



**GOBIERNO DE
MÉXICO**

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



**TECNOLÓGICO
NACIONAL DE MÉXICO**

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO

Doctorado en Ciencias en Ingeniería Electrónica
Especialidad en Control Automático

1er reporte
1° semestre

**Modelado de datos ómicos con técnicas de aprendizaje
profundo mediante redes neuronales**

Presenta:
M.C. Arturo Figueroa Arcos

Directores de tesis:

Dr. Víctor Manuel Alvarado Martínez
Dra. Ma. Guadalupe López López

Revisores de tesis:

Dr. Luis Gerardo Vela Valdés
Dr. Victor Hugo Olivares
Dra. Marisol Cervantes Bobadilla

Cuernavaca, Morelos a 15 de Junio de 2024

Tabla de contenido

Índice de figuras	II
Índice de tablas	III
Nomenclatura	IV
1. Introducción	1
2. Marco teórico	2
2.1. Datos omicos	2
2.2. Tipos de ciencias ómicas	2
2.3. Representación de datos ómicos	3
2.4. Redes neuronales artificiales	4
2.4.1. Componentes de la red neuronal	4
2.4.2. Funciones de activación	5
2.5. Inteligencia artificial	6
2.5.1. Aprendizaje Automático	6
2.5.2. Aprendizaje Profundo	7
2.6. Red Neuronal Convolucional	7
2.7. Red Neuronal Recurrente	9
3. Estado del arte	9
3.1. Datos ómicos	10
3.1.1. Análisis de datos omicos	11
3.1.2. Bases de datos ómicos	12
3.2. Aprendizaje profundo	13
3.2.1. Aprendizaje profundo en ómica	13
4. Propuesta de tesis	17
4.1. Objetivo general	17
4.2. Objetivos específicos	17
4.3. Antecedentes	18
4.4. Planteamiento del problema	18
4.5. Pregunta de investigación	18
4.6. Justificación	18
5. Cronograma de actividades	19

Índice de figuras

2.1. Clasificación de ciencias ómicas y relación entre ellas	2
2.2. Representación de neuronas artificiales de una entrada y múltiples entradas (Hagan et al., 2014).	4
2.3. Redes neuronales artificiales: (a) de capa oculta y (b) múltiples capas.	5
2.4. Subcampos de la inteligencia artificial en el análisis de datos	7
2.5. Diagrama de red convolucional simple.	8
2.6. Diagrama esquemático de operación de convolucion.	8
2.7. Diagrama esquemático de la operación de agrupación.	8
2.8. Estructura de la red neuronal recurrente simple y expandida.	9
5.1. Cronograma de actividades	19

Índice de tablas

0.1. Siglas y acrónimos	IV
2.1. Listado de ciencias ómicas establecidas	3
2.2. Representación de los datos más frecuentemente utilizados	4
3.1. Bases de datos ómicos disponibles de libre acceso	13
3.2. Revisión del estado del arte de algoritmos utilizados en clasificación, identificación, codificación y clasificación	15
3.3. Bases de datos enfocadas en el análisis de datos omicos de tipos de cáncer. .	16
3.4. Revision de articulos enfocados en cancer destacando conjunto de entrenamiento, tipos de datos y base de datos utlizada.	17

Nomenclatura

Tabla 0.1: Siglas y acrónimos

Siglas	Descripción
IA	Inteligencia Artificial
DL	Deep Learning
ML	Machine Learning
CNN	Convolutional Neuronal Network
RNN	Recurrent Neuronal Network
NGS	Next Generation Sequencing
ANN	Artificial Neuronal Network

1. Introducción

En la ingeniería biomédica se aplica tecnología de última generación, para la creación de dispositivos médicos y métodos que permitan contribuir al bienestar humano, para obtener una mejora en la comprensión de los procesos biológicos que suceden en el ser humano. En este campo de estudio intervienen áreas como la ingeniería electrónica, la ingeniería mecánica, la medicina, la biología, la física, entre otros, considerando a la ingeniería biomédica como un campo interdisciplinario.

La generación de datos ómicos ha experimentado un crecimiento exponencial gracias a las tecnologías de secuenciación de alto rendimiento (NGS). Debido a la alta cantidad de datos existentes, se han impulsado métodos analíticos avanzados para interpretar y extraer información biológica significativa. Los datos que se abarcan son de diferentes niveles de organización molecular, como el ADN, el ARN, las proteínas y metabolitos.

Algunos ejemplos de estos datos ómicos son los genómicos que estudian el ADN incluyendo la secuencia, estructura y función, los datos transcriptómicos que estudian el ARN incluyendo su expresión, splicing y modificaciones, los datos proteómicos que estudian las proteínas, incluyendo su estructura, función e interacciones y los datos metabolómicos que estudian las moléculas pequeñas que interactúan en las reacciones químicas de la célula.

El análisis de este tipo de datos cuenta con el potencial para mejorar significativamente nuestra comprensión de la salud y las enfermedades en un ser humano. Los pasos típicos son el preprocesamiento donde se realiza la limpieza y normalización de los datos con el fin de eliminar errores, control de calidad para identificar y eliminar posibles valores atípicos, análisis estadístico para identificar biomarcadores relevantes así como también modelos predictivos y de asociación y por último la interpretación de los resultados con sentido biológico. Algunos de los desafíos que se presentan en el análisis son el alto volumen de datos, complejidad de los datos y la integración de tipo de datos, es por ello por lo que se considera un desafío computacional.

El Aprendizaje Profundo (DL) es un campo de la inteligencia artificial, que ha demostrado ser una herramienta poderosa para el análisis de datos complejos, está basado en la utilización de redes neuronales artificiales (ANN's) con el fin de identificar patrones complejos, realizar análisis no lineales y modelar relaciones a partir de diferentes tipos de datos ómicos.

Las ANN's y DL aplicadas en la ingeniería biomédica representan una oportunidad para los profesionales de la salud, ya que permiten realizar análisis más rápidos de grandes conjuntos de datos e información médica relevante, mejoras en los métodos de diagnóstico y pronóstico de enfermedades, diseño de terapias personalizadas y mejoras para el bienestar humano. Los principales desafíos que se presentan son la necesidad de grandes conjuntos de datos para entrenamiento de las redes, la interpretabilidad de las redes neuronales para comprender su toma de decisiones y el requerimiento computacional para entrenamiento de las redes (Sarmiento-Ramos, 2020).

En este proyecto se enfocará con el propósito de clasificación de tipos y subtipos de fenotipos, así como también en el pronóstico de supervivencia, utilizando de manera preliminar datos omicos centrados en el cáncer como los genómicos, transcriptómicos, proteómicos y metabolómicos. El área de oportunidad que se encuentra es en el preprocesamiento de datos y la codificación debido a la heterogeneidad de los datos procedentes de distintas bases que se encuentran de libre acceso. El propósito de este estudio estará dirigido en el estudio de algoritmos de aprendizaje profundo e implementarlos con enfoque de aplicación en el área biomédica con el fin de dar soporte en la precisión del diagnóstico y predicción de enfermedades.

2. Marco teórico

2.1. Datos omicos

Las ciencias ómicas conocidas también como biomedicina o bioinformática, estudian los procesos biológicos a nivel molecular a través de grandes conjuntos de datos con el fin de diagnosticar, prevenir y predecir enfermedades, así como también en terapias y tratamientos personalizados en pacientes. El término de “omica” se deriva del griego “óma” que significa masa o conjunto (Ravì et al., 2016; Mamoshina et al., 2016). La omica es un campo de la biomédica con una gran extensión y esta se divide en diferentes ramas como la genómica, transcriptómica, proteómica, metabolómica, epigenómica, farmacogenómica y metagenómica.

2.2. Tipos de ciencias ómicas

Las diferentes ramas de la omica se dirigen a distintos niveles moleculares, habitualmente se estudian de manera independiente a la hora de abordar enfermedades y obtener conocimiento médico y científico, cada una ofrece información importante, pero en conjunto las ciencias permiten obtener relaciones de los niveles moleculares y entender su complejidad biológica (Figura 2.1)

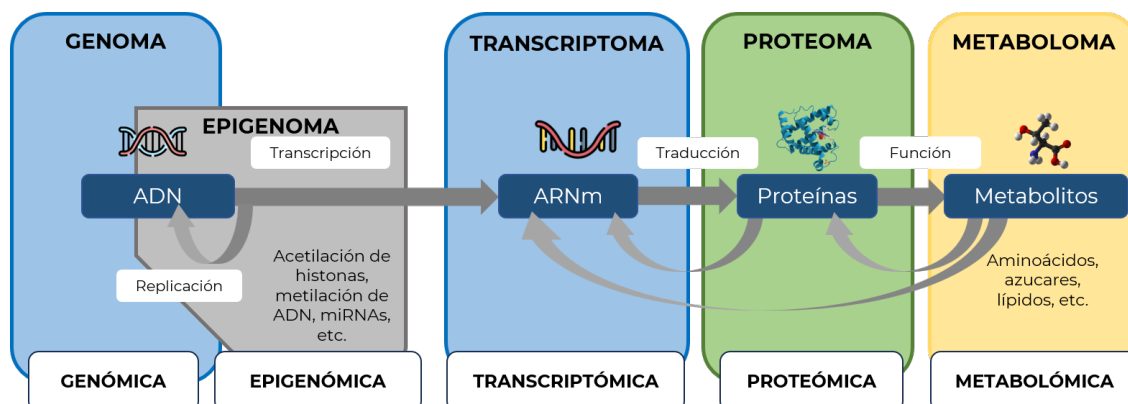


Figura 2.1: Clasificación de ciencias ómicas y relación entre ellas

La genómica es de las primeras ciencias ómicas en reconocerse como tal, esta se encarga del estudio de los genomas, es decir, la totalidad del material genético que tiene un organismo vivo o una partícula viral. El objetivo es identificar alelos genéticos y factores ambientales que contribuyen al desarrollo de enfermedades (Javier et al., 2019).

La transcriptómica estudia el patrón de la expresión genética en un organismo o en células específicas bajo circunstancias concretas, es decir, el conjunto de los ARN mensajeros (ARNm) y no codificantes, a nivel cuantitativo y cualitativo (Davis and Shanley, 2017).

La epigenómica estudia el conjunto de modificaciones reversibles del ADN o de las proteínas asociadas al ADN (como las histonas) que actúan como elementos funcionales de regulación de la expresión génica de una célula sin alterar la secuencia de su ADN (Hasin et al., 2017).

La proteómica estudia el conjunto de las proteínas con sus isoformas y modificaciones postraduccionales expresadas en una celular, tejido u órgano concreto en un momento dado, bajo determinadas condiciones y localización específica, dado que las proteínas median las actividades bioquímicas en una célula (Van Eyk and Snyder, 2018).

La metabólica es la ciencia que estudia el conjunto completo de los metabolitos (intermediarios metabólicos, hormonas y metabolitos secundarios) que se encuentran en un momento dado en una célula, tejido u órgano. Entre los metabolitos estudiados se incluyen desde el oxígeno, los aminoácidos esenciales o las vitaminas (Zhao et al., 2014).

En la tabla 2.1 siguiente se resume la información de las ciencias ómicas establecidas y su definición por área de estudio:

Tabla 2.1: Listado de ciencias ómicas establecidas

Ciencia ómica	Área de estudio
Genómica	Estudio del conjunto del material genético presente en un organismo.
Transcriptómica	Estudio de los perfiles de expresión de los ARN mensajeros, los microARNs y ARN no codificantes.
Epigenómica	Estudio de los elementos que controlan la expresión génica sin modificar la secuencia de nucleótidos del ADN.
Proteómica	Estudio del set completo de proteínas expresadas en un organismo en un tiempo determinado y particular de cada tipo celular o tisular.
Metabolómica	Identificación y cuantificación de productos metabólicos de pequeño tamaño (metabolitos) de un sistema biológico (célula, tejido, fluido biológico u órgano)
Farmacogenómica	Estudio de los genes que afectan a la respuesta de una persona a determinados fármacos
Metagenómica	Estudio del conjunto de microorganismos de una muestra ambiental para proporcionar información de la diversidad ecológica de un ambiente determinado.

2.3. Representación de datos ómicos

En cada ciencia ómica se tiene un tipo distinto de representación de acuerdo al tipo de datos que se extraen, esto es fundamental para el almacenamiento, análisis e interpretación

de este tipo de datos, para permitir a los investigadores extraer características biológicas, en la tabla 2.2 se muestra la forma en que se representa cada dato:

Tabla 2.2: Representación de los datos más frecuentemente utilizados

Tipos de datos	Descripción de la representación
Genómicos	Secuencias de ADN, representadas por cadenas de 4 caracteres.
Transcriptómicos	Perfiles de expresión genética, representados por una tabla o matriz numérica.
Epigenómicos	Señales registradas en un vector que contiene series temporales o perfiles, representadas en una tabla de valores numéricos.
Proteómicos	Secuencias de proteínas, representadas por una cadena de 20 caracteres. Espectrometría de masas, representada por un vector con serie temporal.
Metabólicos	Cálculo de espectro, representado por vector con series temporales. Perfiles, representados en una tabla con valores numéricos.

2.4. Redes neuronales artificiales

Las ANNs son un modelo de algoritmo computacional inspirado en las redes biológicas, con el que se establecen relaciones entre las entradas y salidas. Se caracteriza por ser una herramienta que tiene la capacidad para aprender, procesar y generalizar automáticamente datos, utilizadas en tareas de clasificación y regresión (Hagan et al., 2014). En el propósito de clasificación, los datos de entrada son clasificados en distintas clases, y en la regresión, o aproximación de función, se realiza para predecir un parámetro de salida desconocido (Hagan et al., 2014). Es por esto que las ANNs cuentan con el potencial en aplicaciones de reconocimiento de patrones y predicción de comportamiento.

2.4.1. Componentes de la red neuronal

Una neurona artificial se compone de una o múltiples entradas p , un peso w , un bias o umbral b , un sumador Σ y una función de activación f , como se puede observar en la Figura 2.2.

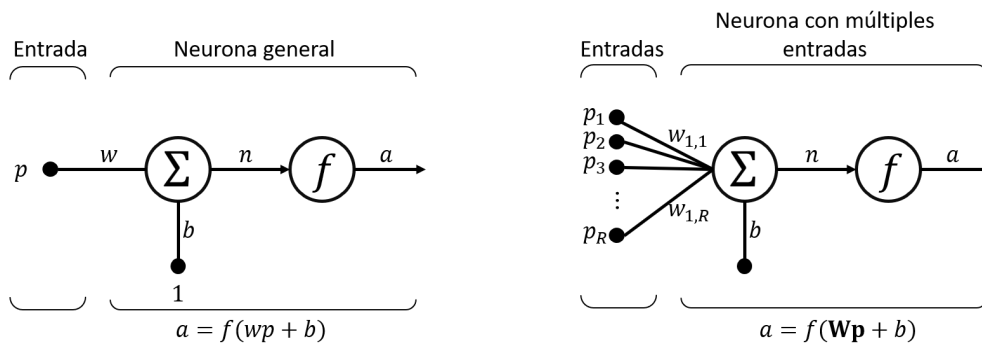


Figura 2.2: Representación de neuronas artificiales de una entrada y múltiples entradas (Hagan et al., 2014).

Normalmente, una neurona con múltiples entradas suele no ser suficiente, por lo que las

ANNs se usan con varias neuronas ubicadas en paralelo formando un “capa”. Las ANNs pueden poseer una capa o múltiples como se observa en la Figura 2.3.

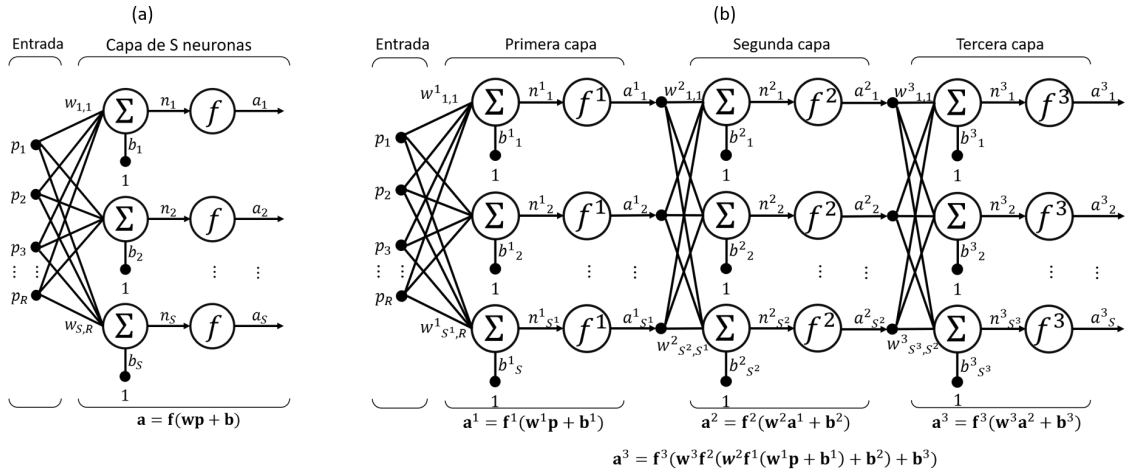


Figura 2.3: Redes neuronales artificiales: (a) de capa oculta y (b) múltiples capas.

El perceptron es conocido por ser el primer modelo de red neuronal, utilizado en clasificación de patrones a partir de entradas escalares binarias o vectores bipolares (Rosenblatt, 1958). El algoritmo de aprendizaje se basa en la regla de Hebb, usando conjunto de datos de entrenamiento para obtener los pesos de la red a través de iteraciones inicialmente con valores aleatorios.

2.4.2. Funciones de activación

Las funciones de activación es una función matemática que determina la salida de una neurona artificial en función de su entrada. Esta función introduce no linealidad en el modelo, lo que permite a las redes neuronales aprender y modelar relaciones complejas entre los datos de entrada y salida (Dubey et al., 2022). Las funciones desempeñan un papel muy crucial en las redes neuronales al aprender características abstractas a través de transformaciones no lineales. Algunas de sus propiedades comunes son las siguientes: a) Debe agregar la curvatura no lineal en el panorama de optimización para mejorar la convergencia del entrenamiento de la red, b) No debería aumentar significativamente la complejidad computacional, c) no debería obstaculizar el flujo del gradiente durante el entrenamiento y d) debería conservar la distribución de datos para facilitar una mejor capacitación de la red.

Existen diferentes tipos de funciones de activación, cada una con sus propias características y aplicaciones. Entre las más comunes se encuentran las siguientes:

- Función tipo escalón: Esta función es la más simple y toma un valor de 1 si la entrada es mayor o igual que un umbral determinado, y 0 en caso contrario.
- Función Sigmoide: Esta función toma un valor entre 0 y 1, asemejándose a una curva en forma de S.

- Función tangente hiperbólica: Esta función es similar a la función sigmoide, pero toma valor entre -1 y 1.
- Función ReLU (Unidad Lineal Rectificada): Esta función toma un valor igual a la entrada si esta es mayor o igual que cero, y 0 en caso contrario.
- Función Leaky ReLU: Esta función es similar a la función ReLU, pero introduce una pequeña pendiente negativa.

2.5. Inteligencia artificial

La inteligencia artificial (IA) es un campo de la ciencia de la computación, que se enfoca en crear máquinas inteligentes que puedan razonar, aprender y actuar de manera autónoma (Rouhiainen, 2018).

Las tecnologías basadas en IA es usada para mejorar y disfrutar una mayor eficiencia en distintos ámbitos de la vida. La aplicación se puede realizar en diversas situaciones y algunas de las más importantes son las siguientes:

- Reconocimiento de imágenes estáticas, clasificación y etiquetado.
- Mejoras del desempeño de la estrategia algorítmica comercial.
- Procesamiento eficiente y escalable de datos de pacientes.
- Mantenimiento predictivo.
- Detección y clasificación de objetos
- Protección contra amenazas de cibernética.

Dentro de la IA se encuentran diversas técnicas para crear sistemas inteligentes, como el aprendizaje automático y aprendizaje profundo que son subcampos que desempeñan un papel fundamental en el análisis de grandes cantidades de datos, identificar patrones y tendencias, de una forma más rápida y precisa (Figura 2.4).

2.5.1. Aprendizaje Automático

El aprendizaje automático (en inglés, *machine learning*) se centra en el desarrollo de sistemas capaces de aprender de conjuntos de datos sin ser programados de manera explícita (Mitchell, 1997). Estos sistemas aprenden y mejoran su rendimiento en una tarea específica a medida que se le presentan más datos. Un resultado típico serían las sugerencias o predicciones en una situación particular (Vieira et al., 2020).

Desde el punto de vista de ingeniería se define como un programa de computador que aprende de una experiencia E , con respecto a una tarea T y una medida de rendimiento R , si su rendimiento en T , medido por R , mejora con la experiencia E y también se puede definir como la ciencia de programar computadores para que aprendan a partir de un conjunto de datos (Géron, 2020).

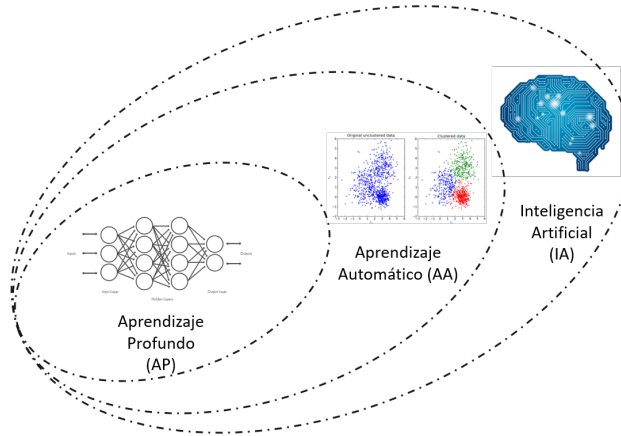


Figura 2.4: Subcampos de la inteligencia artificial en el análisis de datos .

2.5.2. Aprendizaje Profundo

El aprendizaje profundo (en inglés, *deep learning*) es una rama del aprendizaje automático. A diferencia de los algoritmos tradicionales de aprendizaje automático, muchos de los cuales tienen una capacidad finita de aprendizaje independientemente de cuántos datos adquieran, los sistemas de aprendizaje profundo pueden mejorar su rendimiento al poder acceder a un mayor número de datos, o lo que es lo mismo, hacer que la máquina tenga más experiencia (Shinde and Shah, 2018). Una vez que las máquinas han conseguido suficiente experiencia mediante el aprendizaje profundo, pueden ponerse a trabajar para realizar tareas específicas como conducir un coche, detectar hierbas en un campo de cultivo, detectar enfermedades, inspeccionar maquinaria para identificar errores, etc.

El aprendizaje profundo toma los fundamentos teóricos de las ANNs clásicas, pero emplea una gran cantidad de neuronas y capas ocultas, junto con nuevos modelos y paradigmas de entrenamiento ofreciendo una capacidad mucho mayor para aprender a adaptarse y extraer características de datos de entrada de alta complejidad (Schmidhuber, 2015). Las ANNs usadas en el aprendizaje profundo son conocidas como redes neuronales profundas, en inglés *Deep Neuronal Network* (DNNs).

2.6. Red Neuronal Convolucional

Las redes neuronales convolucionales, en inglés, *Convolutional Neuronal Network* (CNN) son un tipo algoritmo propuesto por LeCun en 1989 (LeCun et al., 1989). La aplicación de las CNN se encuentra principalmente en el reconocimiento de voz, reconocimiento facial, reconocimiento de objetos, análisis de movimiento y procesamiento de lenguaje natural.

Las CNN habitualmente constan de múltiples capas de convolución acompañadas de capas de agrupación y una capa de neuronas completamente conectadas, como se muestra en la Figura 2.5.

La función de convolución se utiliza principalmente para extraer diversas características de los datos analizados. En proceso de la convolución cuenta de un núcleo de convolución,

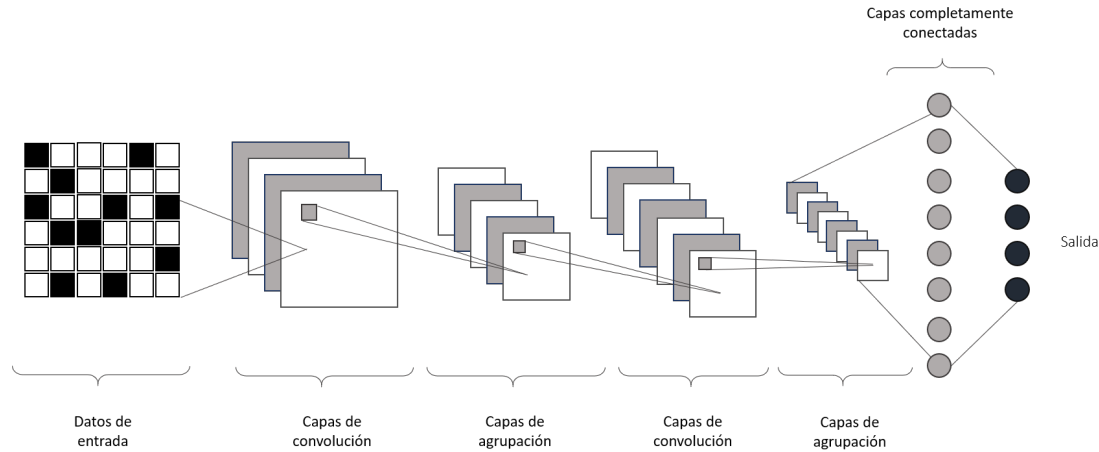


Figura 2.5: Diagrama de red convolucional simple.

el cual se va deslizando en la ventana de entrada, de modo que los parámetros de peso en el núcleo se vayan multiplicando por los píxeles correspondientes. Posteriormente, los resultados siguen la multiplicación. En la Figura 2.6 se muestra el principio de la convolución.

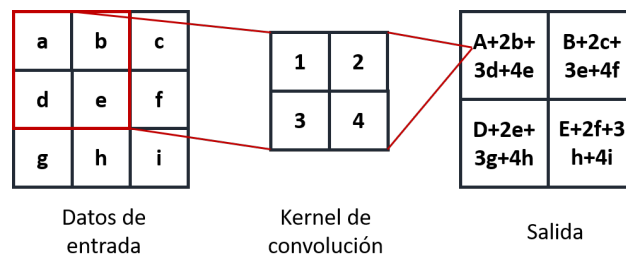


Figura 2.6: Diagrama esquemático de operación de convolucion.

La función de la capa de agrupación es abstraer la señal característica original, lo cual se realiza para reducir en gran medida los parámetros de entrenamiento y también poder reducir el grado de sobreajuste. Las operaciones de agrupación se dividen en dos categorías: agrupación máxima y agrupación media. En la agrupación máxima se toma el valor más grande de un pixel correspondiente como resultado del muestreo, y en la agrupación media se calcula el valor promedio del pixel correspondiente. En la Figura 2.7 se muestra el principio de la agrupación.

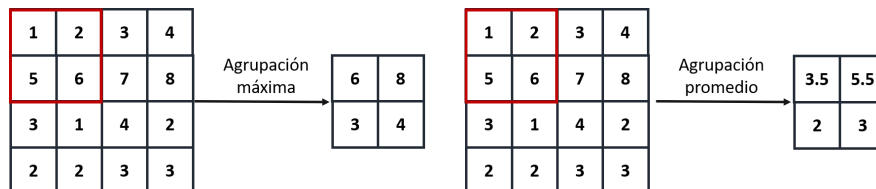


Figura 2.7: Diagrama esquemático de la operación de agrupación.

2.7. Red Neuronal Recurrente

Las redes neuronales recurrentes, en inglés, *Recurrent Neuronal Network* (RNN) es un tipo de algoritmo propuesto en 1980. En los últimos años, las aplicaciones de las RNN han sido en muchos campos como: el procesamiento del lenguaje natural, el reconocimiento de imágenes y reconocimiento de voz.

La característica principal de las RNN es que la entrada de la capa oculta incluye no solo la salida de capa de entrada, sino también la salida de la capa oculta en el último momento. Un modelo RNN simple puede ser expandido a una red compleja. En la Figura 2.8 se puede observar la estructura de una RNN y el mapa de dependencia del orden temporal.

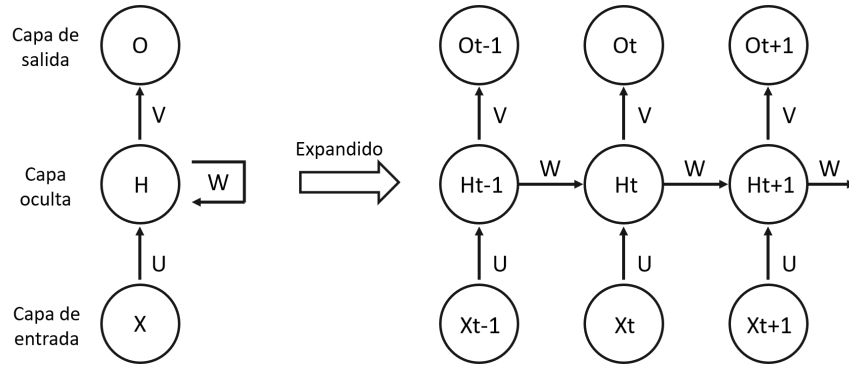


Figura 2.8: Estructura de la red neuronal recurrente simple y expandida.

En la estructura de la RNN, H_t es el estado oculto del tiempo t y O_t representa la salida del tiempo t ; U es el peso directo de la capa de entrada a la capa oculta; W es el peso de la capa oculta a la capa oculta, que es el controlador de memoria de la red que se encarga de programar la memoria; V es el peso de la capa oculta a la capa de salida, y las características aprendidas de la capa oculta pasaran a través de ella nuevamente y como salida final.

3. Estado del arte

Como ya se mencionó en la introducción, los datos ómicos abarcan diferentes conjuntos de datos que miden aspectos biológicos desde el ADN, el ARN, proteínas y metabolitos, en los últimos años se han desarrollado herramientas como la creación de bases de datos públicas para facilitar a los investigadores el acceso al análisis de estos datos y el desarrollo de métodos enfocados en la identificación, integración de datos y predicción.

El aprendizaje profundo ha revolucionado el análisis de datos ómicos, obteniendo información compleja y más precisa de conjuntos masivos de datos. Algunas de las principales aplicaciones en la biomedicina son el descubrimiento de biomarcadores, predicción de fenotipos, integración de datos ómicos, diseño de fármacos y regulación genética.

Se destacan algunas estructuras de ANN's frecuentemente usadas en el análisis de datos ómicos como las redes neuronales convolucionales en la identificación de patrones espaciales, las redes neuronales recurrentes para modelar secuencias temporales, auto codificadores, redes neuronales profundas, redes generativas y codificadores automáticos variacionales.

Los principales desafíos que se presentan son que se requieren grandes conjuntos de datos para entrenamiento y su procesamiento es una tarea que compromete la eficiencia de la red, sesgo de los datos y la codificación de los datos.

En las siguientes secciones se abordará con más detalle sobre el estado del arte de los datos omicos, aprendizaje profundo y sus aplicaciones en el análisis de datos ómicos.

3.1. Datos ómicos

Las ciencias ómicas se refieren a la evaluación integral o global de una colección de características de un ser vivo, el estudio de genes y proteínas han identificado con éxito, características clave que influyen en la salud y la enfermedad. Los datos omicos han sido impulsados, debido al desarrollo y disponibilidad de tecnologías de matrices, espectrometría de masas de alto rendimiento y plataformas de secuenciación (Hasin et al., 2017).

Los sistemas biológicos dependen de la transferencia de información de los ácidos nucleicos a las proteínas y metabolitos para dar forma a la función y fenotipo, es por esto que el estudio de las enfermedades es el resultado de procesos complejos y heterogéneos (Kim and Tagkopoulos, 2018). Los datos omicos son utilizados para el análisis de valores atípicos en una secuencia, lo cual puede indicar la progresión de una enfermedad y permite aprovecharse como indicadores para un diagnóstico temprano, con esto se pueden perfeccionar los enfoques de detección y diagnóstico, así como también identificar y personalizar intervenciones o tratamientos (Krassowski et al., 2020).

Dentro de las ciencias ómicas frecuentemente utilizadas se encuentran: la genómica, transcriptómica, proteómica y la epigenómica, a continuación se detalla un poco más sobre estas ciencias:

La genómica se encarga de la caracterización del contenido genético de un organismo, típicamente consiste en ADN (ARN en algunos virus), la secuenciación del genoma es utilizada para identificación de nuevos genes y variantes genéticas y mediante estudios de asociación de todo el genoma se relacionan variantes genómicas con estados patológicos u otros fenotipos.

La transcriptómica se centra en el estudio del ARN codificante de proteínas (ARN mensajero), las señales transcriptómicas proporcionan información sobre los genes y mecanismos potenciales implicados en un proceso biológico de interés.

La proteómica estudia las proteínas expresadas, las moléculas fundamentales para la vida y su funcionamiento en los organismos vivos. Esta información tiene el potencial para mejorar la comprensión de la biología, el diagnóstico de enfermedades y desarrollo de tratamientos.

La epigenómica aborda la caracterización de todo el genoma, de modificaciones químicas reversibles del ADN o de proteínas asociadas al ADN que afectan la expresión y regulación de genes. Las modificaciones epigenómicas pueden proporcionar información sobre el estado de la enfermedad y/o la exposición ambiental y actuar como rasgos hereditarios.

Uno de los aspectos actuales más destacados en la investigación omica es la integración de datos ómicos (multiómica) de diferentes niveles moleculares con el fin de obtener una visión holística de los sistemas biológicos. Se han encontrado desafíos en la integración de datos como la heterogeneidad de datos, la falta de estándares de datos y la complejidad computacional.

3.1.1. Análisis de datos omicos

La tecnología de análisis de datos omicos ha cambiado la forma en la que se estudian las enfermedades, permitiendo conocer los cambios a nivel molecular que intervienen en diversos padecimientos. A través de los datos omicos se obtiene una comprensión de las causas y mecanismos de las enfermedades, con lo cual se abren oportunidades en aplicaciones de diagnóstico, tratamiento y prevención por medio del descubrimiento de nuevos biomarcadores biológicos (Reel et al., 2021).

Un biomarcador es una sustancia, estructura o proceso que se puede medir en el cuerpo humano o sus productos y puede proporcionar información importante sobre la presencia de una enfermedad o afección (Strimbu and Tavel, 2010). Los biomarcadores moleculares se descubren analizando grandes cantidades de información proporcionada de diferentes ómicas y estos desempeñan un papel importante en la planificación de medidas y decisiones preventivas para los pacientes en el diagnóstico, pronóstico o predicción. Los biomarcadores de diagnóstico se utilizan para determinar la presencia de una enfermedad en un paciente, por otro lado, los biomarcadores de pronóstico brindan información sobre el resultado general con o sin tratamiento. los biomarcadores predictivos se utilizan para identificar el riesgo de sufrir algún resultado (Carlomagno et al., 2017).

Dentro de las principales enfermedades analizadas en omica e identificación de biomarcadores se encuentran:

El cáncer debido a que es una de las principales causas de muerte en el mundo y la omica ha contribuido significativamente a la comprensión de su complejidad (Sarmiento-Ramos, 2020). En el diagnóstico, el análisis ómico ha permitido la identificación de biomarcadores moleculares para la detección de cáncer en una etapa temprana, incluso antes de la presencia de síntomas. En el pronóstico se puede predecir el curso de la enfermedad y la probabilidad de respuesta al tratamiento, la expresión genética de ciertos genes puede usarse para predecir el riesgo de recurrencia del cáncer. En el tratamiento se han impulsado el desarrollo de terapias dirigidas, que se enfocan en las mutaciones genéticas específicas que impulsan el crecimiento del cáncer. Y en la prevención se usa para la identificación de individuos con un alto riesgo de desarrollar cáncer, lo que permite implementar estrategias de prevención personalizar (Munir et al., 2019).

Las enfermedades cardiovasculares, como los ataques cardíacos y los accidentes cerebrovasculares, son otra de las causas de muerte a nivel global. En el diagnóstico se puede predecir el riesgo de presentar enfermedades cardiovasculares en el que se analizan los niveles de lipoproteína de baja densidad y de alta densidad que se asocian a un mayor riesgo de enfermedad cardíaca. En el tratamiento, la omica ha contribuido en el desarrollo de fármacos para implementación en pacientes con dicha patología (Pasha et al., 2020). En la prevención se utiliza para identificar en una persona propensa a desarrollar enfermedades cardiovasculares, lo cual permite implementar estrategias de prevención personalizada, como cambios en estilo de vida, dieta saludable y actividad física para reducir el riesgo en individuos con predisposición genética (Wang et al., 2017b).

Las enfermedades neurodegenerativas, como el Alzheimer y el Parkinson, son un grupo de enfermedades progresivas que afectan el sistema nervioso central y provocan una pérdida progresiva de la función cerebral. La omica en el diagnóstico permite identificar biomarcadores asociados con estas enfermedades, el pronóstico permite conocer la progresión de la enfermedad y predecir la tasa de deterioro cognitivo, en el tratamiento permite el desarrollo de fármacos que permitan tratar estas enfermedades y la prevención para identificar individuos con un alto riesgo de desarrollo de este tipo de enfermedades, lo que permite implementar estrategias de prevención personalizadas (Erdaş et al., 2021).

Enfermedades infecciosas, normalmente causadas por patógenos como virus, bacterias y parásitos, que es un problema de salud pública recurrente e importante, la omica en el diagnóstico permite identificar los biomarcadores por medio de detección genético viral, en el tratamiento para desarrollo de nuevos antibióticos y antivirales más efectivos y en la prevención para identificar a las personas más propensas a desarrollar enfermedades infecciosas e implementar estrategias como las vacunas (Chae et al., 2018).

3.1.2. Bases de datos ómicas

La producción de datos omicos incrementa cada año. es por esto que se han establecido diversas bases de datos bioinformáticas que contienen diferentes tipos de datos moleculares, como secuencias de ADN, perfiles de expresión genética, datos de metilación de ADN y variantes genéticas. La adquisición de datos para entrenamiento y validación de los modelos de aprendizaje profundo ya no se considera un problema, en la Tabla 3.1 siguiente se presentan varias bases de datos de uso común en la rama ómica (Zhang et al., 2019).

Este tipo de datos comúnmente tienen formatos del tipo fasta, fastaq, gff2, bed, etc. Propios de su estándar industrial, para aplicación de aprendizaje profundo puede ser necesario conocer lenguajes de programación como Perl, R o Python para la extracción de información y posteriormente ordenar los datos en una forma que los modelos de aprendizaje profundo puedan interpretar como matrices y vectores.

Actualmente, existe una gran cantidad de datos omicos para diversas aplicaciones en la investigación médica actual impulsadas por tecnología de secuenciación de los cuales destaca

Tabla 3.1: Bases de datos ómicos disponibles de libre acceso

Enfoque de la base de datos	Nombre
Datos de genoma	NCBI Ensembl UCSC
Secuenciación de genoma de tipos de cáncer	TCGA
Secuencias de ácidos nucleicos	ENA GenBank DDBJ
Secuencias de proteínas	Swiss-prot PIRR
Estructura proteínicas	PDB
Clasificación de estructuras proteínicas	SCOPe CATH

la investigación del cáncer que es uno de los mayores proveedores de datos omicos moleculares a gran escala, que proporciona soporte de datos integral desde la perspectiva de diferentes procesos biológicos y ayuda a explorar la patogénesis de todo tipo de cáncer y tumores cancerígenos (Li et al., 2024).

3.2. Aprendizaje profundo

El aprendizaje profundo es un campo de la inteligencia artificial, se ha impulsado debido a la disponibilidad de conjuntos de datos, recursos computacionales y algoritmos innovadores. El campo de aplicación es amplio, pero se tienen investigación importante recientes como:

- En la ciencia para análisis de conjunto de datos científicos y realizar descubrimientos en las áreas de la física, química y biología.
- DL aplicado en atención médica, desarrollando herramientas de diagnóstico y tratamiento para enfermedades como el cáncer, enfermedades cardiacas y enfermedades neurológicas.

3.2.1. Aprendizaje profundo en ómica

El DL en el área de la genómica se ha usado para predecir las unidades funcionales de las secuencias de ADN, predecir el dominio de replicación, predicción del factor de transcripción, el punto de iniciación de la transcripción, el promotor, el potenciador y el sitio de borrado del gen (Quang and Xie, 2019; Umarov and Solovyev, 2017; Zeng et al., 2016; Zhang et al., 2017a; Min et al., 2017; Singh et al., 2019; Lee et al., 2015). En los últimos años, se ha impulsado el uso de redes neuronales convolucionales enfocado en la predicción de promotores, potenciadores, dominios de replicación, detección de supresiones genéticas y diferenciación de exones de intrones. El uso de Redes Neuronales Convolucionales (CNNs) se ha impulsado en los últimos años en la predicción de promotores, potenciadores, dominio de replicación, detección de supresiones genéticas y diferenciación de exones de intrones.

También se destaca el uso de aprendizaje profundo para predicción de la expresión genética. Esto implica en la predicción del gen objetivo, de la función génica, modelado de redes de regulación génica, etc. En estas aplicaciones, los datos de entrenamiento utilizados frecuentemente son: secuencias de ADN y datos de modificación de histonas. Las redes neuronales especializadas para esta aplicación son las CNNs y las RNNs (Quang and Xie, 2016; Raza and Alam, 2016; Zhou and Troyanskaya, 2015; Cuperus et al., 2017; Koh et al., 2017).

Se encuentran trabajos importantes del uso de aprendizaje profundo para explorar genomas y enfermedades epigenéticas y otros campos. Los datos de entrenamiento frecuentemente usados son: el mapa genómico, los perfiles de expresión genética y datos clínicos. En este campo de aplicación los esquemas de redes neuronales es más amplio como CNNs, RNNs, Auto-codificadores, Redes Generativas (Liang et al., 2014; Yousefi et al., 2017; Young et al., 2017).

El DL en la transcriptómica se analiza la estructura de secuencias de ARN como los sitios de unión de RBP, sitios de empalme alternativo y los tipos de ARN. Para el entrenamiento, los datos frecuentemente utilizados son las secuencias de ARN, estructuras secundarias y terciarias de ARN y CLIP-seq. En estas aplicaciones las redes CNN y RNN son las más utilizadas (Xu et al., 2017; Zhang et al., 2017b; Pan and Shen, 2017).

Las aplicaciones más relevantes es la asociación entre el ARN y las enfermedades o entre el ARN y el diseño de fármacos. Para el entrenamiento de DL se utilizan datos de secuencias de ARN (miRNA-seq), transcriptómica en el mapa genético y datos metilación de ARN. Para estas aplicaciones el uso de esquemas de redes neuronales es más amplio como CNNs, RNNs, AE y GANs (Chaudhary et al., 2018; Yu et al., 2018; Aliper et al., 2016; Bhat et al., 2016).

El DL en proteómica se usa en la identificación de estructura de proteínas, como la predicción de la estructura terciaria y secundaria de proteínas, evaluación del modelo de proteínas, predicción del mapa de contacto de proteínas, etc. Los datos que se utilizan en el entrenamiento son la secuencias de aminoácidos, estructuras bidimensionales de proteínas y propiedades fisicoquímicas de aminoácidos. En esta aplicación se encuentran las redes neuronales profundas (DNNs) con un cambio al uso de RNNs (Stahl et al., 2017; Li et al., 2017; Spencer et al., 2014; Heffernan et al., 2015).

También se destaca el uso de DL en la predicción de la función de proteínas, donde los datos utilizados para el entrenamiento del modelo son secuencias de aminoácidos, la estructura de la proteína e interacciones proteína-proteína, el tipo de redes más usadas son las CNNs y las RNNs (Kulmanov et al., 2018; Wang et al., 2017a).

En la siguiente Tabla 3.2 se muestra la revisión de los algoritmos de aprendizaje profundo utilizados para propósitos de clasificación, predicción, codificación de los datos, y el tipo de datos utilizados.

En la revisión anterior se puede observar que las principales aplicaciones de los datos omicos es enfocado en identificación y diagnóstico de cáncer debido es una de las enferme-

Tabla 3.2: Revisión del estado del arte de algoritmos utilizados en clasificación, identificación, codificación y clasificación

Algoritmo de DL	Tipo de datos omicos	Predicción	Clasificación	Codificación	Preprocesamiento	Entrenamiento y validación	Propósito	Referencia
MLP (perceptron multicapa), CNN	Transcriptómico, Metabolómicos	x	x	x	x		Clasificación de etapas de cáncer	(Yu et al., 2019)
XomiVAE	Genómico		x	x	x	x	Clasificación de tipo de tumores	(Withnell et al., 2021)
Auto Encoder	Genómico, transcriptómico		x	x	x	x	Clasificación de tipos de cancer	(Franco et al., 2021)
Deep Prog (CNN y ML)	Genómico, transcriptómico	x	x	x	x	x	Clasificación de tipos de cancer	(Poirion et al., 2021)
FactorNet (CNN))	Genómico, transcriptómico	x			x	x	Predicción de factores de transcripción	(Quang and Xie, 2019)
CNN, AE, RNN	Genómico, epigenómico	x			x	x	Predicción metástasis cáncer	(Albaradei et al., 2021)
DCAP	Genómico, epigenómico	x		x	x	x	Predicción tipos cáncer	(Chai et al., 2021)
VAE	Genómico, transcriptómico		x		x	x	Clasificación de tipos de cáncer	(Leng et al., 2022)
AE, VAE, GAN	Genómico, epigenómico	x		x	x	x	Imputacion de datos faltantes	(Huang et al., 2023)
CNN	Genómico, transcriptómico, proteómico	x	x	x	x	x	Identificacion de cancer y clasificacion	(Chuang et al., 2021)
DCNN	Genómico, transcriptómico	x	x		x	x	Identificación de cáncer y clasificación	(Ma and Zhang, 2018)
DNN, CNN	Genómico, transcriptómico	x					Predicción de expresión genética	(Talukder et al., 2021)
CNN	Genómico	x		x	x		Predicción de variantes no codificantes	(Eraslan et al., 2019)
CNN, RNN, AE, GAN	Genómico, transcriptómico	x	x	x	x		Resolución unicelular	(Erfanian et al., 2023)
DeepMO	Genómico, transcriptómico		x	x	x	x	Clasificación de cáncer	(Li et al., 2020)
MOADLN	Genómico, transcriptómico	x	x		x		Clasificación de subtipos de cáncer	(Gong et al., 2023)
CNN	Genómico, transcriptómico, epigenmico	x	x	x	x	x	Clasificación de cáncer	(Li et al., 2022a)
DeepProg	Genómico, transcriptómico, epigenómico	x	x	x	x		Diagnóstico de cáncer	(Mathema et al., 2023)

dades más importantes actualmente, por esto mismo se ha desarrollado una mayor cantidad de bases de datos. El tipo de algoritmos de aprendizaje profundo más utilizados son el CNN en aplicaciones de predicción y clasificación, seguido de las RNN.

El análisis de datos omicos actualmente se centra en el cáncer y, por lo tanto, se ubica un mayor número de bases de datos genómicos, transcriptómicos, proteómicos y metabolómicos. Esto presenta una ventaja debido a que se tiene un mayor número de datos de entrenamiento y validación para los investigadores que se dedican al desarrollo de algoritmos de aprendizaje profundo. En la Tabla 3.3 se muestran las distintas bases de datos de distintos tipos de cáncer, destacando la TCGA debido a cuenta con una biblioteca para la descarga por lotes de datos y que ha sido de mayor uso en los trabajos de investigación (Li et al., 2022b).

Tabla 3.3: Bases de datos enfocadas en el análisis de datos omicos de tipos de cáncer.

Nombre de la base de datos	Tipo de datos	Acceso a tipos de cáncer	Formatos de descarga
TCGA	Genómica, transcriptómica, proteómica, metabolómica	33 tipos	JSON, TSV
GEO	Genómica, transcriptómica, proteómica	20 tipos	JSON, TSV
ICGC	Genómica, transcriptómica, proteómica, metabolómica	50 tipos	JSON, TSV
CBioPortal	Genómica, transcriptómica, proteómica	200 tipos	JSON, TSV
NCI	Genómica, transcriptómica, proteómica	30 tipos	JSON, TSV
CBDiscovery	Genómica, transcriptómica, proteómica	25 tipos	JSON, TSV

De esta revisión se logra precisar sobre el tipo de datos que se seleccionaran para el diseño del algoritmo de red neuronal, debido a que los datos enfocados en el cáncer cuentan con el mayor número de bases de datos del cual adquirir conjuntos de entrenamiento. En la Tabla 3.4 se toman trabajos con este enfoque.

El creciente éxito de aprendizaje profundo ha impulsado el desarrollo de software de código abierto, en las que se encuentran las más populares como tensorflow, Caffe, Torch y CNTK. Estas herramientas son compatibles con CPU multinucleo y GPU multinucleo(Shi et al., 2016; Liu et al., 2020).

Tensorflow desarrollador por Google integra unidades mas comunes en el marco del aprendizaje profundo. Soporta redes actualizadas como CNN y RNN con diferentes configuraciones. Este marco esta diseñado para ofrecer flexibilidad, portabilidad y alta eficiencia del hardware equipado.

Caffe desarrollador por Berkeley Vision and Learning Center (BVLC) y es de código abierto desde 2014. Caffe puede procesar 40 millones de imágenes al día con la version acelerada por GPU en una sola tarjeta GPU NVIDIA k40 o Titan. Con integración cuDNN, se consigue otra aceleración 1.3k (Chethur et al., 2014).

CNTK es un conjunto de herramientas de redes computacionales unificadas desarrolladas por Microsoft Research, que admite muchas redes neuronales populares. Este marco con múltiples GPU cuenta con un rendimiento bastante aceptable comparándolo con otros(Huang, 2015).

Tabla 3.4: Revision de articulos enfocados en cancer destacando conjunto de entrenamiento, tipos de datos y base de datos utilizada.

Proposito	Conjuto de datos	Genomica	Transcriptómi-ca	Base de datos	Otros	Referencia
Pronostico de supervivencia	15 tipos	x	x	TCGA		(Huang et al., 2023)
Clasificación de cáncer	11 tipos		x	TCGA	Datos clinicos	(Chuang et al., 2021)
Clasificación de cáncer	33 tipos	x	x	TCGA		(Franco et al., 2021)
Predicción de supervivencia	32 conjuntos multiomicos	x	x	TCGA		(Chuang et al., 2021)
Clasificación de cáncer	33 tipos	x	x	TCGA		(Withnell et al., 2021)
Clasificación de cáncer y agrupamiento de cáncer	5 tipos	x	x	TCGA		(Chuang et al., 2021)
Clasificación de cáncer	33 tipos	x	x	TCGA		(Zhang et al., 2019)
Predicción de supervivencia	1 tipo	x	x	TCGA		(Tong et al., 2020)

Torch es un marco de computación científica que proporciona estructuras de datos para los componentes más útiles en algoritmos de aprendizaje automático, como tensores multi-dimensionales y operaciones matemáticas sobre ellos.

4. Propuesta de tesis

4.1. Objetivo general

- Modelar datos ómicos (genómicos, transcriptómicos y proteómicos) utilizando técnicas de aprendizaje profundo, empleando redes neuronales de diseño propio, y configurar la salida del modelo para diferentes propósitos de clasificación y predicción.

4.2. Objetivos específicos

- Codificar datos genómicos, transcriptómicos y proteómicos para ser alimentados en las redes neuronales.
- Implementar redes neuronales convolucionada (CNN) y recurrente (RNN), y entrenarlas.
- Proponer y diseñar una red neuronal propia a partir de las dos anteriores.
- Configurar en cada caso la salida, si es un clasificador, un predictor, un regresor o un generador de señal.
- Enfocar el modelado de las redes neuronales para fines biomédicos.
- Validar los resultados con las bases de datos e incluyendo opinión de especialistas.

4.3. Antecedentes

El aprendizaje profundo actualmente tiene como base de procesamiento a las redes neuronales, en CENIDET se cuentan con trabajos relacionados con la aplicación de algoritmos de redes neuronales artificiales.

Uno de los principales intereses es la participación con especialistas que participan en el área médica en el centro oncológico de San Peregrino Cancer Center, donde se tiene un convenio de participación. De misma manera se tiene una colaboración con la Universidad de Grenoble, donde se ha estado trabajando con datos omicos.

4.4. Planteamiento del problema

La conjunción del aprendizaje profundo en el área biomédica recientemente está dando resultados, como es una tecnología relativamente nueva, existen múltiples problemáticas por abordar como la alta dimensionalidad de datos, datos desequilibrados, explicabilidad de los modelos, estandarización de datos de las bases públicas, la imputación de datos y la clasificación errónea.

Las áreas de oportunidad que se plantean abordar son la codificación de datos en un formato que pueda ser interpretado y analizado por el modelo de red neuronal, donde se considera la normalización de los datos adquiridos, imputación y reducción de dimensiones con el fin de aumentar la precisión del modelo de DL.

Los algoritmos de aprendizaje profundo en aplicaciones de clasificación y predicción utilizando datos omicos están lejos de ser óptimos debido a la complejidad de los tipos de datos y los problemas antes mencionados, lo que se busca abordar es en la propuesta de un nuevo algoritmo que tome las ventajas que tienen otros e incorporarlas, ya que como se propone vincular con el área médica se requiere una alta precisión en las respuestas que se obtienen del modelo.

4.5. Pregunta de investigación

¿Es posible proponer un modelo de red neuronal basado en aprendizaje profundo que ayude a aumentar la precisión en la clasificación/predicción de un fenotipo utilizando datos genómicos, transcriptómicos y genómicos que pueda utilizarse en el área biomédica?

4.6. Justificación

El uso de datos omicos de diversas fuentes públicas presenta problemas como la heterogeneidad y datos desequilibrados (datos faltantes y/o mal etiquetados), con el uso de algoritmos de aprendizaje profundo enfocados en la imputación y codificación se podría mejorar la eficiencia de predicción y clasificación del modelo para pronóstico del cáncer. La integración con el área biomédica supone una ventaja en el área biomédica para una evaluación temprana en

pacientes con un tipo de cáncer y determinar la progresión con el cual se pueden establecer tratamientos adecuados.

5. Cronograma de actividades

En esta sección se muestra es cronograma de las actividades que de manera preliminar se planean abordar los 4 años de estancia en CENIDET.

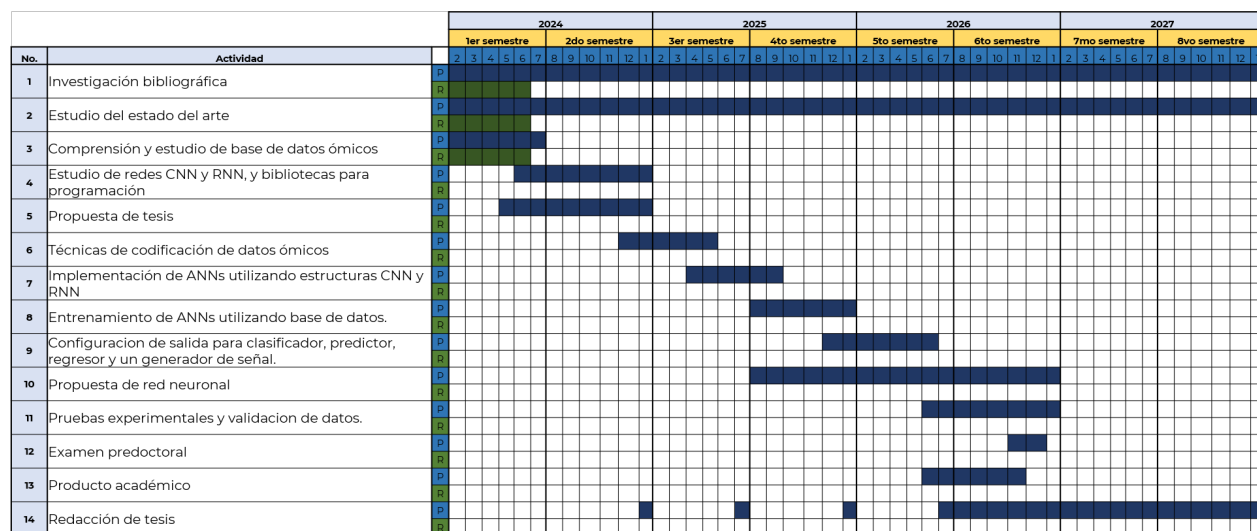


Figura 5.1: Cronograma de actividades

Referencias

- Albaradei, S., Thafar, M., Alsaedi, A., Van Neste, C., Gojobori, T., Essack, M., and Gao, X. (2021). Machine learning and deep learning methods that use omics data for metastasis prediction. *Computational and structural biotechnology journal*, 19:5008–5018.
- Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., and Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7):2524–2530.
- Bhat, R. R., Viswanath, V., and Li, X. (2016). Deepcancer: Detecting cancer through gene expressions via deep generative learning. *arXiv preprint arXiv:1612.03211*.
- Carlomagno, N., Incollingo, P., Tammara, V., Peluso, G., Rupealta, N., Chiacchio, G., Sandoval Sotelo, M. L., Minieri, G., Pisani, A., Riccio, E., et al. (2017). Diagnostic, predictive, prognostic, and therapeutic molecular biomarkers in third millennium: a breakthrough in gastric cancer. *BioMed research international*, 2017.
- Chae, S., Kwon, S., and Lee, D. (2018). Predicting infectious disease using deep learning and big data. *International journal of environmental research and public health*, 15(8):1596.
- Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., and Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in biology and medicine*, 134:104481.
- Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259.
- Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., and Shelhamer, E. (2014). cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*.
- Chuang, Y.-H., Huang, S.-H., Hung, T.-M., Lin, X.-Y., Lee, J.-Y., Lai, W.-S., and Yang, J.-M. (2021). Convolutional neural network for human cancer types prediction by integrating protein interaction networks and omics data. *Scientific reports*, 11(1):20691.
- Cuperus, J. T., Groves, B., Kuchina, A., Rosenberg, A. B., Jojic, N., Fields, S., and Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5 untranslated regions from 500,000 random sequences. *Genome research*, 27(12):2015–2024.
- Davis, M. and Shanley, T. (2017). The missing-omes: Proposing social and environmental nomenclature in precision medicine. *Clinical and translational science*, 10(2):64.
- Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108.
- Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403.

- Erdaş, Ç. B., Sümer, E., and Kibaroglu, S. (2021). Neurodegenerative disease detection and severity prediction using deep learning approaches. *Biomedical Signal Processing and Control*, 70:103069.
- Erfanian, N., Heydari, A. A., Feriz, A. M., Iañez, P., Derakhshani, A., Ghasemigol, M., Farahpour, M., Razavi, S. M., Nasser, S., Safarpour, H., et al. (2023). Deep learning applications in single-cell genomics and transcriptomics data analysis. *Biomedicine & Pharmacotherapy*, 165:115077.
- Franco, E. F., Rana, P., Cruz, A., Calderon, V. V., Azevedo, V., Ramos, R. T., and Ghosh, P. (2021). Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. *Cancers*, 13(9):2013.
- Géron, A. (2020). Aprende machine learning con scikit-learn, keras y tensorflow. *España: Anaya*.
- Gong, P., Cheng, L., Zhang, Z., Meng, A., Li, E., Chen, J., and Zhang, L. (2023). Multi-omics integration method based on attention deep learning network for biomedical data classification. *Computer Methods and Programs in Biomedicine*, 231:107377.
- Hagan, M., Demuth, H., Beale, M., and De Jesús, O. (2014). *Neuronal Network Design*. Martin Hagan, 2nd edition.
- Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, 18:1–15.
- Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., and Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific reports*, 5(1):11476.
- Huang, L., Song, M., Shen, H., Hong, H., Gong, P., Deng, H.-W., and Zhang, C. (2023). Deep learning methods for omics data imputation. *Biology*, 12(10):1313.
- Huang, X. (2015). Microsoft computational network toolkit offers most efficient distributed deep learning computational performance.
- Javier, B., Anna, G., and Núria, M. (2019). Fundación Instituto Roche. Accessed Mayo 15, 2024.
- Kim, M. and Tagkopoulou, I. (2018). Data integration and predictive modeling methods for multi-omics datasets. *Molecular omics*, 14(1):8–25.
- Koh, P. W., Pierson, E., and Kundaje, A. (2017). Denoising genome-wide histone chip-seq with convolutional neural networks. *Bioinformatics*, 33(14):i225–i233.
- Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. (2020). State of the field in multi-omics research: from computational needs to data mining and sharing. *Frontiers in Genetics*, 11:610798.

- Kulmanov, M., Khan, M. A., and Hoehndorf, R. (2018). Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.
- Lee, B., Lee, T., Na, B., and Yoon, S. (2015). Dna-level splice junction prediction using deep recurrent neural networks. *arXiv preprint arXiv:1512.05135*.
- Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., Wang, M., Zhang, Z., He, S., and Bo, X. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome biology*, 23(1):171.
- Li, H., Hou, J., Adhikari, B., Lyu, Q., and Cheng, J. (2017). Deep learning methods for protein torsion angle prediction. *BMC bioinformatics*, 18:1–13.
- Li, M., Guo, H., Wang, K., Kang, C., Yin, Y., and Zhang, H. (2024). Avbae-modfr: A novel deep learning framework of embedding and feature selection on multi-omics data for pan-cancer classification. *Computers in Biology and Medicine*, page 108614.
- Li, T., Tong, W., Roberts, R., Liu, Z., and Thakkar, S. (2020). Deep learning on high-throughput transcriptomics to predict drug-induced liver injury. *Frontiers in bioengineering and biotechnology*, 8:562677.
- Li, Y., Wu, X., Yang, P., Jiang, G., and Luo, Y. (2022a). Machine learning for lung cancer diagnosis, treatment, and prognosis. *Genomics, Proteomics and Bioinformatics*, 20(5):850–866.
- Li, Z., Ma, Z., Zhou, Q., Wang, S., Yan, Q., Zhuang, H., Zhou, Z., Liu, C., Wu, Z., Zhao, J., et al. (2022b). Identification by genetic algorithm optimized back propagation artificial neural network and validation of a four-gene signature for diagnosis and prognosis of pancreatic cancer. *Heliyon*, 8(11).
- Liang, M., Li, Z., Chen, T., and Zeng, J. (2014). Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(4):928–937.
- Liu, J., Li, J., Wang, H., and Yan, J. (2020). Application of deep learning in genomics. *Science China Life Sciences*, 63:1860–1878.
- Ma, S. and Zhang, Z. (2018). Omicsmapnet: Transforming omics data to take advantage of deep convolutional neural network for discovery. *arXiv preprint arXiv:1804.05283*.
- Mamoshina, P., Vieira, A., Putin, E., and Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Molecular pharmaceuticals*, 13(5):1445–1454.

- Mathema, V. B., Sen, P., Lamichhane, S., Orešič, M., and Khoomrung, S. (2023). Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. *Computational and Structural Biotechnology Journal*, 21:1372–1382.
- Min, X., Zeng, W., Chen, S., Chen, N., Chen, T., and Jiang, R. (2017). Predicting enhancers with deep convolutional neural networks. *BMC bioinformatics*, 18:35–46.
- Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, 18(3):11–11.
- Munir, K., Elahi, H., Ayub, A., Frezza, F., and Rizzi, A. (2019). Cancer diagnosis using deep learning: a bibliographic review. *Cancers*, 11(9):1235.
- Pan, X. and Shen, H.-B. (2017). Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18:1–14.
- Pasha, S. N., Ramesh, D., Mohmmad, S., Harshavardhan, A., et al. (2020). Cardiovascular disease prediction using deep learning techniques. 981(2):022006.
- Poirion, O. B., Jing, Z., Chaudhary, K., Huang, S., and Garmire, L. X. (2021). Deepprog: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome medicine*, 13:1–15.
- Quang, D. and Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11):e107–e107.
- Quang, D. and Xie, X. (2019). Factornet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47.
- Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21.
- Raza, K. and Alam, M. (2016). Recurrent neural network based hybrid model for reconstructing gene regulatory network. *Computational biology and chemistry*, 64:322–334.
- Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49:107739.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.
- Rouhiainen, L. (2018). Inteligencia artificial. *Madrid: Alienta Editorial*, pages 20–21.
- Sarmiento-Ramos, J. L. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, 19(4):1–18.

- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.
- Shi, S., Wang, Q., Xu, P., and Chu, X. (2016). Benchmarking state-of-the-art deep learning software tools. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, pages 99–104. IEEE.
- Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBE)*, pages 1–6. IEEE.
- Singh, S., Yang, Y., Póczos, B., and Ma, J. (2019). Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology*, 7(2):122–137.
- Spencer, M., Eickholt, J., and Cheng, J. (2014). A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 12(1):103–112.
- Stahl, K., Schneider, M., and Brock, O. (2017). Epsilon-cp: using deep learning to combine information from multiple sources for protein contact prediction. *BMC bioinformatics*, 18:1–11.
- Strimbu, K. and Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463–466.
- Talukder, A., Barham, C., Li, X., and Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, 22(3):bbaa177.
- Tong, L., Mitchel, J., Chatlin, K., and Wang, M. D. (2020). Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC medical informatics and decision making*, 20:1–12.
- Umarov, R. K. and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PloS one*, 12(2):e0171410.
- Van Eyk, J. E. and Snyder, M. P. (2018). Precision medicine: role of proteomics in changing clinical management and care. *Journal of proteome research*, 18(1):1–6.
- Vieira, S., Pinaya, W. H. L., and Mechelli, A. (2020). Main concepts in machine learning. In *Machine learning*, pages 21–44. Elsevier.
- Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., and Xu, D. (2017a). Musitedeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, 33(24):3909–3916.
- Wang, J., Ding, H., Bidgoli, F. A., Zhou, B., Iribarren, C., Molloy, S., and Baldi, P. (2017b). Detecting cardiovascular disease from mammograms with deep learning. *IEEE transactions on medical imaging*, 36(5):1172–1181.

- Withnell, E., Zhang, X., Sun, K., and Guo, Y. (2021). Xomivae: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in bioinformatics*, 22(6):bbab315.
- Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017). Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to esc fate decision. *Nucleic acids research*, 45(21):12100–12112.
- Young, J. D., Cai, C., and Lu, X. (2017). Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC bioinformatics*, 18:5–17.
- Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Velazquez Vega, J. E., Brat, D. J., et al. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):1–11.
- Yu, H., Samuels, D. C., Zhao, Y.-y., and Guo, Y. (2019). Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC genomics*, 20:1–12.
- Yu, L., Sun, X., Tian, S., Shi, X., and Yan, Y. (2018). Drug and nondrug classification based on deep learning with various feature selection strategies. *Current Bioinformatics*, 13(3):253–259.
- Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting dna–protein binding. *Bioinformatics*, 32(12):i121–i127.
- Zhang, S., Hu, H., Jiang, T., Zhang, L., and Zeng, J. (2017a). Titer: predicting translation initiation sites by deep learning. *Bioinformatics*, 33(14):i234–i242.
- Zhang, Y.-z., Yamaguchi, R., Imoto, S., and Miyano, S. (2017b). Sequence-specific bias correction for rna-seq data using recurrent neural networks. *BMC genomics*, 18:1–6.
- Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., and Peng, S. (2019). Deep learning in omics: a survey and guideline. *Briefings in functional genomics*, 18(1):41–57.
- Zhao, Y.-Y., Cheng, X.-l., and Lin, R.-C. (2014). Lipidomics applications for discovering biomarkers of diseases in clinical chemistry. *International review of cell and molecular biology*, 313:1–26.
- Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934.