	CÓDIGO: 76621650450	
	VERSIÓN: 4	
	FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021	

## INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD

### Red neuronal convolucional para la curación de secuencias de ADN de LTR

#### retrotransponibles en plantas

Santiago Alba Iriarte

Ingeniería Biomédica

Simón Orozco Arias

Tutor

Reinel Tabares Soto

co-Tutor

PROCESO DE INVESTIGACIÓN II

UNIVERSIDAD AUTÓNOMA DE MANIZALES


PERIODO 3, AÑO 2021

MANIZALES

	CÓDIGO: 76621650450	
	VERSIÓN: 4	
	FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV///2021	

## TABLA DE CONTENIDO

<b>1. INTRODUCCIÓN .....</b>	<b>3</b>
<b>2. REFERENTE TEÓRICO .....</b>	<b>4</b>
<b>3. PROBLEMA DE INVESTIGACIÓN Y JUSTIFICACIÓN .....</b>	<b>5</b>
<b>4. OBJETIVOS.....</b>	<b>6</b>
<b>5. METODOLOGÍA.....</b>	<b>7</b>
<b>6. INFORME DE RESULTADOS .....</b>	<b>9</b>
<b>6.1 Cronograma de actividades.....</b>	<b>9</b>
<b>6.2 Resultados Parciales.....</b>	<b>10</b>
<b>7. CONCLUSIONES.....</b>	<b>16</b>
<b>8. RECOMENDACIONES .....</b>	<b>16</b>
<b>9. ANEXOS.....</b>	<b>17</b>
<b>10. BIBLIOGRAFÍA.....</b>	<b>17</b>

		
	<b>INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD</b>	
	<b>CÓDIGO: 76621650450</b> <b>VERSIÓN: 4</b> <b>FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021</b>	


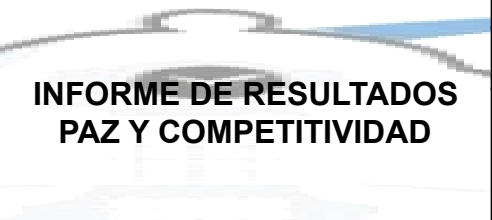
## 1. INTRODUCCIÓN

Los elementos transponibles, también son conocidos como “genes saltarines”, estos son secuencia de ADN con la capacidad de reubicarse de posición en el genoma. Desde sus inicios se realizan estudios en verificar y observar sus impactos en la evolución de los organismos eucariotas, por el hecho de tener la capacidad de cambiar de posición en el genoma, ocasionando alteraciones en las formas de los genes, causar mutaciones y siendo potencialmente valiosos para la evolución[1]. Actualmente existen dos principales clases, los **retrotransposones** y los **transposones de ADN**, el primero son secuencias de ADN, con la capacidad de “copiar y pegar” su propia información genética y llevarla a otro sitio del genoma, y esto lo hace imitando algunos de los virus de ADN, entre los cuales podemos encontrar los bicatenario retrotranscrito, este transcribe su información a ARN, para luego una transcriptasa reversa, permitirle convertir el ARN en ADN en cualquier región del genoma. La segunda gran clase transponible son los transposones de ADN. Estos poseen los mecanismos de transposición. Uno de ellos es el de “cortar y pegar”, en el que el transposón se separa de un lugar del genoma y se inserta en otro. [1]

Por otro lado, se han llevado a cabo diferentes almacenamientos y recolección de secuencias de ADN, con el propósito de ser usado por investigadores, en el campo de la bioinformática, por lo tanto se han llevado propuestas para clasificar secuencias no homólogas. por medio de elementos transponibles (TERL), que pre procesa y transforma las secuencias unidimensionales en un espacio de datos bidimensional (es decir, datos similares a imágenes de las secuencias) y lo aplica a redes neuronales convolucionales profundas.

El método consolidado por investigadores para clasificar varias superfamilias de secuencias de ET, proporciona buenos resultados, un ejemplo claro fue el artículo[2], el cual nos da una investigación base de la construcción de una red neuronal para tareas de bioinformática, la cual fue tuneada antes de llevar a cabo los diferentes experimentos de la investigación y mejorar su resultado y aplicarla a la curación secuencias de elementos retrotransponibles.

Esta investigación tiene como enfoque tomar como etapa inicial la red mencionada anteriormente, y aplicar dicha red a una base de datos consolidada por el grupo el semillero, la cual sea apropiada para detectar anomalías en la cadena de ADN y en caso de que encuentre algún nucleótido distinto a los generales [A, C, G, T], determinarlo como “N”, lo anterior con el fin de que sea útil la investigación para la revisión de las secuencias de LTR

		
	<b>CÓDIGO: 76621650450</b>	
	<b>VERSIÓN: 4</b>	
<b>INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD</b>		<b>FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021</b>

en plantas, y tener catalogado dicho proceso y preprocesamiento de los datos para crear una matriz bidimensional y poder utilizar redes neuronales convolucionales(CNNs), además se propende evaluar los diferentes hiper-parámetros para proporcionar mejoramiento de un proceso teórico en la práctica.

## 2. REFERENTE TEÓRICO

Actualmente por medio de los avances computacionales se han llevado diferentes desarrollo de herramientas en bioinformática para la identificación de diferentes tipos de genomas en secuenciación y optimizaciones en la clasificación, es por ende que el estudio de los retrotransposones LTR juegan un papel fundamental en la evolución y la diversidad genética, por lo que nace la importancia de entender su función y profundidad respecto a las variaciones que pueden presentar por lo que el objetivo principal es reducir el tiempo de ejecución. Para un curador automático de diferentes bibliotecas de retrotransposones LTR de plantas, basados en Deep Learning (DL), en donde se han llevado proyectos con métricas de F1-score del 91,18% , utilizando cuatro genomas diferentes, además los mejores resultados poseen un rendimiento del 93,6% de F1-score, y con un tiempo de ejecución de 22,61 segundos para la predicción por la red neuronal, se evidencia que con el aumento de la secuenciación del genoma completo, es necesario automatizar el proceso de análisis.[3]

Por otra parte, abordar la detección y clasificar elementos transponibles (TEs) supone tediosas tareas que implican métodos bioinformáticos, es por esto que , se han evaluados mediante técnicas de ML sobre conjuntos de datos de TÉ, lo que han demostrado que la selección de las métricas que miden el rendimiento de los modelos poseen características específicas y , aunque la forma más utilizada para comparar las medidas es mediante un análisis empírico, estas propiedades se calculan sobre la base de datos, si una medida determinada cambia su valor bajo ciertas modificaciones en la matriz de confusión, brindan parámetros comparativos independientes de los conjuntos de datos. Por lo que se han analizado 26 métricas diferentes utilizadas en clasificaciones binarias, multiclase y jerárquicas, a través de diferentes fuentes, los hallazgos utilizando conjuntos de datos de TÉ disponibles libremente y algoritmos de ML comúnmente utilizados. han demostrado que las métricas más adecuadas para las tareas de TÉ deben ser la puntuación F1 y el accuracy. [4]

Además se toma en consideración que, cada día publican nuevos proyectos de secuenciación masiva (es decir, que pretenden secuenciar miles de individuos). Sin embargo, no existen suficientes herramientas automáticas para analizar esta gran cantidad

	CÓDIGO: 76621650450	
	VERSIÓN: 4	
	FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021	

## INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD


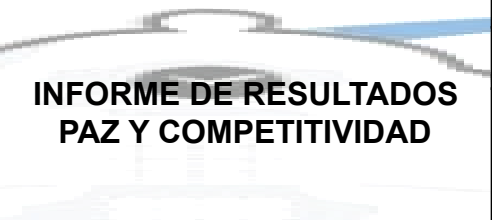
de información genómica.respecto a los retrotransposones LTR, debido a que son las secuencias repetitivas más frecuentes en los genomas de las plantas, sin embargo, su detección y clasificación se suele realizar mediante programas semiautomáticos y manuales que suelen ser muy demorados y consumen fuera de tiempo gran computo. A pesar de la disponibilidad de varias herramientas bioinformáticas para detectarlos y clasificarlos, ninguna de estas ha brindado de manera individual resultados precisos. por lo que se utilizaron algoritmos de Machine Learning como lo son k-men para clasificar los retrotransposones LTR son resultados de F1-Score del 95%, lo cual en su propone un estudio base inicial y automático para analizar secuencias.[5]

Los elementos transponibles (TEs) han demostrado que debido a su número natural de repeticiones y a su gran diversidad estructural, la identificación y clasificación han brindado un campo de investigación amplio, demostrando un campo importante ya que permitiría la regulación de los genes, así como en la adaptación y la evolución de los mismos , por lo que estas clases son cruciales para comprender mejor las funciones del genoma y su evolución. Actualmente se han desarrollado diferentes software para los procesos de detección y clasificación de TE, pero a esto abarcan diferentes dificultades o problemas como lo son la precisión y la velocidad de los análisis. [6]

Adicional a lo anterior, añadiendo una etapa previa, cuyo objetivo es detectar la necesidad de una revisión, posteriormente se formulan y ejecutan ecuaciones de búsqueda en varias bases de datos bibliográficas, analizando publicaciones y preguntas de investigación de diferentes investigadores, con lo cual se sustentan varios enfoques de ML en otros problemas bioinformáticos con soluciones prometedoras, en donde se encuentran algoritmos y arquitecturas disponibles en la literatura, centrados específicamente en los TEs. A pesar de representar la mayor parte del ADN de muchos organismos, sólo se encontraron 35 artículos y se clasificaron como relevantes en TE o campos relacionados donde se conmesura que el ML es un poderoso y activo instrumento que puede, ser útil y su utilización en los análisis de TE es todavía limitada, solo ha sido posible constatar que el uso de ML para los análisis de TE (detección y clasificación) lo cual es un problema abierto y se ha convertido en un nuevo campo de investigación.[7]

### 3. PROBLEMA DE INVESTIGACIÓN Y JUSTIFICACIÓN

Dentro de los procesos de secuenciación de ADN, enfocados a los elementos transponibles los cuales son genes capaces de modificar su posición de manera significativa

		
	<b>CÓDIGO: 76621650450</b> <b>VERSIÓN: 4</b>	
	<b>INFORME DE RESULTADOS</b> <b>PAZ Y COMPETITIVIDAD</b> <b>FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021</b>	

en una cadena de ADN, alterando procesos y comportamientos de la misma, dichos cambios generan alteraciones a nivel de los genes con lo cual en el proceso macro de crecimiento de algún ser vivo, este puede verse con inconvenientes para realizar un proceso de evolución natural; se tiene en cuenta que dichos procesos de secuenciación requieren de ciertas bases en cuanto a clasificación y forma.[8][9]

La clasificación de las secuencias de ADN, el proceso tiene en cuenta que la cadena o secuencia debe sufrir ciertos cambios con el fin de quedar con un preproceso antes de la clasificación de sus componentes o elementos fundamentales y que la secuencia se encuentre con características uniformes o dentro de un marco estándar. Dentro del , pretratamiento de la secuencia y por medio de, herramientas de bioinformática es posible usar diferentes técnicas y métodos para realizar lo que se conoce como “curación de secuencias de ADN”, técnica usada en varios procesos para análisis de los elementos retrotransposones (RTS). [10] [11]

La desventaja frente a los métodos actuales es que, aunque hay técnicas, instrumentos y herramientas, para realizar la curación, este proceso conlleva uso de tiempo considerable, requiere en su mayoría métodos manuales, los cuales al proceso le restan precisión, credibilidad, y optimización. Cabe resaltar que el proceso se muestra ineficiente debido a que el material genético tiene un amplio volumen y el protocolo de procesamiento es extenso y si bien entra en un proceso estándar, es complejo y tedioso. [7][12]

Tomando en cuenta un punto de vista cuantitativo y cualitativo del proceso se pretende por medio de la consolidación de diferentes bases de datos analizar los parámetros, con los cuales crear una herramienta con el objetivo de encontrar anomalías en cadenas de ADN en plantas, pero con una barrera o fronteras en cuanto a un vacío en el conocimiento y ejecución frente a un tipo de datos, del cual se tiene desconocimiento y poco sustento teórico, por ende el medio por el cual se evalúa es a través de análisis previos y formación controlada o por pruebas de la arquitectura deseada. Es por este motivo que se plantea el diseño y adaptación de las redes ya existentes, con el objetivo de aplicarlo a la bioinformática, en este caso a la curación de secuencias de ADN de LTR retro transponibles en plantas, con el objetivo de brindar a futuros investigadores, previos estudios y ampliar los conocimientos en este campo, y proporcionar confiabilidad y generalización frente a nuevas secuencias, validando la investigación presente en esta investigación. [13]

#### 4. OBJETIVOS

**Objetivo General:** Desarrollar modelos computacionales de aprendizaje profundo y procesamiento de datos, utilizando redes neuronales convolucionales para catalogar de

	CÓDIGO: 76621650450	
	VERSIÓN: 4	
	FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021	

forma rápida, económica y precisa, una secuencia de ADN de LTR retro transponibles en planta donde se encuentren anomalías.

#### **Objetivos Específicos:**

- Establecer un flujo de trabajo para el afinamiento de hiper-parámetros de una red neuronal convolucional para el filtrado de secuencias de ADN
- Diseñar una red neuronal convolucional para la curación automática de LTR retrotransposones de genomas de plantas.
- Evaluar el desempeño y generalización del modelo, brindando calidad en la clasificación de una nueva secuencia de ADN

## **5. METODOLOGÍA**


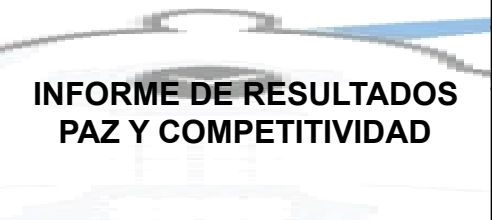
Con el fin de dar solución al problema de investigación, Inicialmente se planteó la búsqueda del estado del arte de diferentes autores con el objetivo de contribuir a sus investigaciones tomando las redes neuronales preexistentes con el motivo de mejorarlas y adaptar su modelo a nuestra problemática.[2]

Dentro de las rúbricas de búsqueda se toman en cuenta datos o papers de revistas indexadas, y las claves de búsqueda tanto en inglés y español como: Elementos transponibles y redes neuronales, secuencias de ADN y métodos de machine learning, proceso de secuenciación de ADN, métodos de clasificación de ADN y elementos retrotransponibles. Para Luego se conllevar a la búsqueda de bases de datos, de las grandes familias de cadenas de ADN de elementos Retro Transponibles, y utilizarlas con las diversas arquitecturas encontradas y mejoradas en esta investigación con la red neuronal.

Una vez concluidos los literales anteriores, el siguiente paso se delimitó en observar las métricas y diferir qué procesos y mejoras podría darle a la red para optimizarla, y subir sus porcentajes de precisión con los diversos métodos, de tuneo de hiperparametros que podemos realizarle a la red, gracias a la API de Keras.

Como paso final sería utilizar las métricas empleadas por diversos investigadores tal como el f1-score, y la matriz de confusión para evaluar el modelo y observar qué tan bien está generalizando nuestra red frente a nuevas secuencias de datos de ADN.

Por otro lado el Autor Ping Shung, y su artículo “Accuracy, precision, recall or f1?”[14]. nos habla de las diferencias métricas que debemos de tener en cuenta, a la hora de evaluar un modelo las cuales son:

		
	<b>INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD</b>	
	<b>CÓDIGO: 76621650450</b> <b>VERSIÓN: 4</b> <b>FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021</b>	

Acuracy: mide el porcentaje global de muestras que el modelo ha clasificado correctamente.

Acuracy:  $TP + TN$

Donde TN es Verdadero Negativo, TP es Verdadero Positivo, FP es Falso Positivo, FN corresponde a Falso Negativo.



Recall: Esta métrica nos va a informar sobre la **cantidad de datos aciertos** que el modelo es capaz de identificar.

Precisión: Es la fracción de todas las instancias relevantes dividida por las instancias obtenidas. Se utiliza para medir la calidad del modelo, identificando las predicciones positivas que fueron realmente correctas.

F1-Score: se utiliza para combinar las medidas de precisión y recall medidas en un solo valor. Esto es práctico porque facilita la comparación del rendimiento combinado de precisión y exhaustividad (recall) entre varias soluciones, independientemente de si el conjunto de pruebas está equilibrado o no.

Las métricas anteriores fueron utilizadas en la presente investigación con el objetivo de presentar un trabajo con diferentes resultados, pre procesamientos y tuneos de parámetros realizados, y la mejora que se ha venido realizando a la red tomada del estado del arte.

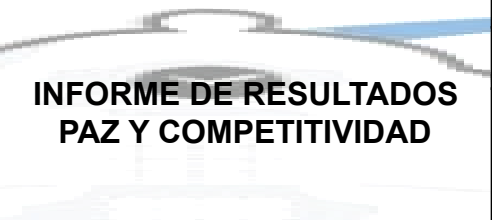


		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

## 6. INFORME DE RESULTADOS

### 6.1 Cronograma de actividades

Actividades	% de Cumplimiento
Apoyar las pruebas de validación de los modelos computacionales implementados.	85%
Documentar los resultados y métodos empleados para la mejora de la red propuesta.	85%
Elaborar una recopilación de información de los modelos encontrados en la literatura, para el análisis y comparación de resultados.	70%
Elaborar el protocolo para el flujo de la información desde y hacia la red.	72%
Elaborar el protocolo de los datos de la biblioteca de cadena de ADN, para ser empleadas en el modelo.	70%
Elaborar un modelo de viabilidad económica para la implementación.	60%
Elaborar un artículo donde se provee de toda la información obtenida.	70%

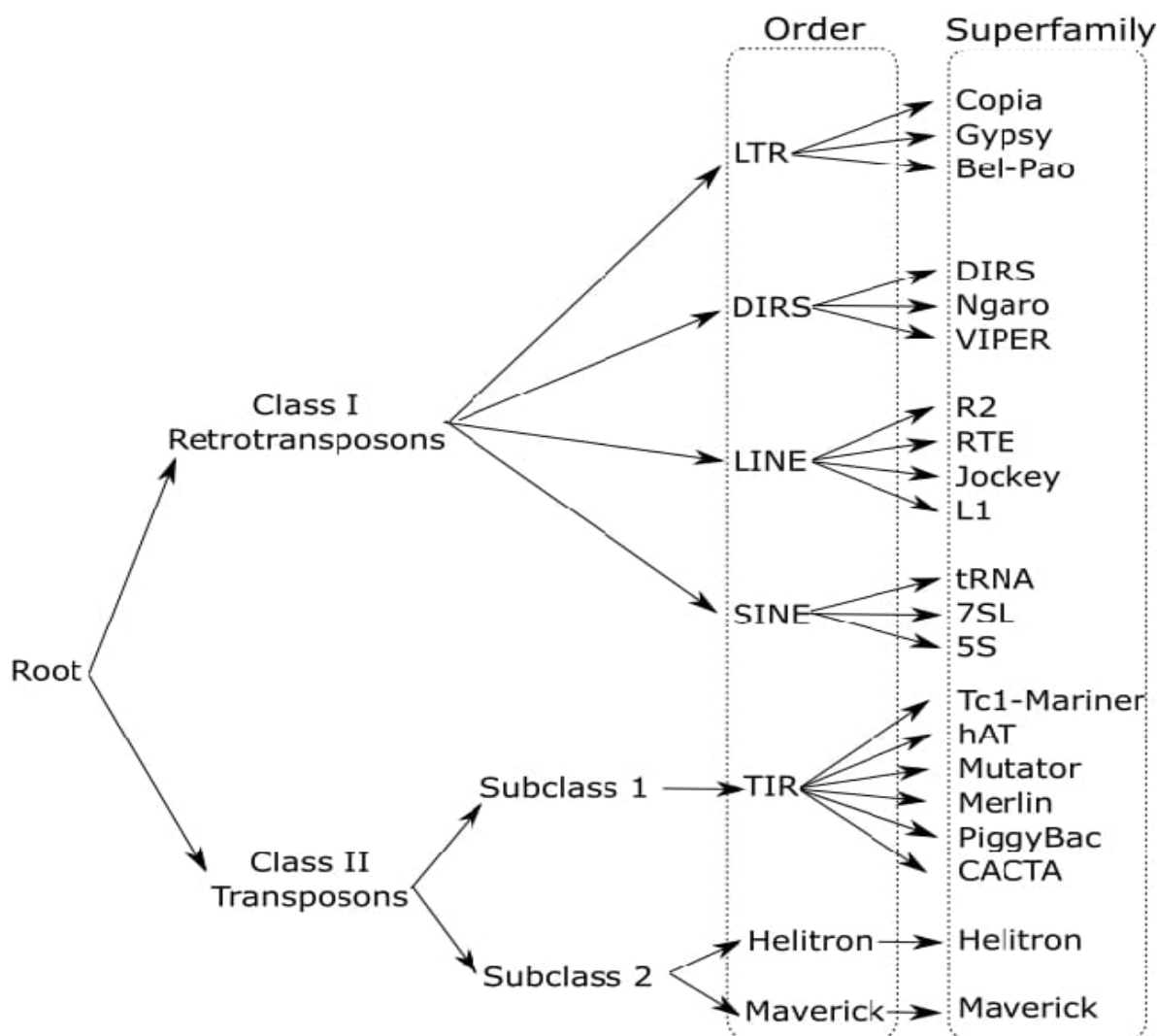
		
	<b>CÓDIGO: 76621650450</b> <b>VERSIÓN: 4</b>	
	<b>INFORME DE RESULTADOS PAZ Y COMPETITIVIDAD</b> <b>FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV///2021</b>	

## 6.2 Resultados Parciales.

Se realizó una búsqueda general en tres principales bases de datos biológicos con acceso público, estas fueron Repbase [15], repeatDB [16] y PGSB [17], de las cuales se extrajeron aproximadamente cien mil elementos retrotransponibles pertenecientes a los genomas de diversas especies de plantas. Además, debido a que los elementos retrotransponibles para algunas especies no se encontraban disponibles, EDTA [18] y LTR\_STRUCTURE [18] fueron utilizados para extraer las secuencias de LTR-RT de estas especies. Para la clasificación de los elementos los filtros de InpactorDB [19] fueron utilizados, en donde las secuencias que lograron pasar la totalidad de los filtros se les consideraba como elementos LTR-RT completos y, los elementos que no lograron pasar la totalidad de los filtros fueron designados como elementos incompletos, en donde, los elementos podían ser filtrados debido a que poseían inserciones de otros elementos ya fueran retrotransponibles o transponibles y la longitud de los elementos era superior en comparación a las longitudes reportadas en la literatura; utilizando una tolerancia del 20% para la clasificación.

Con base a la anterior información de las bases de datos, se optó inicialmente en realizar un estudio de tuneo de hiper parámetros con una secuencia de 10000 datos, esto debido a que computacionalmente será más fácil de procesar la información y poder obtener resultados de forma más rápido, para luego probar con la base de datos completa, y tener definido y cumplir con uno de los objetivos que fue diseñar una base de datos estable para la realización de las diversas investigaciones y metodologías planteadas para la curación de secuencias ADN de elementos Retro transponibles.

En la realización de la investigación cabe recalcar que existen 2 tipos de clases transponibles, tal como se explicó en la introducción, por tal motivo se aclara que la investigación se enfocó principalmente en los elementos Retro transponibles, en donde dependiendo del comportamiento de los elementos se van a clasificar de diferentes maneras, de los cuales los más encontrados en las cadenas de ADN son el LTR, el cual fue este el utilizado y las grandes superfamilias utilizadas para nuestra investigación fueron la copia y la Gypsy. En la imagen 1 se evidencia un poco más a detalle las clasificación de los elementos transponibles en las secuencias de ADN en plantas.

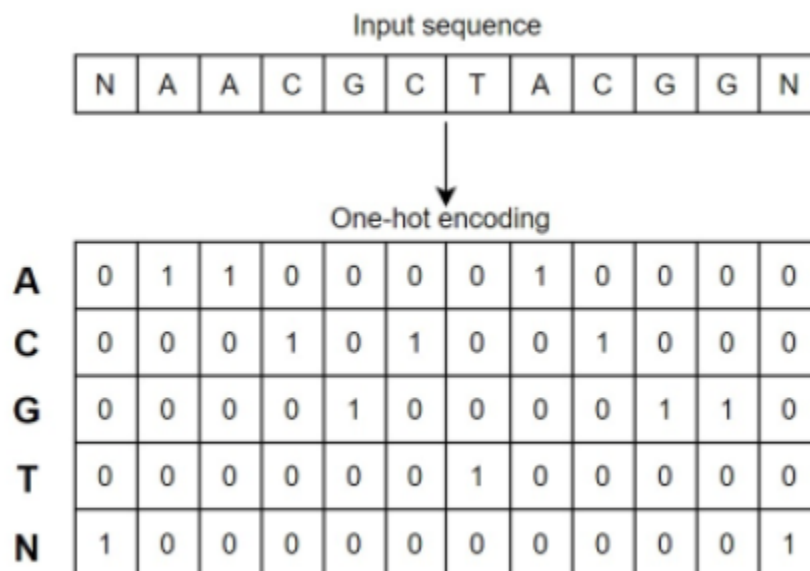


**Figura 1. Clasificación de elementos Transponibles.[2]**

**Pre-Proceso:**

Inicialmente se realizó un ajuste a los datos con el objetivo de poder utilizar CNNs, para este tomamos las bibliotecas o cadenas de ADN, y luego por medio de una matriz de 5xN; debido a que los primeros 4 nucleótidos [A, C, G, T], son los que van a brindar información a nuestra investigación, por lo que se recorre toda la secuencia y una vez coincida el nucleótido con el nucleótido de la matriz se le da un 1, a este valor y a los demás cero, y en caso de que esté en la cadena un elemento diferente a los 4 propuestos

hiran en el último de la fila llamado N. En la figura 2 se evidencia el Método de one encoding, en una pequeña secuencia de ADN, evidenciando la formación de la matriz Nx5.





**Figura 2. Codificación de la secuencia de ADN.[2]**

### **Tuneo de Hiperparametros:**

Los hiper parámetros han permitido ha muchos investigadores actualmente mejorar el acierto de precisión de los modelos diseñados por los mismos, aunque su proceso es algo lento y tedioso ha demostrado resultados muy significativos e importantes para diferentes investigadores, es por este motivo que se realizaron en esta investigación diferentes pasos, para llevar a cabo dichos tuneos de hiperparametros.

Como paso inicial, se realizó una búsqueda intensiva de los optimizadores, del modelo para lo cual en la tabla 1, se observa los optimizadores proporcionados por Keras, y su respectivo valor, utilizando la métrica F1-score.

Optimizadores	Loss	F1 score
Adam	0,3652	0,8527
Adadelta	0,83	0,4856
Adagrad	0,8344	0,624
Adamax	0,3729	0,8492
Ftrl	0,6932	0,4722

		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

Nadam	0,343	0,8601
RMSprop	0,3812	0,8492
SGD	0,8287	0,6359

**Tabla 1. Búsqueda de los mejores Optimizadores.**


Una vez sabiendo elegido el mejor optimizador, el siguiente paso, fue buscar las capas densamente conectadas, para lo cual en la tabla 2 se ilustra las capas densas probadas y su respectiva métrica, en este caso Accuracy, para las 3 carpetas[Train, Validation and Test].

En donde las capas densas que mayor porcentaje fueron las de 32 y 16.

Train					Test	
Fully conected	acc	loss	val acc	val loss	acc	loss
32	0,89	0,3	0,88	0,36	0,8875	0,3588
32,16	0,92	0,24	0,88	0,35	0,884	0,37
32,16,12	0,921	0,26	0,874	0,4	0,873	0,41
32,16,12,9,6	0,913	0,288	0,862	0,431	0,865	0,42
64	0,936	0,19	0,87	0,39	0,866	0,402
64,32	0,922	0,27	0,863	0,43	0,875	0,436
64,32,25	0,923	0,243	0,882	0,34	0,886	0,342
128	0,918	0,27	0,87	0,37	0,86	0,4
128,64	0,915	0,28	0,86	0,4	0,875	0,39
512	0,912	0,34	0,87	0,43	0,87	0,44
512, 256	0,92	0,3	0,88	0,42	0,88	0,42
512,256,204	0,9	0,34	0,885	0,405	0,883	0,41
512,256,204,153,102	0,903	0,37	0,86	0,46	0,867	0,46
1024	0,903	0,405	0,88	0,46	0,885	0,452
1024,512	0,912	0,39	0,877	0,48	0,88	0,49
1024,512,204	0,896	0,45	0,86	0,52	0,865	0,52

**Tabla 2. Búsqueda de las capas densamente conectadas.**

Estas series de pasos nos han proporcionado una vista previa de la mejora que puede tener el modelo de Deep Learning al modificar diferentes hiperparametros, aunque es algo costoso computacionalmente, y tedioso, los resultados proporcionados podrán brindar un

		CÓDIGO: 76621650450
		VERSIÓN: 4
		FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV///2021

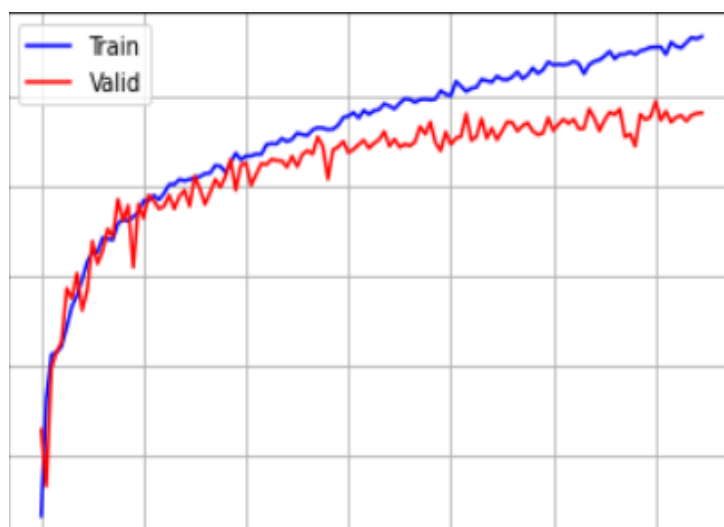
paso fundamental para la curación de secuencias de elementos retro transponibles en plantas .

En la tabla 3, se evidencia el porcentaje obtenido por el momento por el modelo, siendo del 90%. Aunque es un porcentaje bueno nuestro objetivo es superar aún más este porcentaje y probar los diferentes hiperparametros faltantes para cumplir con nuestro objetivo y problemática de investigación.


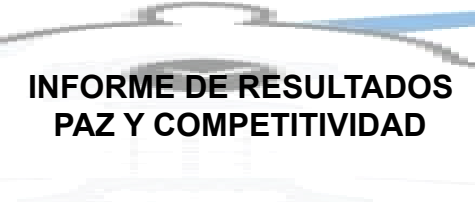
Train					Test
	acc	loss	val_acc	val_loss	acc
	0,934	0,233	0,8945	0,34	0,9

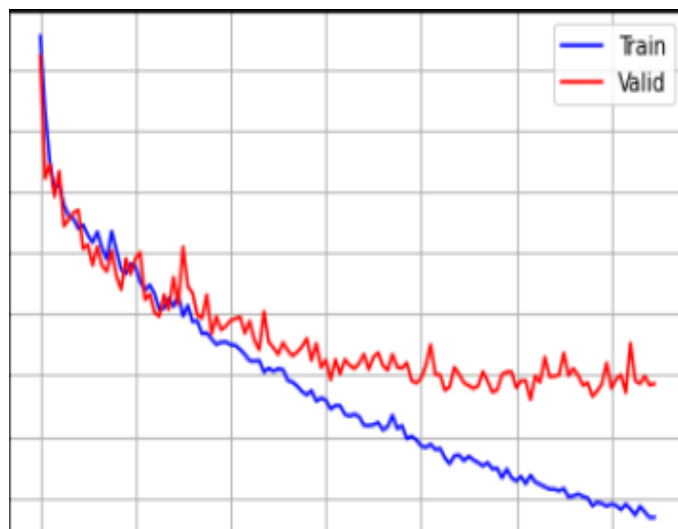
**Tabla 3. Resultado mejor modelo obtenido**

Y a continuación observamos en las figuras 3 y 4. Sus gráficas de Entrenamiento y pérdida del respectivo modelo. Donde se evidencia que el modelo está aprendiendo correctamente, sin presentar Overfitting durante su entrenamiento, pero se empieza a estancar en un determinado porcentaje, en donde se pretende mejorar este caso, y proporcionar una mejora significativa al modelo durante su entrenamiento.



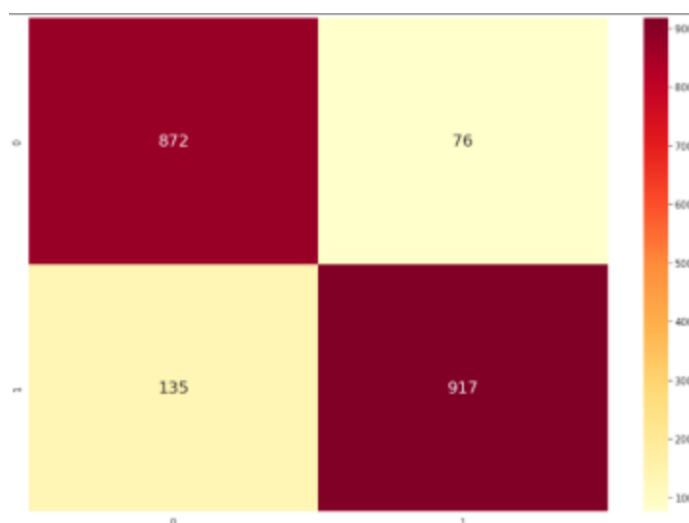
**Figura 3. Gráfica de entrenamiento del mejor modelo obtenido.**



		CÓDIGO: 76621650450
		VERSIÓN: 4
		FECHA ELABORACIÓN DEL DOCUMENTO: 30/NOV//2021



**Figura 4. Gráfica de la pérdida o error del mejor modelo obtenido.**

Al igual que sus métricas utilizadas para evaluar el modelo. En la figura 5 se evidencia su matriz de confusión observando la incidencia de aciertos positivos que tiene una clase respecto la otra, a simple vista el modelo está clasificando correctamente, pero sigue aún presentando errores respecto a algunas secuencias de ADN.



		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

**Figura 5. Matriz de confusión del modelo.**

Por último en la tabla 4, se evidencia las métricas utilizadas y su resultado, en donde ambas métricas conllevaron a tener resultados similares, e igual al 0.903. Dando a conocer que este porcentaje es el final y el correspondiente a nuestro modelo entrenado.

<b>Métrica</b>	<b>Resultado</b>
Accuracy	0.903
F1- Score	0.903
Recall	0.903
Precisión	0.9037

**Table 4. Métricas Obtenidas del mejor modelo.**

## 7. CONCLUSIONES

Se concluye que la arquitectura de da Cruz puede ser utilizada para clasificar los LTR retrotransposones, además se implementa y se observa que su rendimiento puede ser mejorado por medio de la regularización de bias y kernel con L1 y L2 en la capas de la arquitectura resultan ser un buen método para disminuir el sobre ajuste del modelo, además de la disminución del Learning Rate a medida que pasa las épocas, ya que mostró mayor eficiencia que métodos convencionales como el Checkpoint.

La profundidad de clasificación de los LTR retrotransposones afecta el rendimiento de la arquitectura puesto que es un problema más general y por ende más complicado de clasificar. Además el tuneo de hiperparametros como la expansión de la base de datos han brindado una mejora al modelo siendo de vital importancia su estudio, y aplicaciones práctica para la curación de elementos trasponibles, brindando un soporte, optimización en cuanto a tiempo y nuevas técnicas y mejoras que se le puede proporcionar a un modelo de CNNs.

## 8. RECOMENDACIONES

Una recomendación para la comunidad UAM es el uso de nuestras tecnologías actuales y herramientas computacionales a nuestro alcance, para el estudio de adquisición



		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

de nuevas técnicas y enseñanzas para ampliar los conocimientos, para la toma de decisiones de nuestro día a día, tomando como problemática diversos problemas actualmente como lo son, la enseñanza y las diversas herramientas computacionales para la realización de las investigaciones que lo requieran, como lo es esta misma, con el objetivo de mejorar nuestra sociedad, y nuestro vivir. Unificándolas cada día con los estudios de otros investigadores con el propósito de solucionarlo.

## 9. ANEXOS

Experimentos CNN\_3

## 10. BIBLIOGRAFÍA

[1] Genotipia.com. 2020. Elementos Transponibles: Los “Saltimbanquis” Del Genoma -. [online] Available: <https://genotipia.com/elementos-transponibles/>. [Accessed 28 October 2021].


[2] da Cruz, Murilo Horacio Pereira et al. 2021. “TERL: Classification of Transposable Elements by Convolutional Neural Networks.” *Briefings in bioinformatics* 22(3).

[3] Orozco-Arias, Simon et al. 2022. “Deep Neural Network to Curate LTR Retrotransposon Libraries from Plant Genomes.” In *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 85–94.

[4] Orozco-Arias, Simon et al. 2020. “Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements.” *Processes*, doi:10.3390/pr8060638.

[5] Orozco-Arias, Simon et al. 2021. “K-Mer-Based Machine Learning Method to Classify LTR-Retrotransposons in Plant Genomes.” *PeerJ*.

[6] Orozco-Arias, Simon, Gustavo Isaza, and Romain Guyot. 2019. “Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning.” *International Journal of Molecular Sciences*. Doi.: 10.3390/ijms20153837

		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

[7] Orozco-Arias, Simon, Gustavo Isaza, Romain Guyot, and Reinel Tabares-Soto. 2019. “A Systematic Review of the Application of Machine Learning in the Detection and Classification of Transposable Elements.” *PeerJ* .

[8] Raúl Castanera Andrés, 2017, “Transposable elements in basidiomycete fungi”, *Dialnet*, Universidad Pública de Navarra , España.

[9] Enrique Navas Pérez, 2018, “Una nueva domesticación molecular en el origen de los euterios,” Universidad de Barcelona, España.

[10] Carlos A. M. E., Lizeth V. R. M., Jhenifer F. S. 2011, “Tecnologías bioinformáticas para el análisis de secuencias de ADN,” *Bioinformatics Technologies for the Analysis of DNA sequences*.

[11] Khan Academy, “secuenciación del ADN” .[online] Available: <https://es.khanacademy.org/science/ap-biology/gene-expression-and-regulation/biotechnology/a/dna-sequencing>. [Accessed 1 December 2021].



[12] Bousios A., Minga E., Kalitsou N., Pantermali M., Tsaballa A., Darzentas N. 2012. “MASiVEDb: the Sirevirus Plant Retrotransposon Database”. *BMC Genomics*, Doi: <https://doi.org/10.1186/1471-2164-13-158>

[13] Negi P., Rai A. N., Suprasanna P. 2016. “Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response”. *Front. Plant Sci*, Doi: <https://doi.org/10.3389/fpls.2016.01448>

[14 ]Ping Shung, K., 2018. Accuracy, precision, recall or f1? Towards data science

[15] Bao W, Kojima KK, Kohany O. 2015. “Repbase Update, a database of repetitive elements in eukaryotic genomes,” *Mob DNA*, 6:4–9. Doi: <https://doi.org/10.1186/s13100-015-0041-9>

[16] Amselem J, Cornut G, Choisine N, et al. 2019. “RepetDB: A unified resource for transposable element references”. *Mob DNA*. 4–11. Doi: <https://doi.org/10.1186/s13100-019-0150-y>

		<b>CÓDIGO:</b> 76621650450
		<b>VERSIÓN:</b> 4
		<b>FECHA ELABORACIÓN DEL DOCUMENTO:</b> 30/NOV//2021

[17] Spannagl M, Nussbaumer T, Bader KC, et al. 2016. “PGSB plantsDB: Updates to the database framework for comparative plant genome research”. *Nucleic Acids Res* 44:D1141–D1147. Doi: <https://doi.org/10.1093/nar/gkv1130>

[18] Ou, S., Su, W., Liao, Y. *et al.* 2019. “Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline”. *Genome Biol* 20, 275. Doi: <https://doi.org/10.1186/s13059-019-1905-y>

[19] McCarthy EM, McDonald JF. 2003. “LTR STRUC: A novel search and identification program for LTR retrotransposons”. *Bioinformatics*. 19:362–367. Doi: : <https://doi.org/10.1093/bioinformatics/btf878>

[20] Orozco-Arias S, Jaimes PA, Candamil MS, et al. 2021. “InpactorDB : A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning”. *MDPI Genes*. 12:17. Doi: <https://doi.org/https://doi.org/10.3390/genes12020190>