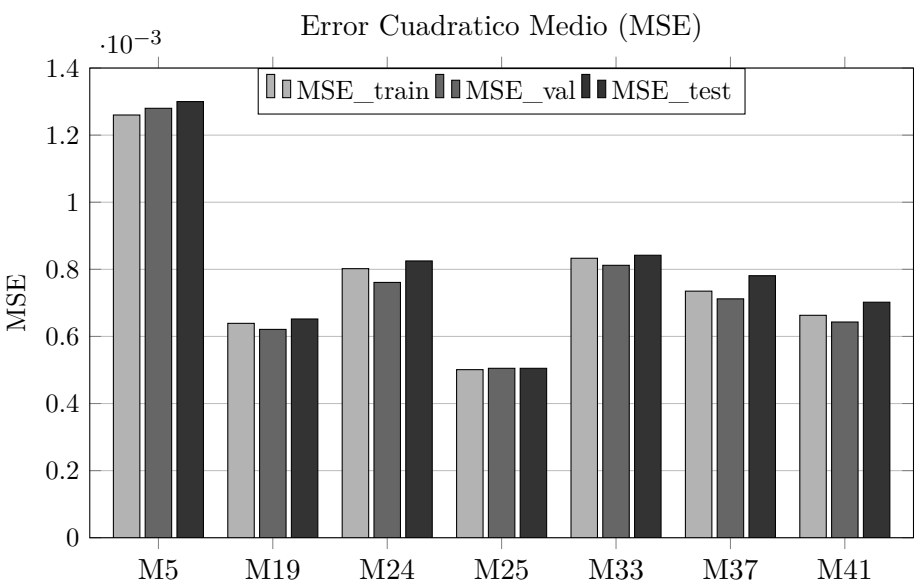


Figura 9.2: Gráfica de barras de los errores de entrenamiento (MSE_train), validación (MSE_val) y prueba (MSE_test) de los modelos LSTM.

Capítulo 10

Monitoreo en línea.

Para la simulación del monitereo en línea de un cultivo microbiano se realizaron un total de 32 predicciones. Cada predicción con un intervalo de una hora de acuerdo a los datos del archivo que se utilizó (cinética de la figura 8.3). Al mismo tiempo se muestran los datos de laboratorio pertenecientes a la cinética y con estos datos se genera un modelo logístico de acuerdo con la ecuación 6.5, ajustando el modelo mediante el uso de la librería *scikit-learn*.

En la figura 10.1 se presentan 4 predicciones realizadas durante el monitoreo. En la gráfica de la predicción a la primer hora se pudo observar que los datos del modelo LSTM generan una curva que en un principio tiene una forma similar a la generación de producto que se espera, sin embargo difiere de los datos del modelo de regresión en algunas regiones, después de los 550 minutos para ser mas preciso. Esta diferencia puede ser causada debido a los picos de glucosa que se hayan podido presentar durante ese tiempo, tal como se muestra en la gráfica de la figura 8.3.

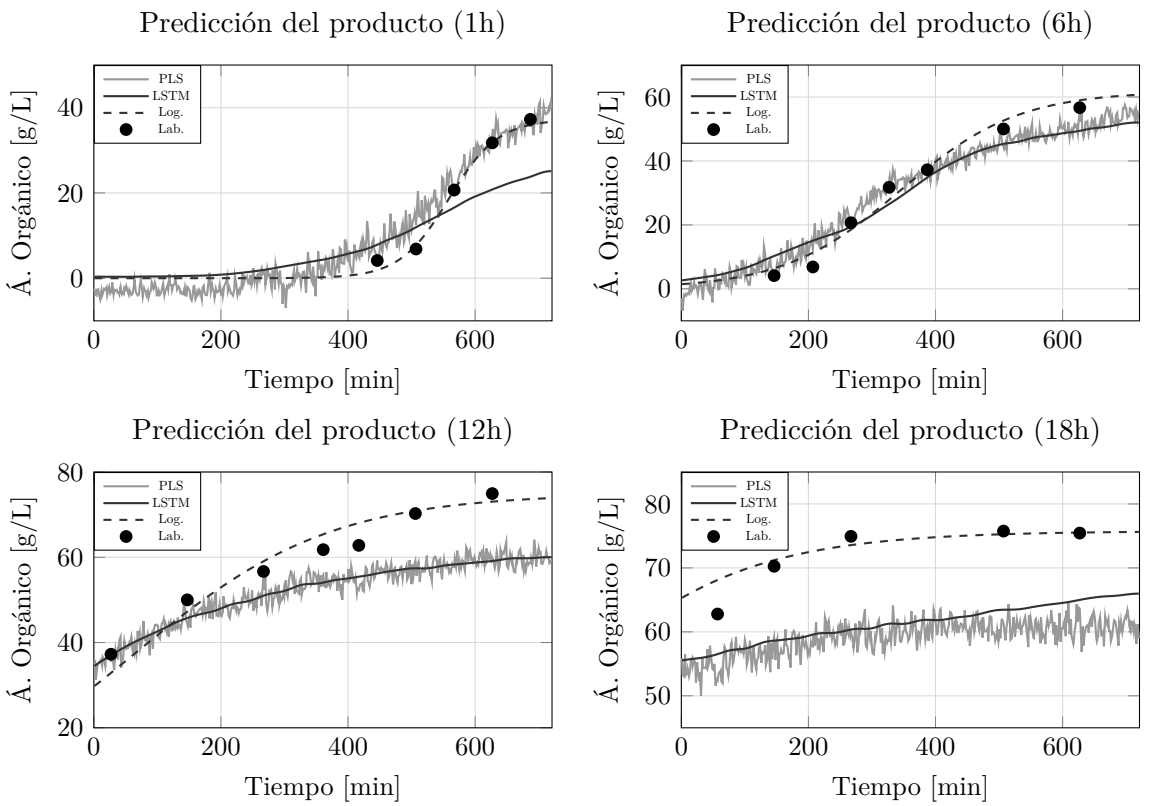


Figura 10.1: Comparación de las predicciones del modelo LSTM con los datos del modelo de regresión (PLS), los datos de laboratorio (Lab.) y los datos del modelo logístico (Log.) generado a partir de los datos de laboratorio. Las predicciones se realizaron a 1, 6, 12 y 18 horas de haber iniciado el monitoreo.

Como se pudo observar, alrededor de las seis y doce hora de monitoreo, las predicciones del modelo LSTM se ajustan muy bien al comportamiento de los datos del modelo de regresión. Además, se pudo destacar la importancia de la generación de un buen modelo de regresión pues se observó una clara diferencia en ambas gráficas, por un lado los datos de laboratorio, del modelo logístico, del modelo de regresión y del modelo de predicción se ajustan muy bien durante la fase de crecimiento del cultivo, sin embargo, existe una clara diferencia entre estos datos al llegar a la fase estacionaria que ocurre al rededor de las 24 horas.

En la última predicción se pudo observar más claramente la diferencia entre el modelo logístico, los datos de laboratorio y la predicción junto a los datos del modelo de regresión. Es importante notar la separación que se da entre el modelo de predicción y los datos del modelo de regresión después de los 600 minutos, al observar la gráfica de la figura 8.3 se puede decir que se llega a una fase estacionaria, sin embargo, recordando que es un cultivo en continuo esta fase estacionaria podría darse debido a un cambio en el ambiente en el que se encuentran las bacterias lo cual causa que la producción de ácido orgánico se anule. Debido a que el modelo de predicción solo analizó los datos de concentración de la glucosa, el ácido orgánico y la biomasa, no podría predecir este tipo de sucesos, lo cual claramente genera una diferencia entre la predicción y lo medido por el modelo de regresión.

En la tabla 10.1 se realizó una recopilación de los errores calculados de todas predicciones realizadas durante la simulación del monitoreo en línea, se pueden observar los valores mínimos acumulados del error en diferentes momentos de la predicción, dado que es un cultivo en continuo concuerda el hecho de que este valor aumente conforme más datos se tomen pues la predicción se va diferenciando cada vez más debido a los picos de glucosa, además de la dispersión que presentan los datos del modelo de regresión. Se pudo observar un valor de error frecuente que varia entre los valore de 1.6 y 3.9, así como un valor máximo de 6.

De la misma forma en la tabla 10.2 se presentan los errores del modelo LSTM con respecto a los datos de laboratorio, cabe recalcar que a diferencia de los datos del modelo de regresión o

Tabla 10.1: Comparación de los valores mínimos (MIN), modales (MOD) y máximos (MAX) del MAE, MAPE y RMSE de los datos predichos hasta 3, 6, 9 y 12 horas con respecto a los datos del modelo de regresión M58.

	Valores	**3hrs**	**6hrs**	**9hrs**	**12hrs**
MAE	**MIN**	1.194	1.2549	1.2945	1.2924
	MOD	1.5862	1.6127	3.6566	3.886
	MAX	5.8999	5.5481	4.6292	4.7083
MAPE	**MIN**	0.0208	0.0214	0.0225	0.0235
	MOD	0.9328	0.5399	0.3696	0.2839
	MAX	10.9642	6.2437	4.1871	3.148
RMSE	**MIN**	1.5053	1.5286	1.6226	1.6165
	MOD	1.9172	1.9129	4.2968	4.3423
	MAX	6.4479	6.1407	5.2328	6.1282

Tabla 10.2: Comparación de los valores mínimos (MIN), modales (MOD) y máximos (MAX) del MAE, MAPE y RMSE de los datos predichos hasta 3, 6, 9 y 12 horas con respecto a los datos de laboratorio.

	Valores	3hrs	6hrs	9hrs	12hrs
MAE	**MIN**	0.5235	1.4261	2.2966	3.6012
	MOD	6.4962	6.8661	4.925	5.8641
	MAX	14.4597	13.98	14.1243	13.7842
MAPE	**MIN**	0.0153	0.0303	0.0431	0.0602
	MOD	0.1376	0.2291	0.1578	0.1522
	MAX	1.4824	2.4153	1.4199	1.1644
RMSE	**MIN**	0.5408	1.9301	3.0339	5.1331
	MOD	6.5723	6.9736	9.6865	8.038
	MAX	14.4597	14.0345	14.1731	13.8478

del modelo logístico, los datos de laboratorio son muy pocos en comparación. Se observó un valor mínimo de 0.5, valores frecuentes con un intervalo entre 4.9 y 6.9, además de un valor máximo de 14.5.

Por otro lado la tabla 10.3 presenta los errores con respecto al modelo logístico, estos valores podrían considerarse como lo mas cercano a los errores con respecto a los valores reales de la concentración del producto. Aquí se puede encontrar un valor mínimo de 1.9, los valores mas frecuentes en un intervalo entre 9.1 y 12.8, además de un valor máximo de 55.8.

Tabla 10.3: Comparación de los valores mínimos (MIN), modales (MOD) y máximos (MAX) del MAE, MAPE y RMSE de los datos predichos hasta 3, 6, 9 y 12 horas con respecto a los datos del modelo logístico.

	Valores	3hrs	6hrs	9hrs	12hrs
MAE	**MIN**	4.2642	3.3058	2.4961	1.9358
	MOD	12.8546	14.4973	11.0989	9.1391
	MAX	55.8067	48.0716	39.3654	32.8071
MAPE	**MIN**	0.0618	0.0479	0.0362	0.0281
	MOD	0.139	0.1243	0.1081	0.0913
	MAX	0.9885	0.9646	0.8995	0.7863
RMSE	**MIN**	4.3131	3.4891	2.8995	2.5155
	MOD	12.922	14.808	11.9576	10.4808
	MAX	55.9663	48.7648	41.7201	36.6525

Capítulo 11

Conclusiones

En este proyecto se buscó diseñar un modelo de red neuronal que determinara el tiempo en el que el cultivo alcanzara un determinado nivel de concentración de los compuestos de interés, en este caso se consideró un ácido orgánico. Se logró configurar un modelo LSTM para predecir hasta doce horas de información de concentración del producto en cuestión. Con diferencia a lo esperado, con esta información se puede concluir si el cultivo alcanzará un cierto nivel de concentración dentro de las doce horas, hasta 100 gramos por litro, y también se puede determinar el tiempo en que posiblemente alcanzará ese valor, pues el modelo arroja los valores de concentración que tendrá el cultivo minuto a minuto de las doce horas de predicción, estos son 720 datos en total.

La implementación de este modelo de red neuronal en los equipos de monitoreo continuo por medio de espectroscopia Raman puede tener muchos beneficios a largo plazo, dentro de los cuales se encuentra la reducción del tiempo de cultivo. Al realizar predicciones por medio del modelo LSTM se puede determinar el momento en el que sera necesario detener la producción y comenzar con un nuevo cultivo, o por otro lado los tiempos en el que es necesario introducir mas fuente de carbono para mantener una producción constante.

Los cultivos microbianos son muy sensibles a los cambios de su entorno, requieren de una fuente de carbono abundante para poder generar sus productos, además de condiciones ambientales óptimas que deben ser controladas y monitoreadas en todo momento. Los modelos LSTM permiten la posibilidad de realizar predicciones a partir de los datos que se estén recibiendo de la cinética que se este monitoreando en ese preciso momento. Además, cada uno de los modelos generados se puede mejorar al implementar la técnica de entrenamiento de ajuste fino (*fine-tuning*) con la cual es posible presentar nuevas cinéticas que no se hayan considerado durante el entrenamiento. Esta técnica de entrenamiento combinada con la configuración del modelo para leer los datos de otros sensores como el de pH y temperatura, podrían permitir predecir posibles errores durante el cultivo, al ser entrenadas con las cineticas que hayan presentado algunos problemas debido al cambio en algún parámetro.

Es importante enfatizar en la importancia de los datos de entrenamiento. Uno de los problemas principales que se presentaron durante el desarrollo del proyecto fue la cantidad de datos con los que se contaba para la generación de los modelos de regresión y el entrenamiento del modelo de predicción. Tal como se puede apreciar en la figura 10.1 el desempeño del modelo de predicción esta íntimamente relacionado con el desempeño de los modelos de regresión que se utilicen para generar los datos de las cinéticas microbianas durante el monitoreo en línea. Cabe recalcar que cada una de las cinéticas debe tener las mismas características, al menos en este caso pues aquellos cultivos que se hayan realizado con características diferentes pueden afectar en gran medida la generación de los modelos.

Parte V

Apéndices

Apéndice A

Espectros Raman

Con propósito ilustrativo, en este apéndice se presenta la figura A.1 en donde se muestran los espectros Raman correspondientes a un cultivo microbiano de la bacteria *E. coli*. En este cultivo se utilizó glucosa como sustrato para los microorganismo.

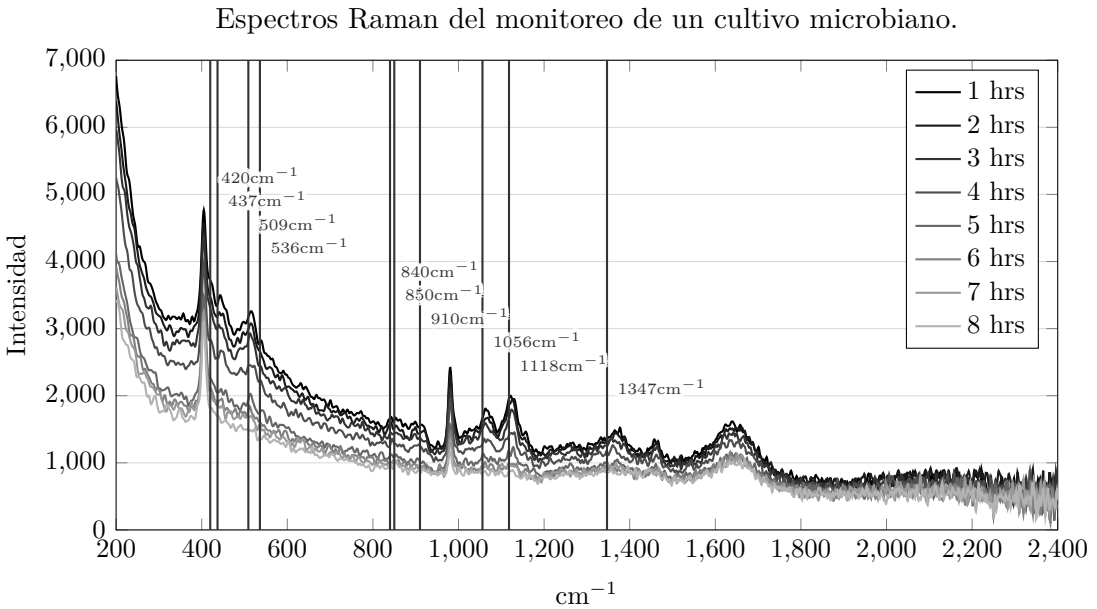


Figura A.1: Espectros Raman correspondientes al monitoreo de un cultivo microbiano. Las lineas verticales indican los picos representativos de la glucosa [7].

Los espectros mostrados corresponden a 8 mediciones realizadas durante el monitoreo del cultivo. Al observar los picos representativos de la glucosa se puede apreciar la atenuación de la intensidad para los respectivos números de onda en los espectros debido al agotamiento del sustrato a lo largo de la evolución microbiana.

Referencias

[1] S. M. Ewanick, W. J. Thompson, B. J. Marquardt y R. Bura, «Real-time understanding of lignocellulosic bioethanol fermentation by Raman spectroscopy,» *Biotechnology for biofuels*, vol. 6, n.º 1, págs. 1-8, 2013.

[2] E. Smith y G. Dent, *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons, 2005.

[3] R. Rodriguez-Diaz, T. Wehr y S. Tuck, *Analytical techniques for biopharmaceutical development*. Taylor & Francis US, 2005.

[4] R. Luo, J. Popp y T. Bocklitz, «Deep Learning for Raman Spectroscopy: A Review,» *Analytica*, vol. 3, n.º 3, págs. 287-301, 2022.

[5] E. Z. Panagou, C. C. Tassou, E. K. Saravanos y G.-J. E. Nychas, «Application of neural networks to simulate the growth profile of lactic acid bacteria in green olive fermentation,» *Journal of food protection*, vol. 70, n.º 8, págs. 1909-1916, 2007.

[6] D. A. Skoog, F. J. Holler y T. A. Nieman, *Principios de análisis instrumental*. Cengage Learning México^ eD. FDF, 2008.

[7] A. D. Shaw, N. Kaderbhai, A. Jones, A. M. Woodward, R. Goodacre y J. J. Rowland, «Noninvasive, on-line monitoring of the biotransformation by yeast of glucose to ethanol using dispersive Raman spectroscopy and chemometrics,» *Applied spectroscopy*, vol. 53, n.º 11, págs. 1419-1428, 1999.

[8] L. R. Sadergaski, T. J. Hager y H. B. Andrews, «Design of Experiments, Chemometrics, and Raman Spectroscopy for the Quantification of Hydroxylammonium, Nitrate, and Nitric Acid,» *ACS omega*, vol. 7, n.º 8, págs. 7287-7296, 2022.

[9] A. Asongo, M. Barma y H. Muazu, «Machine Learning Techniques, methods and Algorithms: Conceptual and Practical Insights,» *Int. J. Eng. Res. Appl*, vol. 11, págs. 55-64, 2021.

[10] H. Witjes, M. Van den Brink, W. Melssen y L. Buydens, «Automatic correction of peak shifts in Raman spectra before PLS regression,» *Chemometrics and Intelligent Laboratory Systems*, vol. 52, n.º 1, págs. 105-116, 2000.

[11] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[12] P. Wang, L. Guo, Y. Tian, J. Chen, S. Huang y C. Wang, «Discrimination of blood species using Raman spectroscopy combined with a recurrent neural network,» *OSA continuum*, vol. 4, n.º 2, págs. 672-687, 2021.

[13] W. Zeng, J. Huang, Z. Xia, Z. Li y H. Qu, «Classification of pathogenic bacteria by Raman spectroscopy based on recurrent neural network,» en *4th Optics Young Scientist Summit (OYSS 2020)*, SPIE, vol. 11781, 2021, pág. 1 178 102.

[14] J. R. Ferraro, *Introductory raman spectroscopy*. Elsevier, 2003.

[15] A. Campion, «Raman spectroscopy,» *Vibrational spectroscopy of molecules on surfaces*, págs. 345-415, 1987.

[16] J. N. Miller y J. C. Miller, «Statistics and chemometrics for analytical chemistry,» *Signal*, vol. 100, pág. 95, 1998.

[17] S. Wold, M. Sjöström y L. Eriksson, «PLS-regression: a basic tool of chemometrics,» *Chemometrics and intelligent laboratory systems*, vol. 58, n.º 2, págs. 109-130, 2001.

[18] S. Haykin, *Neural networks and learning machines, 3/E*. Pearson Education India, 2009.

[19] C. M. Bishop y N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[20] A. R. S. Parmezan, V. M. Souza y G. E. Batista, «Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model,» *Information sciences*, vol. 484, págs. 302-337, 2019.

[21] X.-Y. Qian y S. Gao, «Financial series prediction: Comparison between precision of time series models and machine learning methods,» *arXiv preprint arXiv:1706.00948*, págs. 1-9, 2017.

[22] A. Graves y N. Jaitly, «Towards end-to-end speech recognition with recurrent neural networks,» en *International conference on machine learning*, PMLR, 2014, págs. 1764-1772.

[23] J. Monod, «The growth of bacterial cultures,» *Annual review of microbiology*, vol. 3, n.º 1, págs. 371-394, 1949.

[24] J. Van Impe, F. Poschet, A. Geeraerd y K. Vereecken, «Towards a novel class of predictive microbial growth models,» *International journal of food microbiology*, vol. 100, n.º 1-3, págs. 97-105, 2005.

[25] M. H. Zwietering, I. Jongenburger, F. M. Rombouts y K. Van't Riet, «Modeling of the bacterial growth curve,» *Applied and environmental microbiology*, vol. 56, n.º 6, págs. 1875-1881, 1990.

[26] M. G. Jiménez-Escamilla, C. Garibay-Origel y M. A. Borja-Salin, «Modelo bioquímicamente estructurado para la estimación de la eficiencia de una celda de combustible microbiana,» *Revista internacional de contaminación ambiental*, vol. 34, n.º 2, págs. 331-345, 2018.

[27] D.-K. Kim, J. M. Park, H. Song e Y. K. Chang, «Kinetic modeling of substrate and product inhibition for 2, 3-butanediol production by Klebsiella oxytoca,» *Biochemical engineering journal*, vol. 114, págs. 94-100, 2016.

[28] K. H. Chu, «Fitting the Gompertz equation to asymmetric breakthrough curves,» *Journal of Environmental Chemical Engineering*, vol. 8, n.º 3, pág. 103 713, 2020.

[29] N. K. Afseth, V. H. Segtnan y J. P. Wold, «Raman spectra of biological samples: A study of preprocessing methods,» *Applied spectroscopy*, vol. 60, n.º 12, págs. 1358-1367, 2006.

[30] W. Peng y Q. Ni, «A hybrid SVM-LSTM temperature prediction model based on empirical mode decomposition and residual prediction,» en *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2020, págs. 1616-1621.