# Proyecto:
# Clasificación de audio.

Universidad Nacional Autónoma de México

Licenciatura en Ciencia de datos

- Martiñón Luna Jonathan José
- Tapia López José de Jesús

# Objetivos

## Sensar

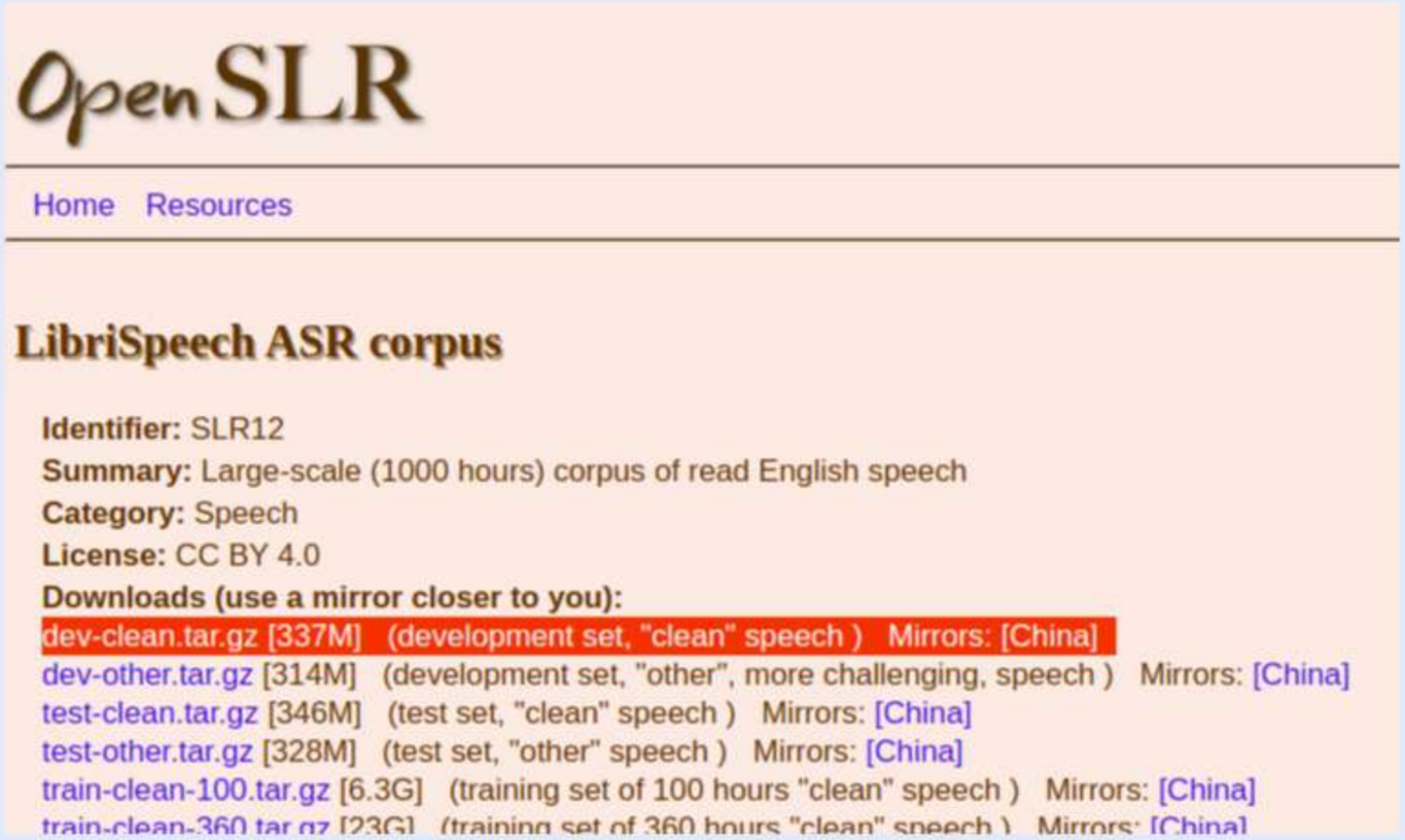
A través de un audio, poder identificar la cantidad de personas que se encuentren en ciertas áreas.

## Simplificar

Dejar de lado el uso de sensores o contadores manuales y dejar el trabajo a un simple micrófono. Queremos resolver este problema usando herramientas de Aprendizaje Profundo.

1

Se descargó el conjunto de audios individuales de: https://www.openslr.org/12



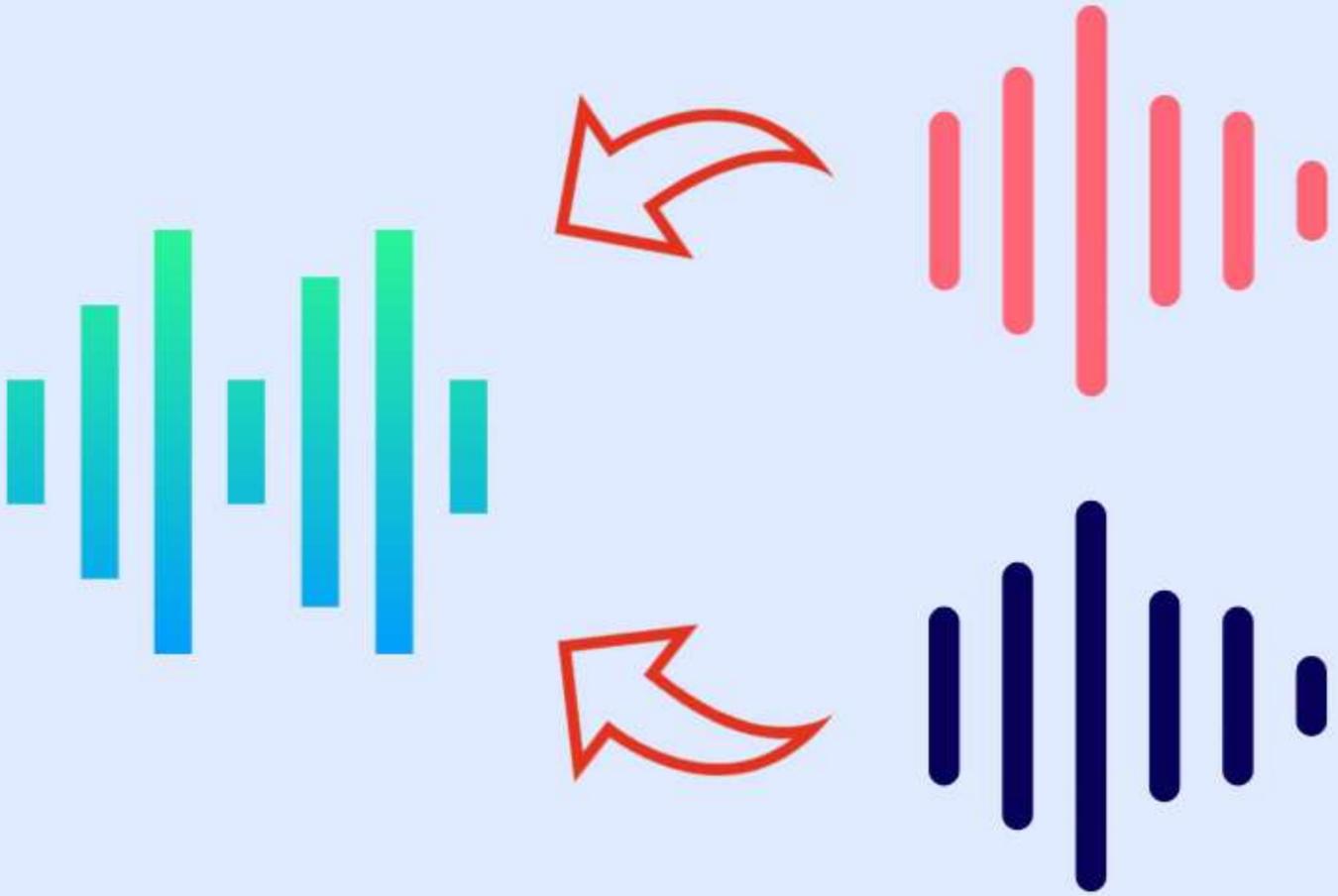
OpenSLR

Home Resources

## LibriSpeech ASR corpus

**Identifier:** SLR12
**Summary:** Large-scale (1000 hours) corpus of read English speech
**Category:** Speech
**License:** CC BY 4.0
**Downloads (use a mirror closer to you):**
dev-clean.tar.gz [337M] (development set, "clean" speech ) Mirrors: [China]
dev-other.tar.gz [314M] (development set, "other", more challenging, speech ) Mirrors: [China]
test-clean.tar.gz [346M] (test set, "clean" speech ) Mirrors: [China]
test-other.tar.gz [328M] (test set, "other" speech ) Mirrors: [China]
train-clean-100.tar.gz [6.3G] (training set of 100 hours "clean" speech ) Mirrors: [China]
train-clean-360.tar.gz [23G] (training set of 360 hours "clean" speech ) Mirrors: [China]

# 2 Mezcla y etiquetado de audio Hombre - Mujer



# Datos

Separación en:

- Entrenamiento
- Validación
- Prueba

| | Entrenamiento | Validación | Prueba |
|---|---|---|---|
| Peso | 3.5 GB | 364 MB | 364 MB |
| Registros | 100,000 | 10,000 | 10,000 |

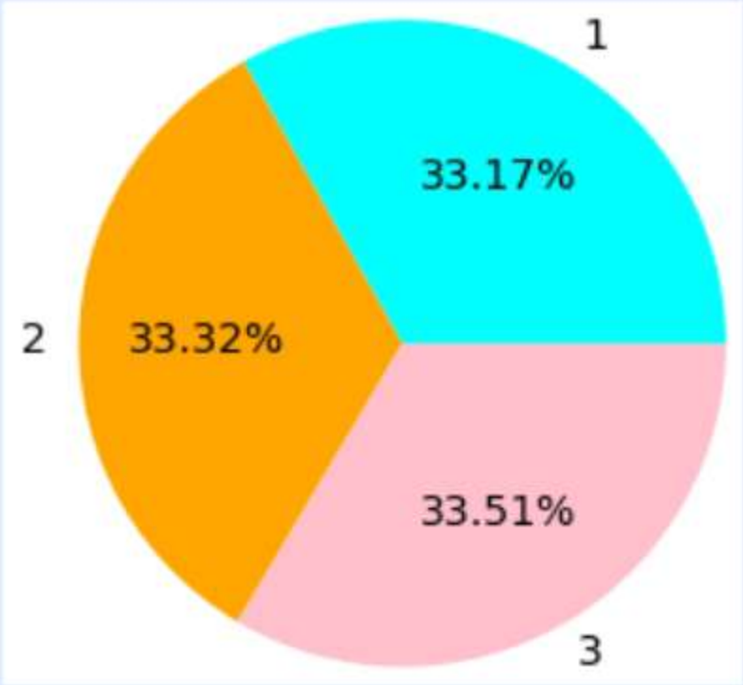
# Ejemplo

Datos

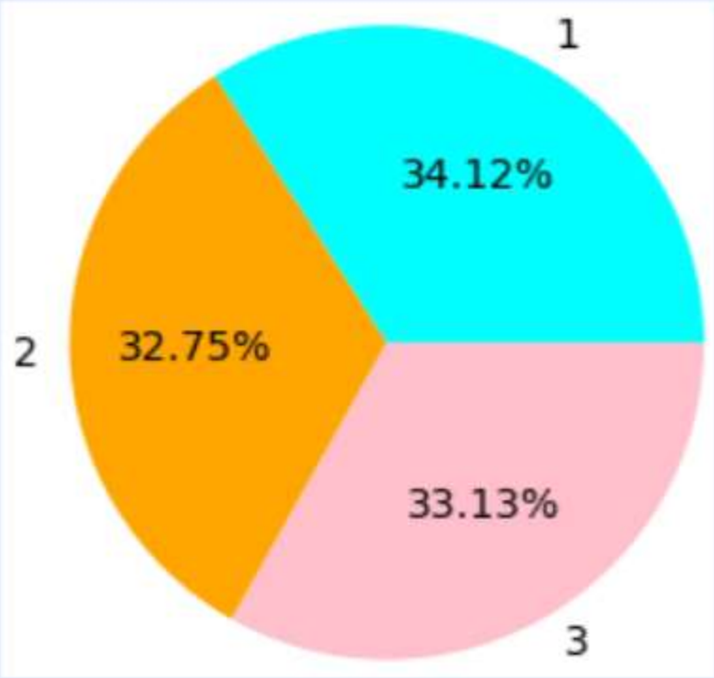
| | Waveform | Speakers | Speakers sex | F | M |
|---|---|---|---|---|---|
| Registro | Tensor[[0.0001,...,0.003]] | 3 | MMF | 1 | 2 |

# Distribución de clases

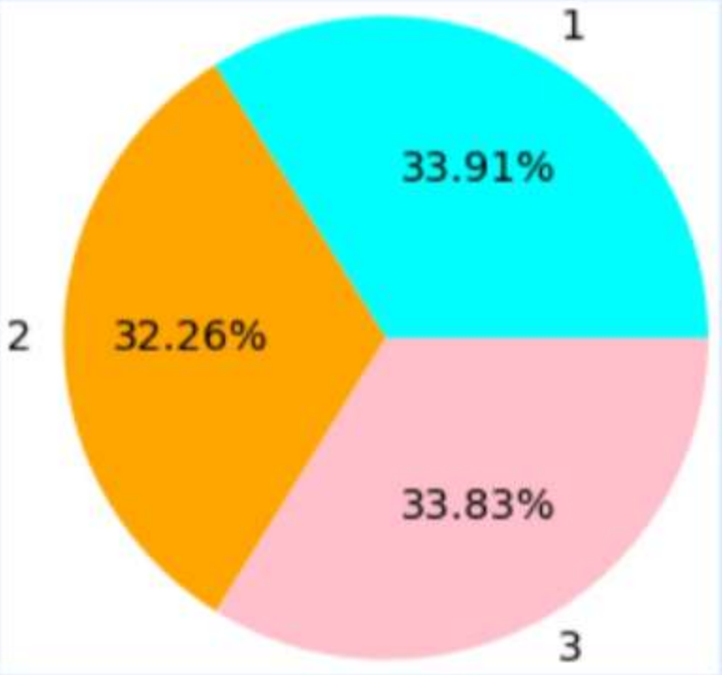
Datos




Entrenamiento

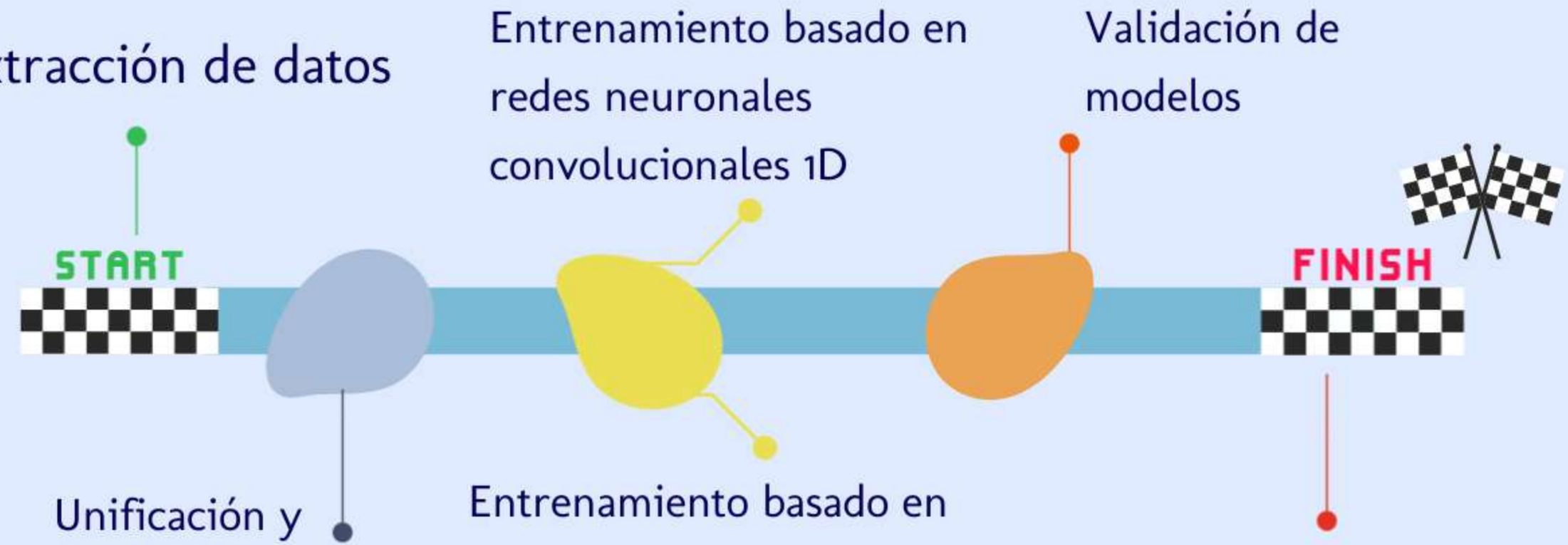



Validación



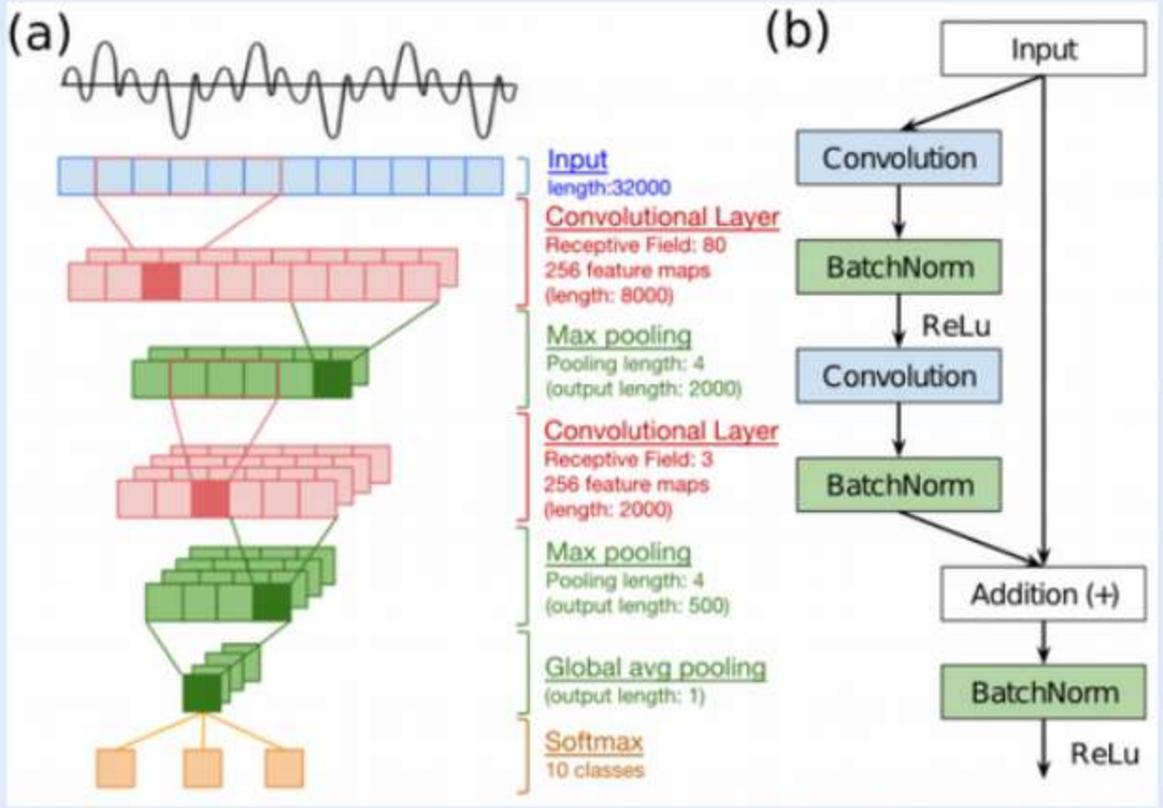

Prueba

# Forma de Trabajo



# RNC 1D

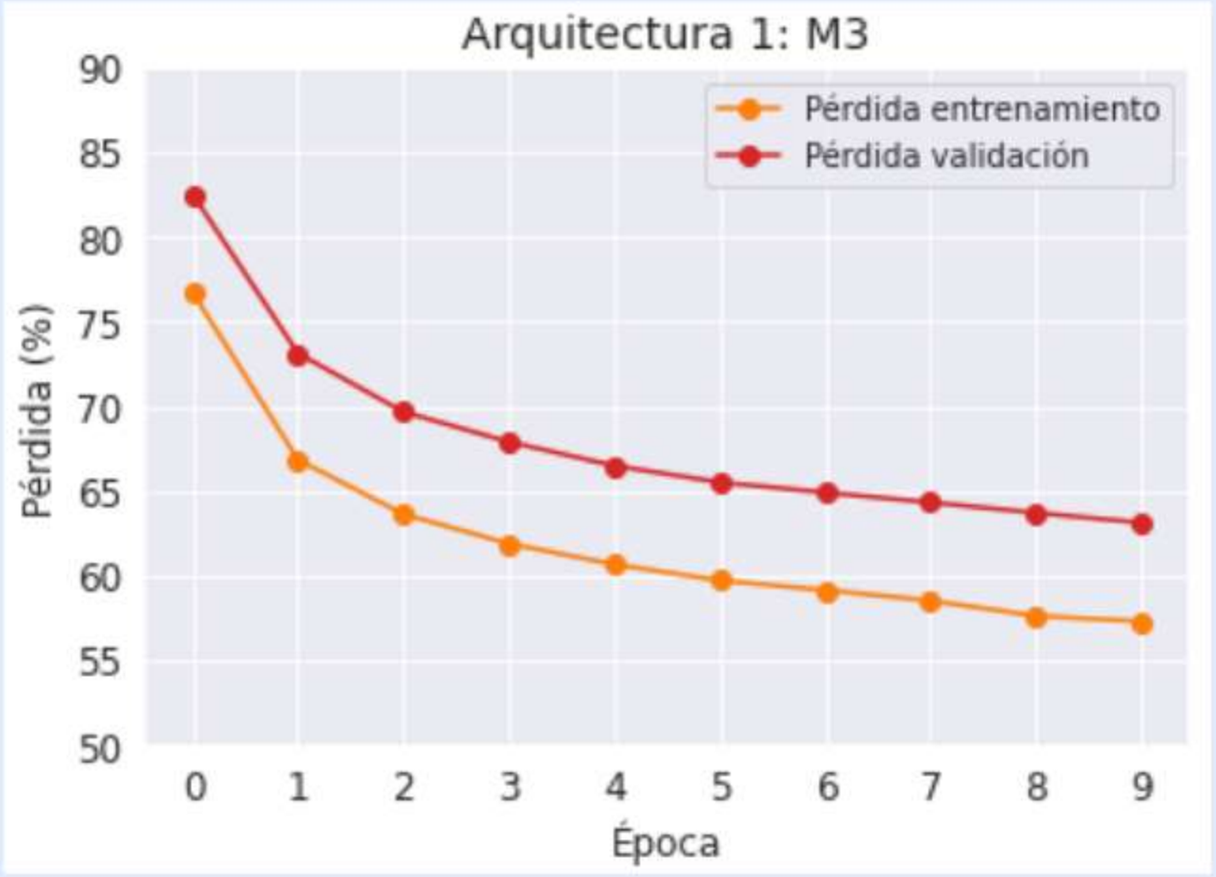
- stride en la primera capa: 16 (de 4 en el artículo)
- número de canales producidos por la primera convolución: 32 (de 128 para M3 y M5; de 64 en la M11 y M18 en el artículo)

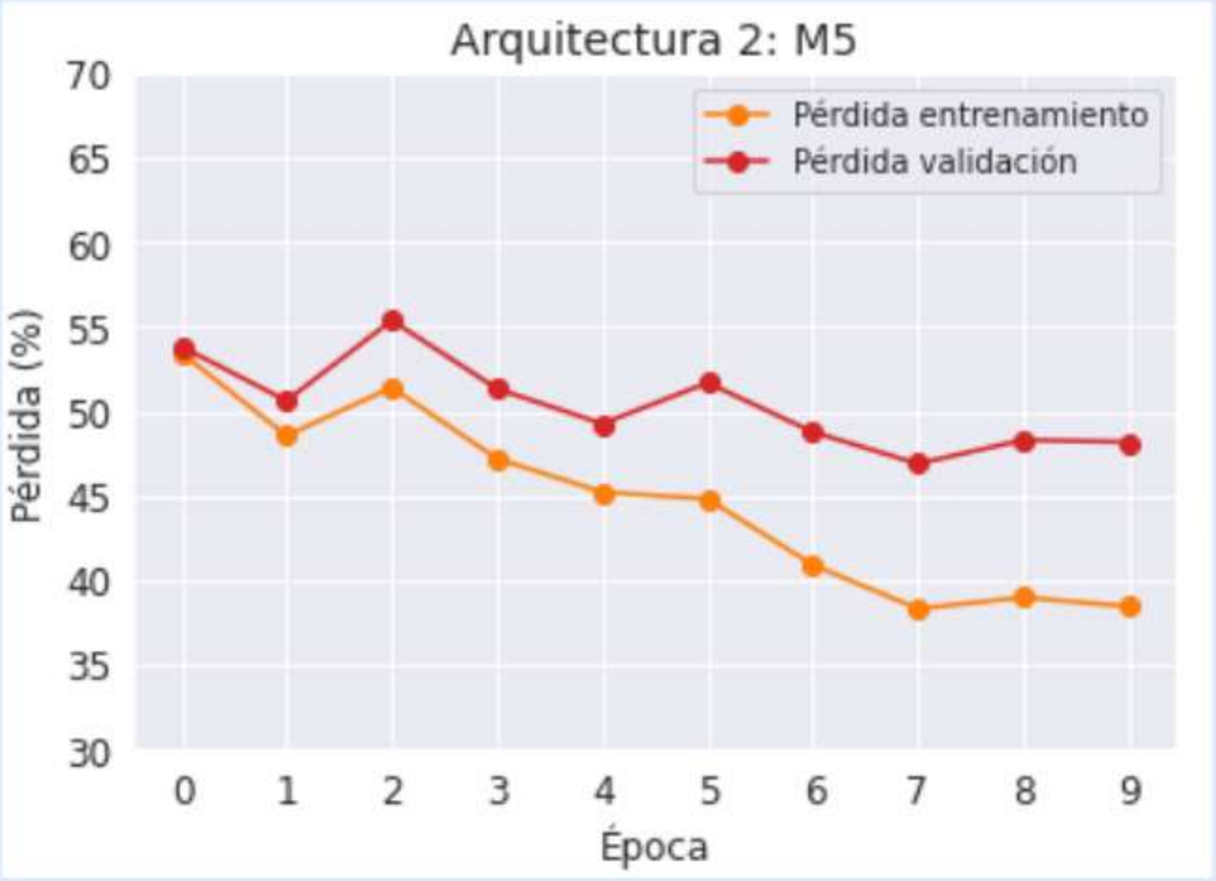



| M3 (0.2M) | M5 (0.5M) | M11 (1.8M) | M18 (3.7M) | M34-res (4M) |
|---|---|---|---|---|
| Input: 32000x1 time-domain waveform | | | | |
| [80/4, 256] | [80/4, 128] | [80/4, 64] | [80/4, 64] | [80/4, 48] |
| Maxpool: 4x1 (output: 2000 × n) | | | | |
| [3, 256] | [3, 128] | [3, 64] × 2 | [3, 64] × 4 | $\begin{bmatrix} 3, 48 \\ 3, 48 \end{bmatrix} \times 3$ |
| Maxpool: 4x1 (output: 500×n) | | | | |
| | [3, 256] | [3, 128] × 2 | [3, 128] × 4 | $\begin{bmatrix} 3, 96 \\ 3, 96 \end{bmatrix} \times 4$ |
| | Maxpool: 4x1 (output: 125 × n) | | | |
| | [3, 512] | [3, 256] × 3 | [3, 256] × 4 | $\begin{bmatrix} 3, 192 \\ 3, 192 \end{bmatrix} \times 6$ |
| | Maxpool: 4x1 (output: 32 × n) | | | |
| | | [3, 512] × 2 | [3, 512] × 4 | $\begin{bmatrix} 3, 384 \\ 3, 384 \end{bmatrix} \times 3$ |
| Global average pooling (output: 1 × n) | | | | |
| Softmax | | | | |

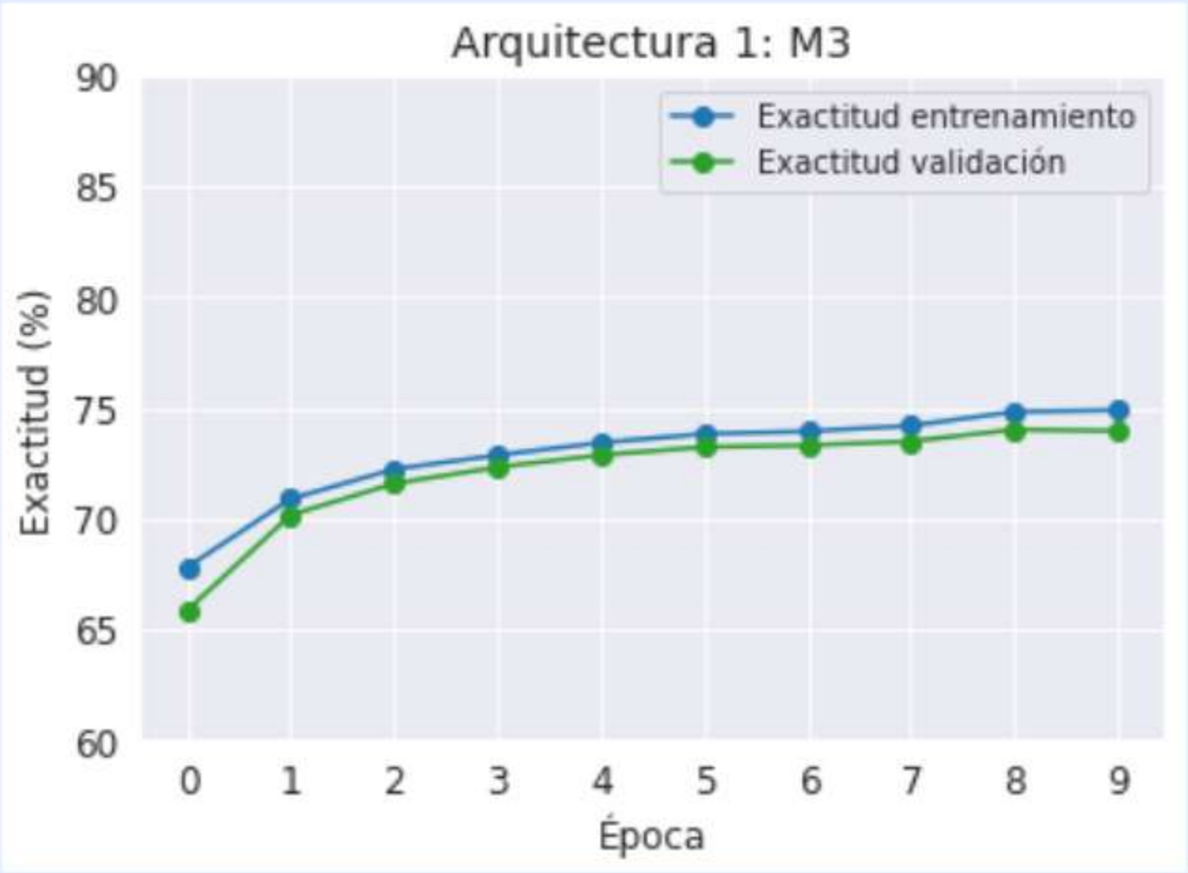
Dai, We. et al. *Very Deep Convolutional Neural Networks For Raw Waveforms*. 2016.

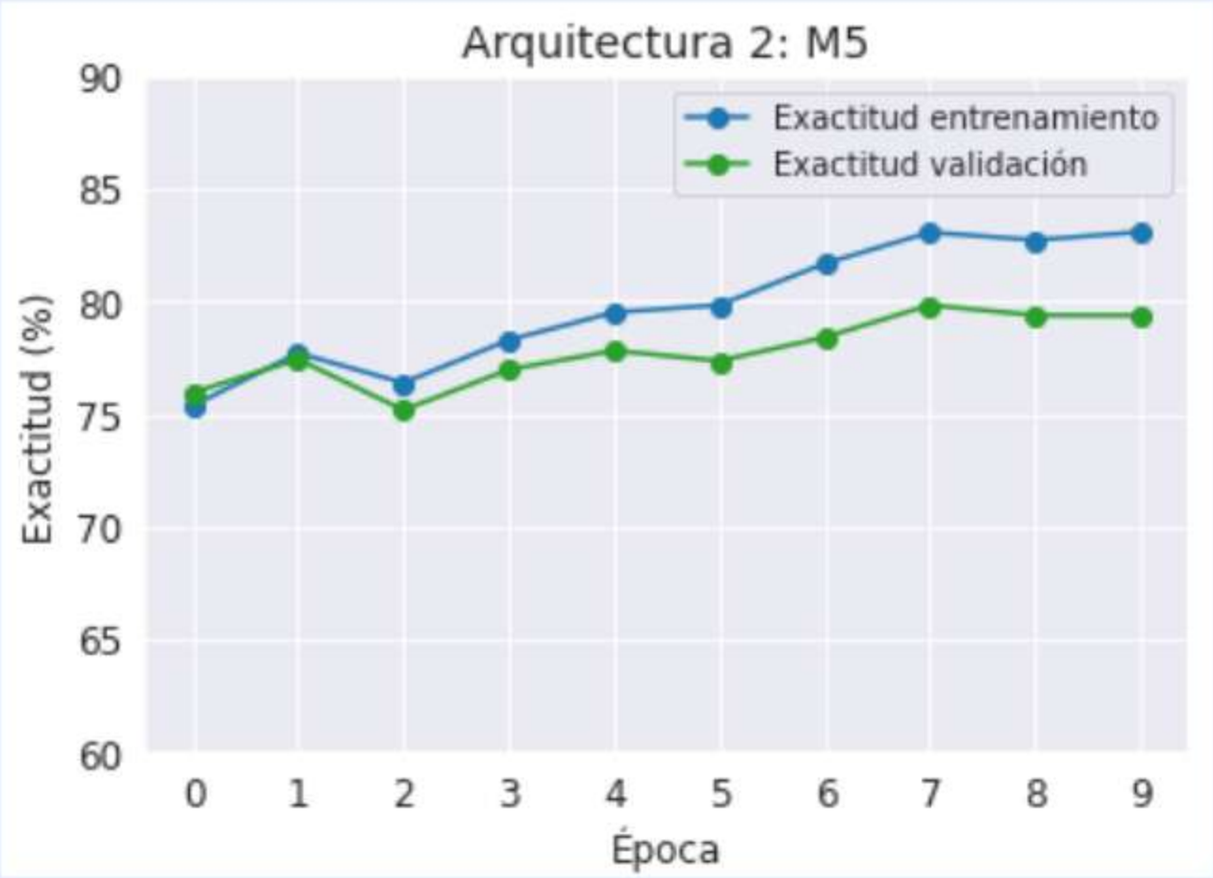
# M3y M5: Pérdida



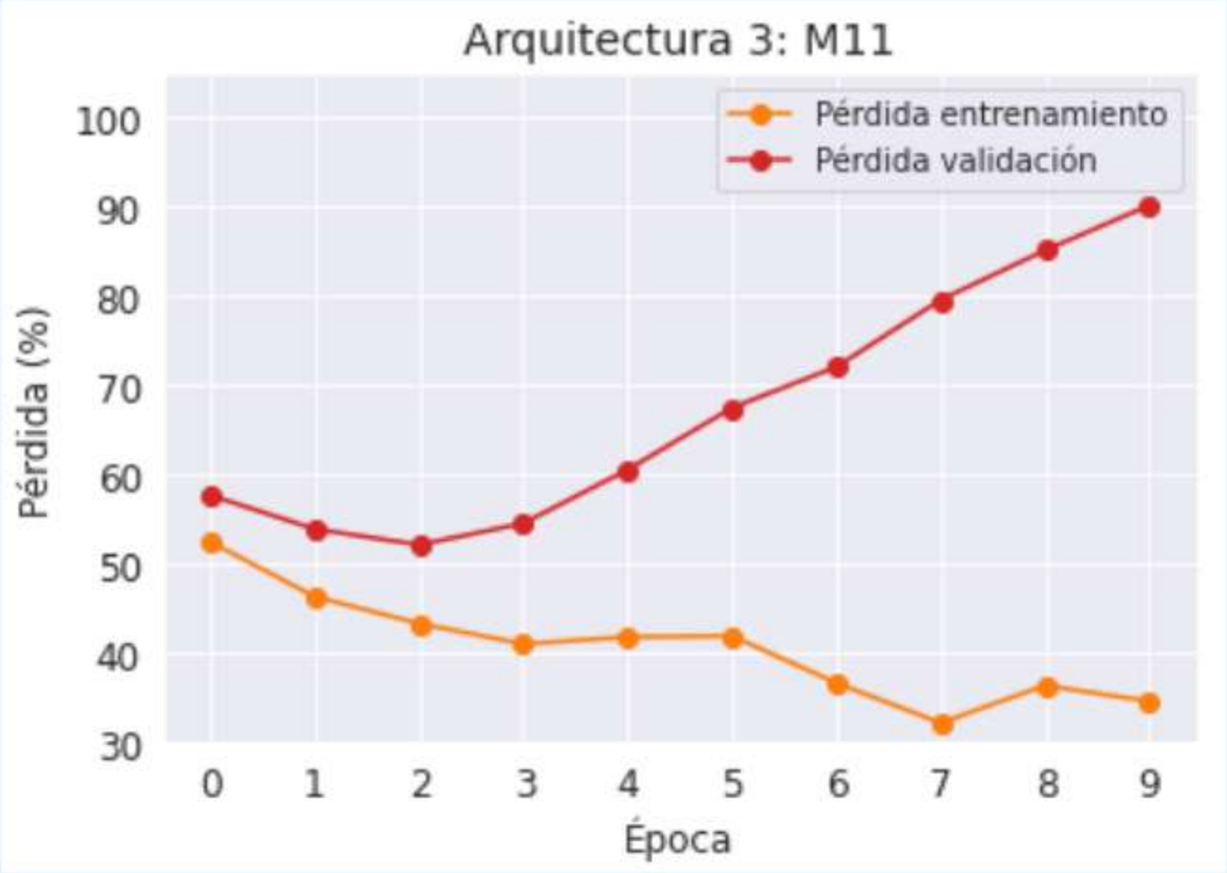



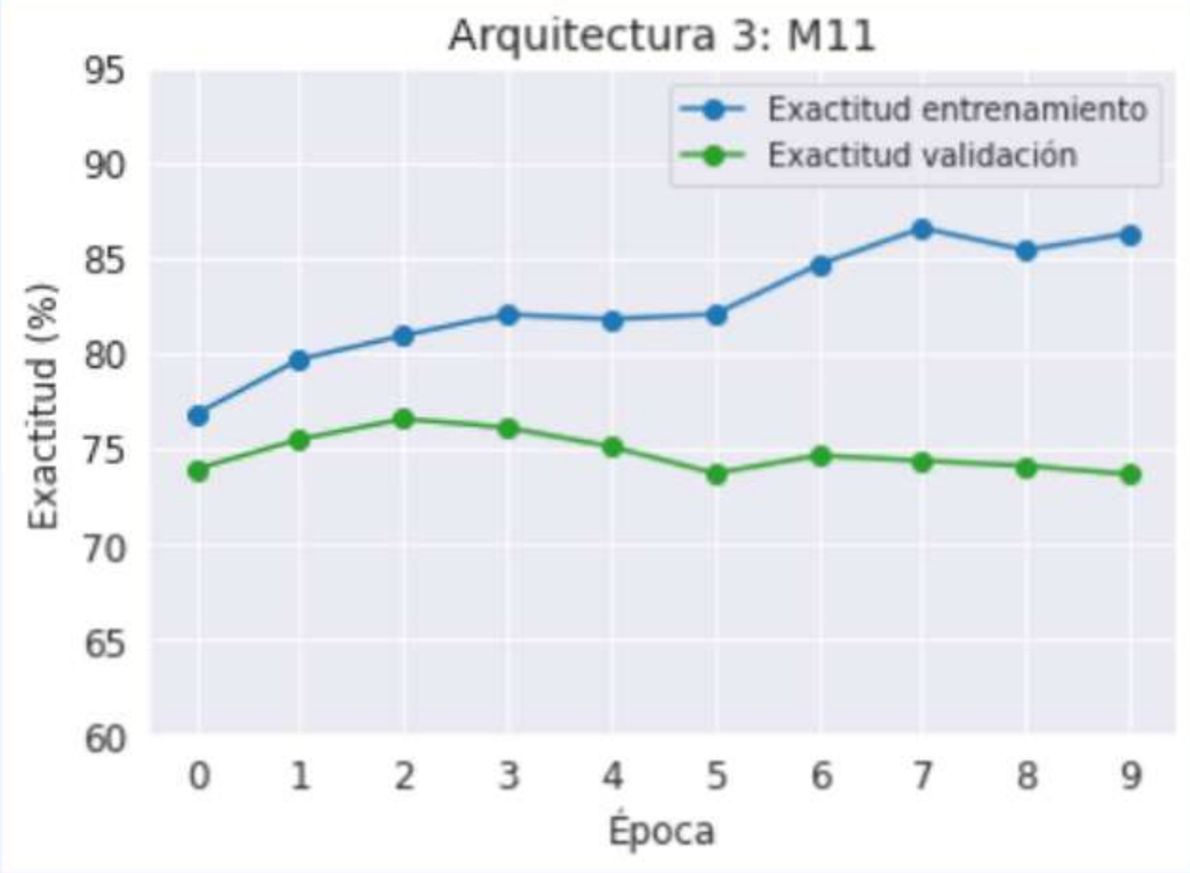
# M3y M5: Exactitud



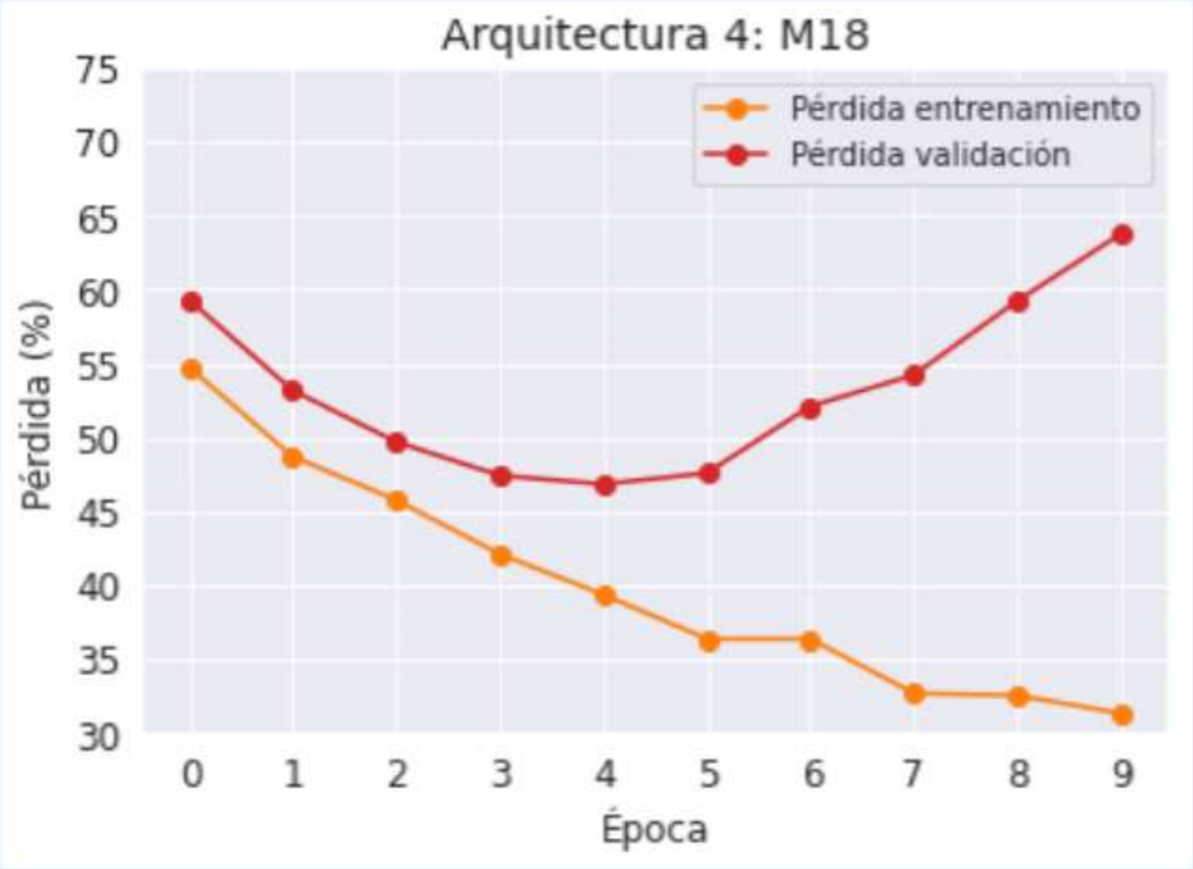



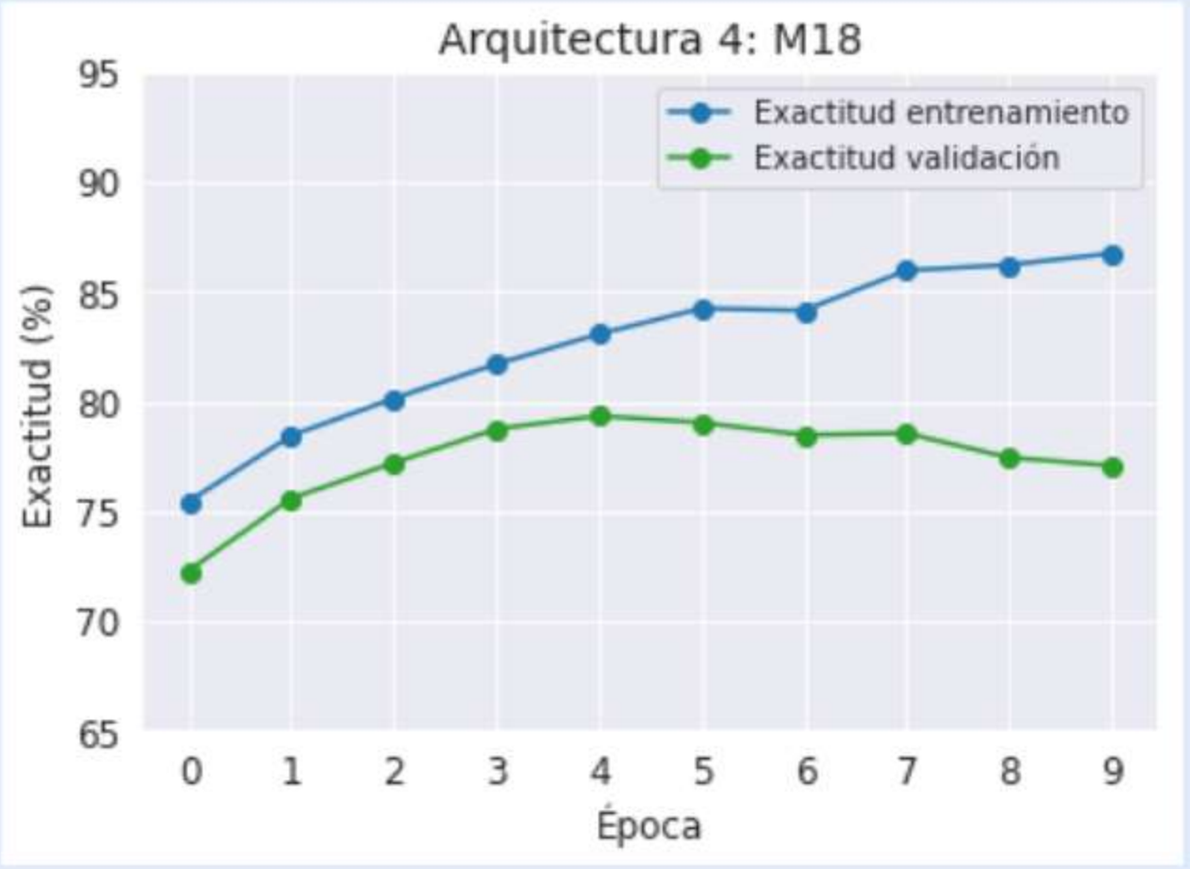
# M11


Arquitectura 3: M11
Pérdida entrenamiento
Pérdida validación
Pérdida (%)
100
90
80
70
60
50
40
30
0
1
2
3
4
5
6
7
8
9
Época


Arquitectura 3: M11
Exactitud entrenamiento
Exactitud validación
Exactitud (%)
95
90
85
80
75
70
65
60
0
1
2
3
4
5
6
7
8
9
Época

# M18


Arquitectura 4: M18
Pérdida entrenamiento
Pérdida validación
Pérdida (%)
Época
75
70
65
60
55
50
45
40
35
30
0
1
2
3
4
5
6
7
8
9


Arquitectura 4: M18
Exactitud entrenamiento
Exactitud validación
Exactitud (%)
Época
95
90
85
80
75
70
65
0
1
2
3
4
5
6
7
8
9

# Comparaciones de modelos Conv1D

| Arquitectura | # Paráms. | Tiempo |
|---|---|---|
| **M3** | 5,923 | 24 minutos |
| **M5** | 37,507 | 36 minutos |
| **M11** | 449,059 | 50 minutos |
| **M18** | 924,643 | 80 minutos |

# Mejor solución con Conv1D: M5

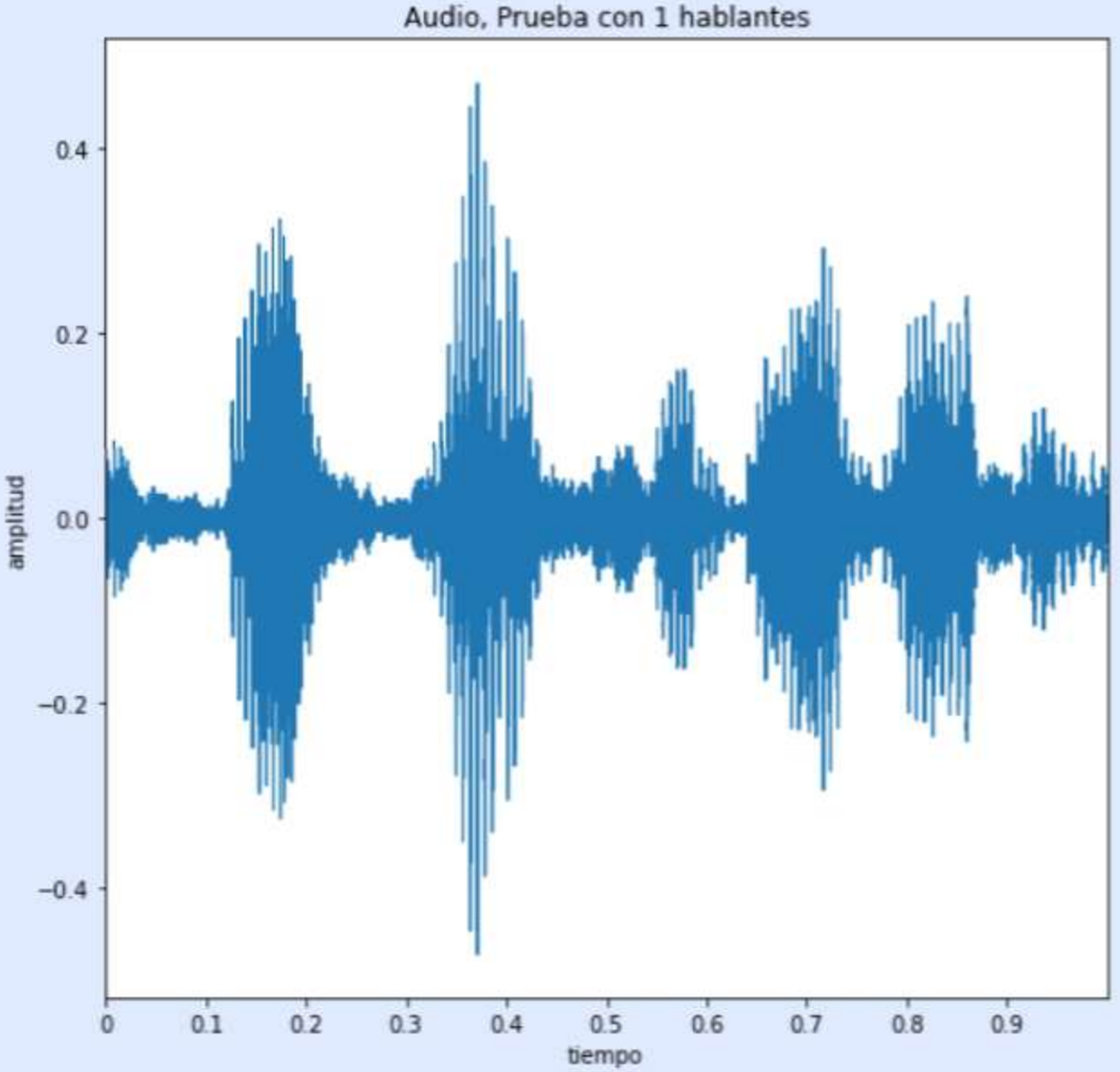
En la última época, es con la que tenemos la mayor exactitud en los datos de validación (79.37%) y una menor pérdida (48.17%) en estos mismos.

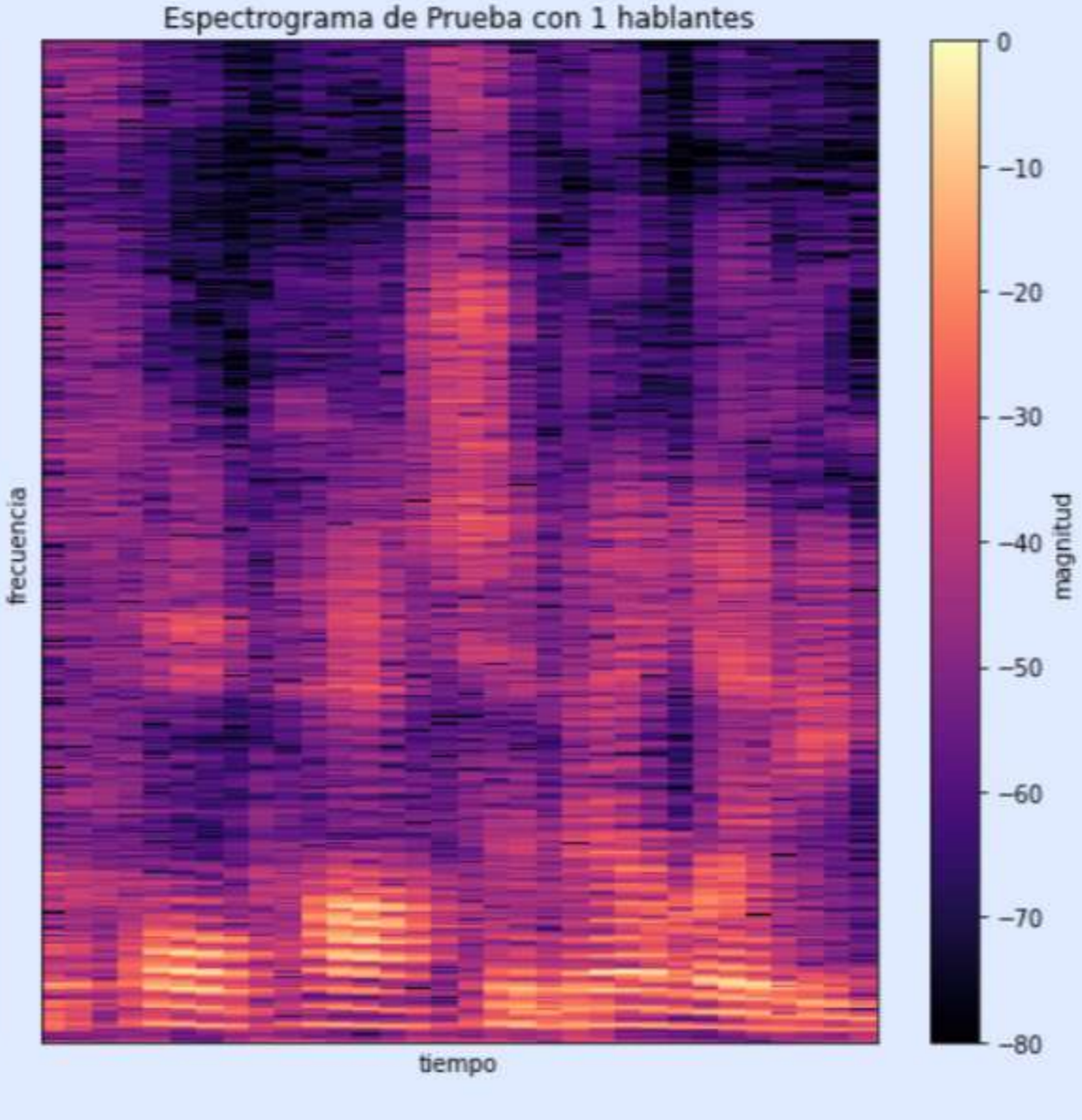
Consideramos que realmente no se sobreajusta.

Su número de parámetros es relativamente pequeño; su tiempo de ejecución es relativamente adecuado.

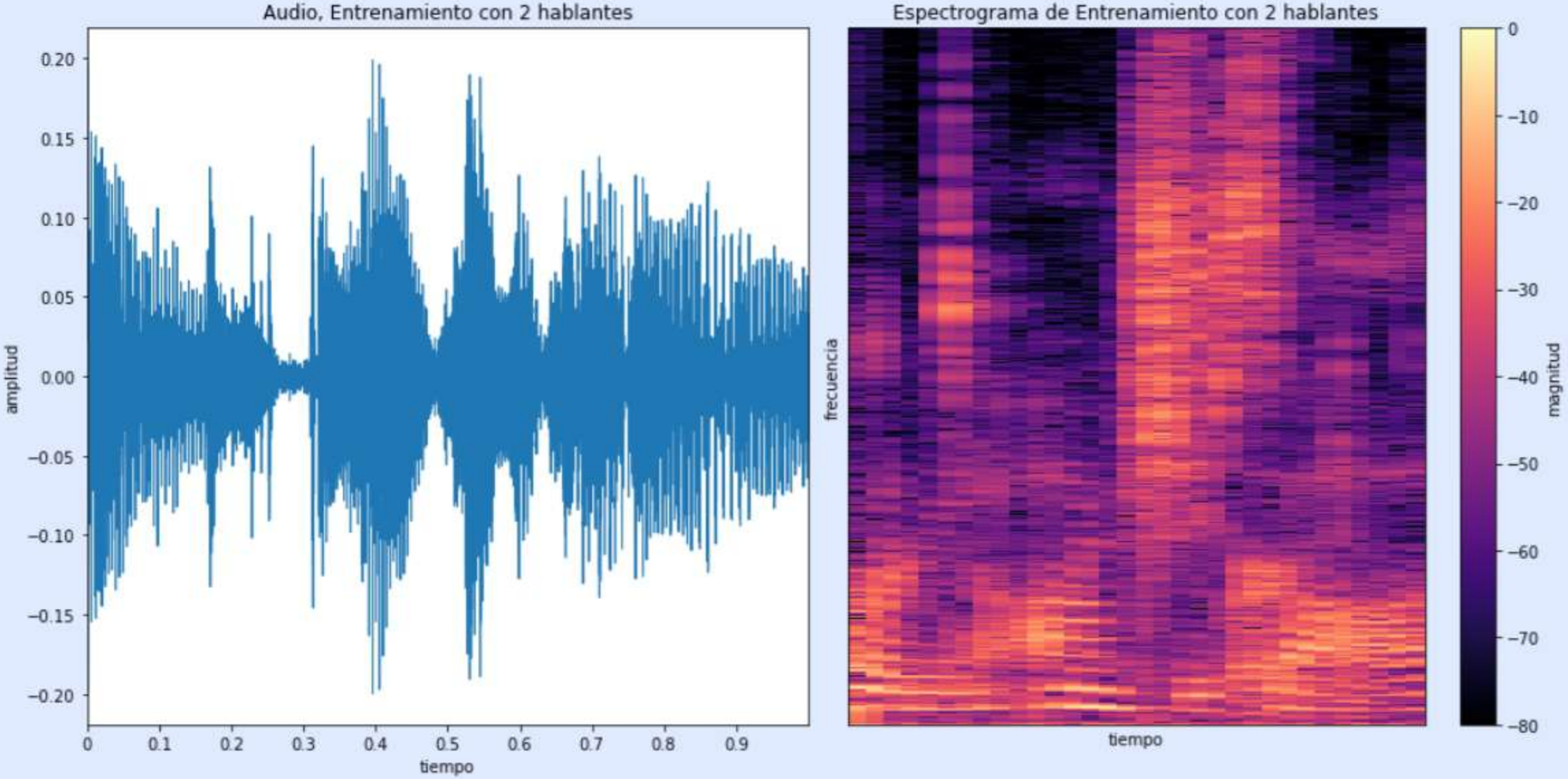
# Predicciones con M5

# Espectrogramas y Convolucionales

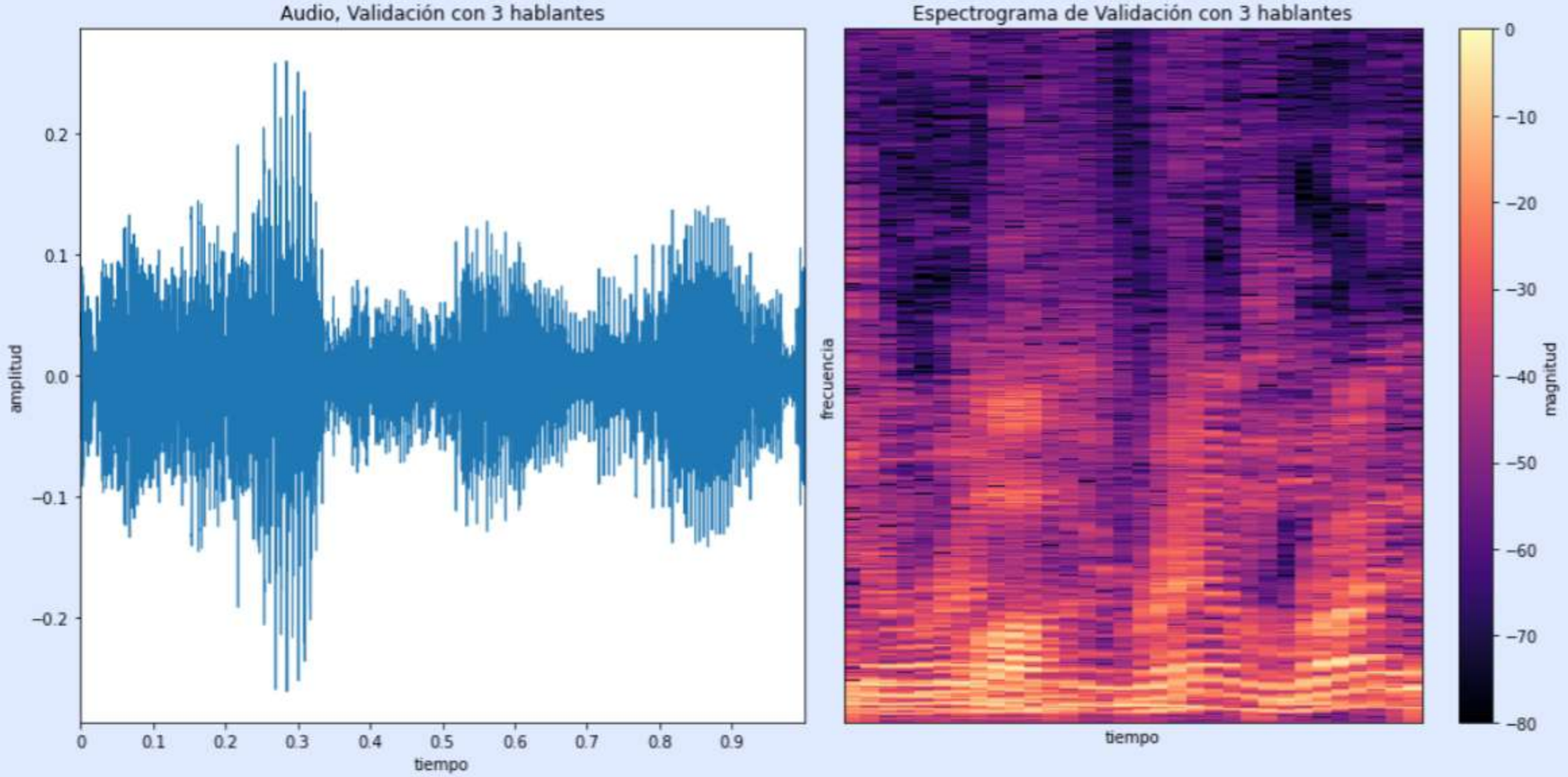


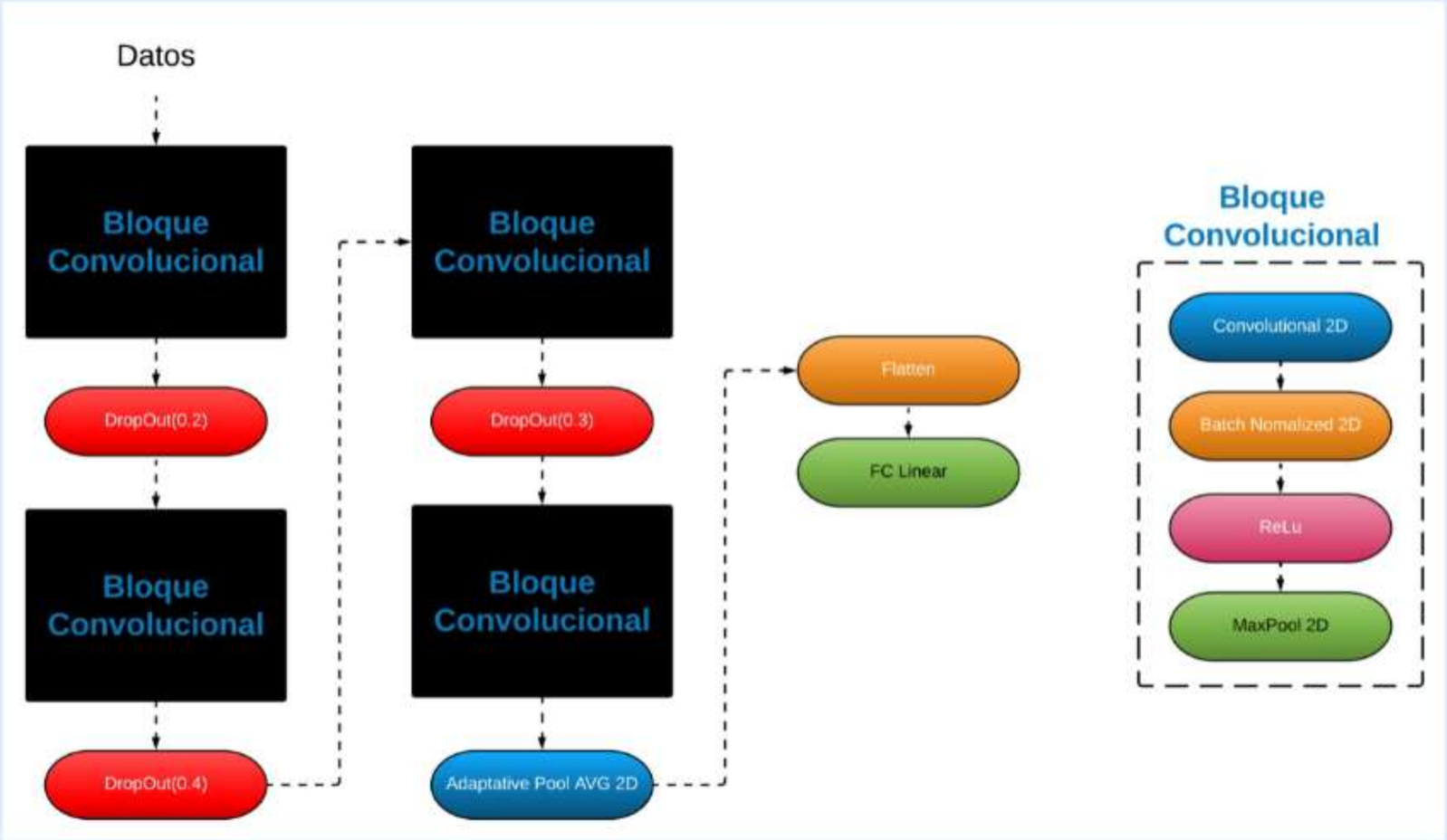
# Espectrogramas y Convolucionales



# Espectrogramas y Convolucionales

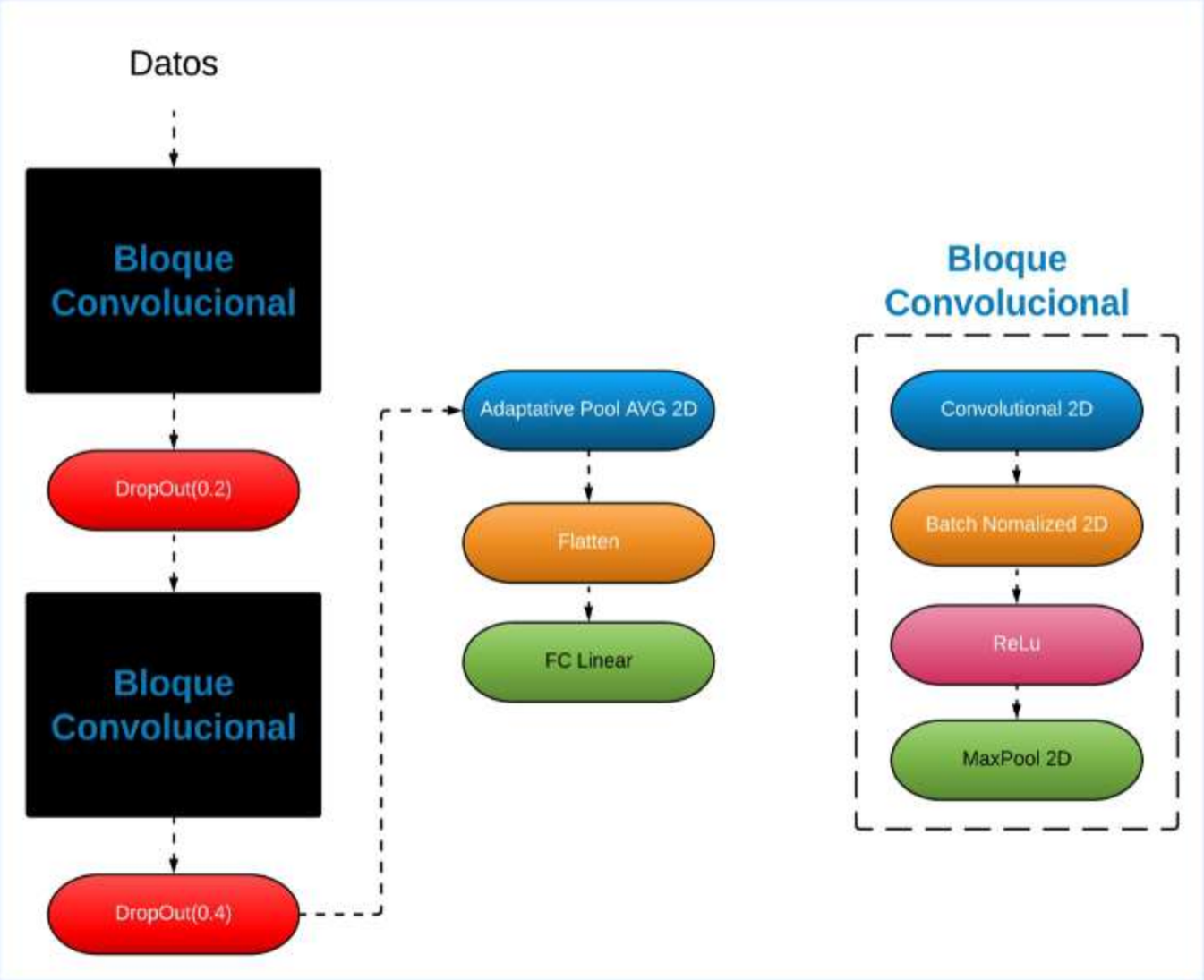


# Primer Intento (Arquitectura 2)



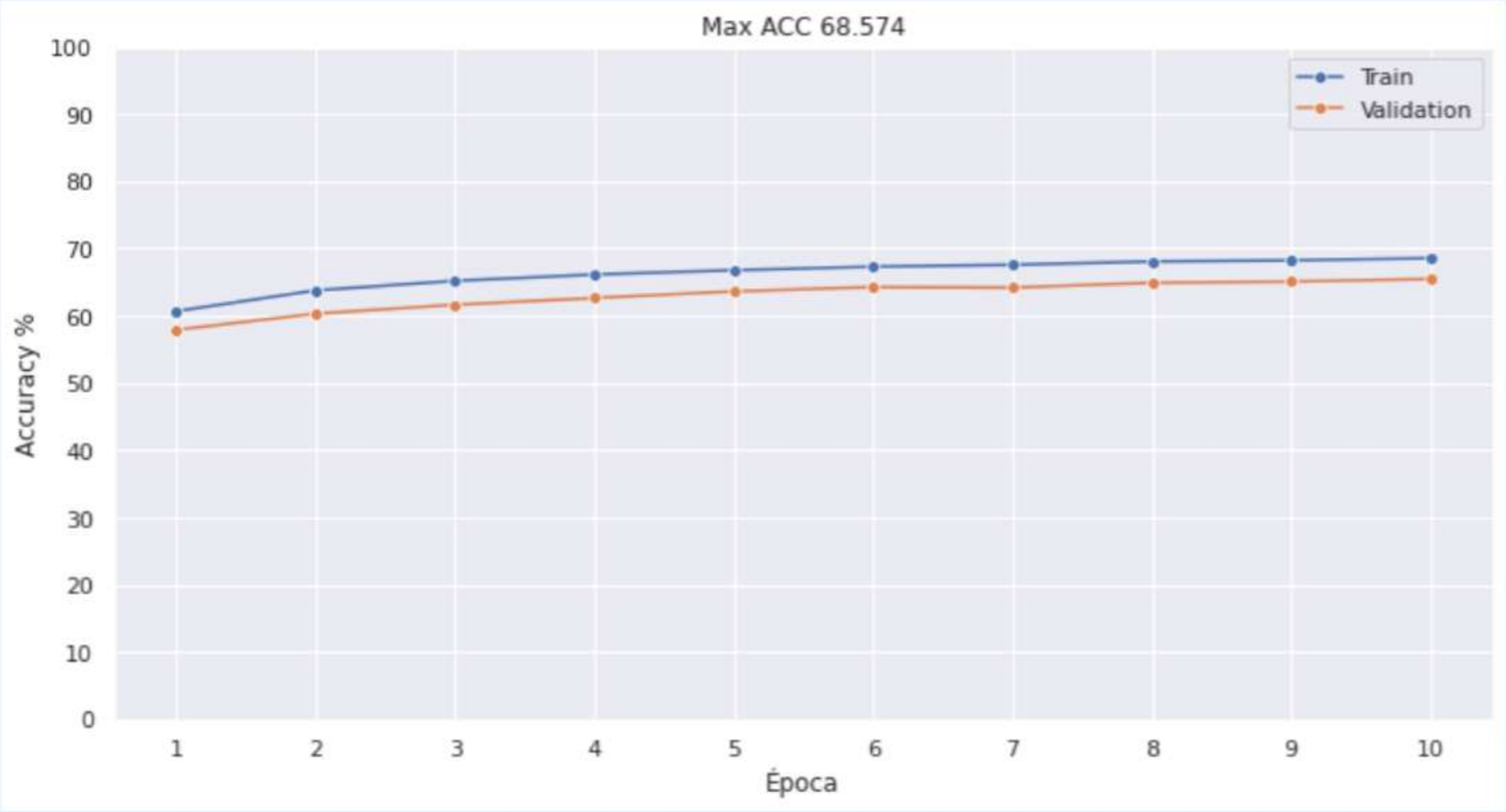

- Modelo inicial. Basado en la libreta 2f
- Corrió bien la primera ocasión. Después ya no.
- Dió pie a la siguiente arquitectura.
- Se añadieron DropOuts

# Arquitectura 1

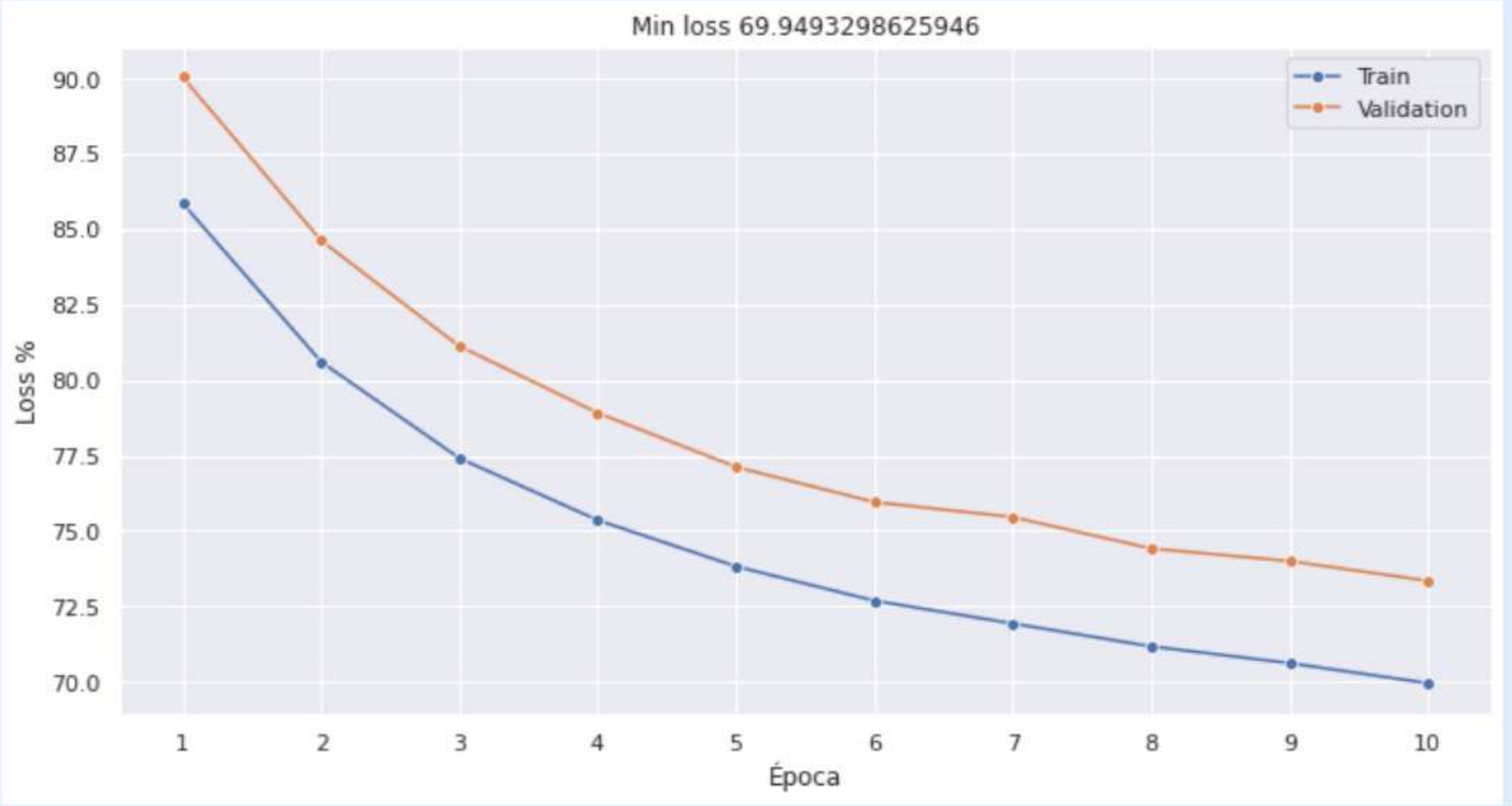



- Funcionó sin mayor problema
- Se pudo ejecutar más de una sola ocasión
- 2 DropOuts
- Accuracy: 68.574
- Loss: 69.949

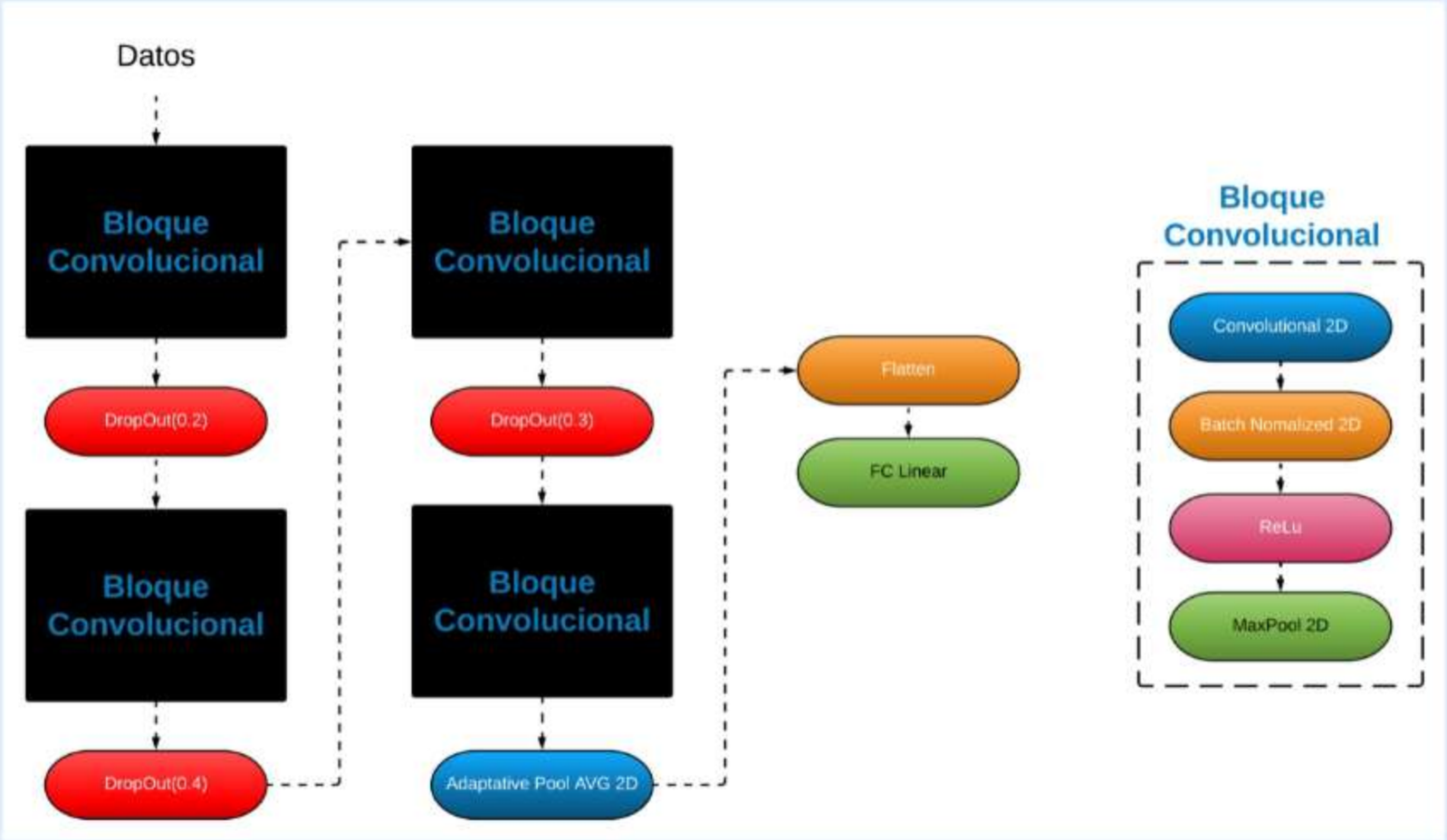
# Arquitectura 1. ACCURACY



# Arquitectura 1. LOSS

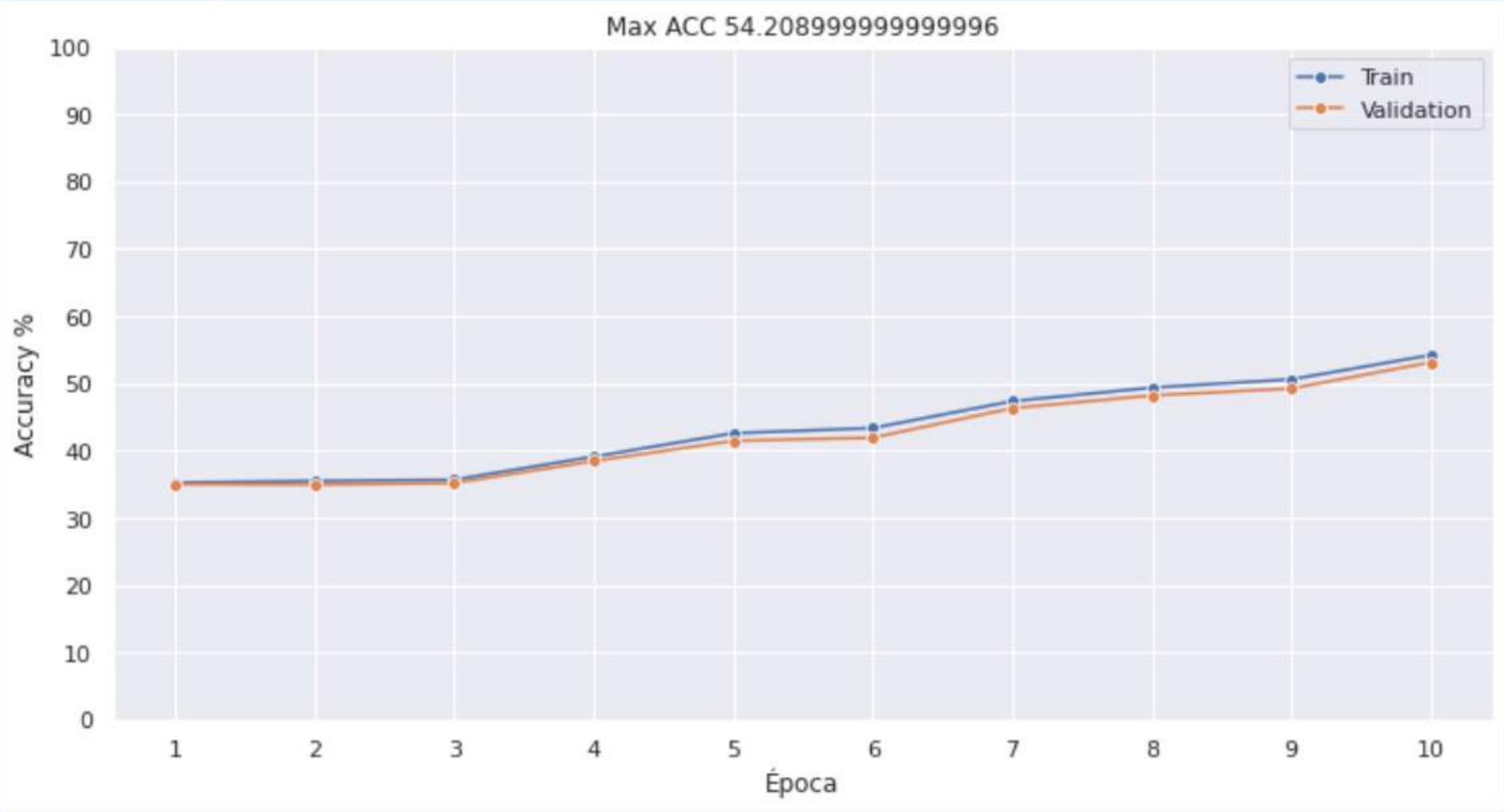


# Arquitectura 2

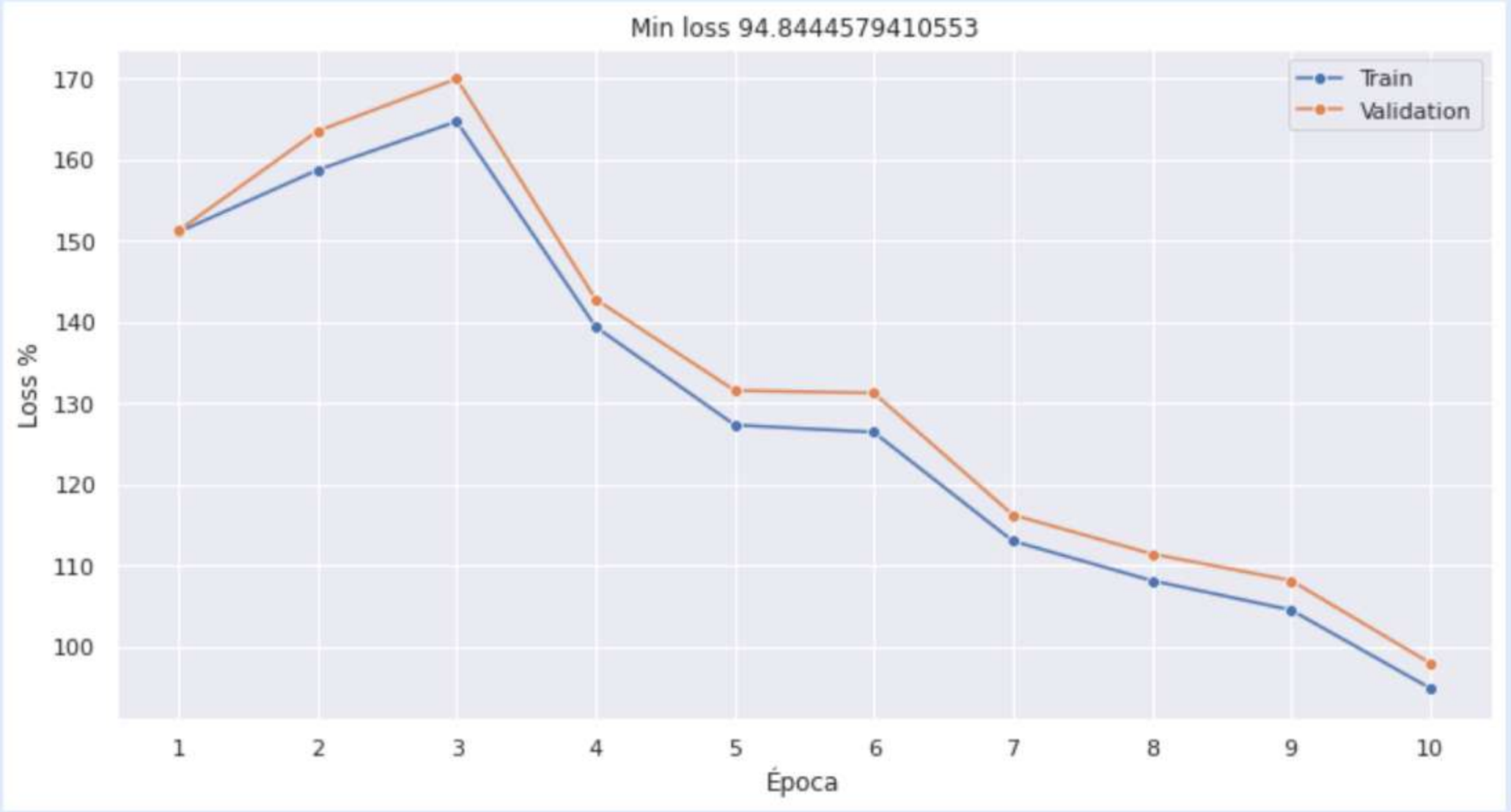



- Modelo inicial. Basado en la libreta 2f
- Curiosamente, corrió sin mayor problema.
- 3 DropOuts
- Accuracy: 54.20
- Loss: 94.84

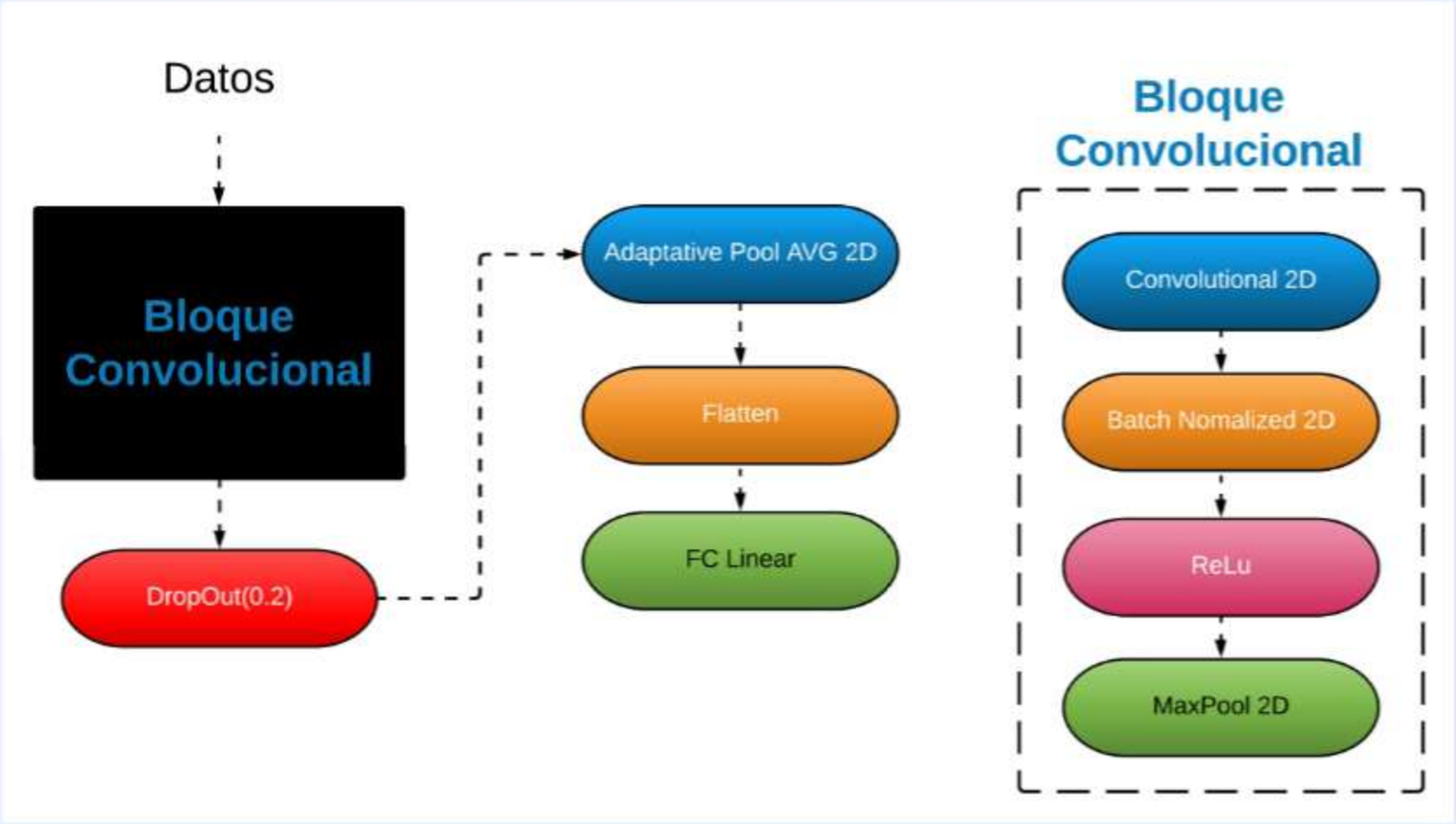
# Arquitectura 2. ACCURACY



# Arquitectura 2. LOSS

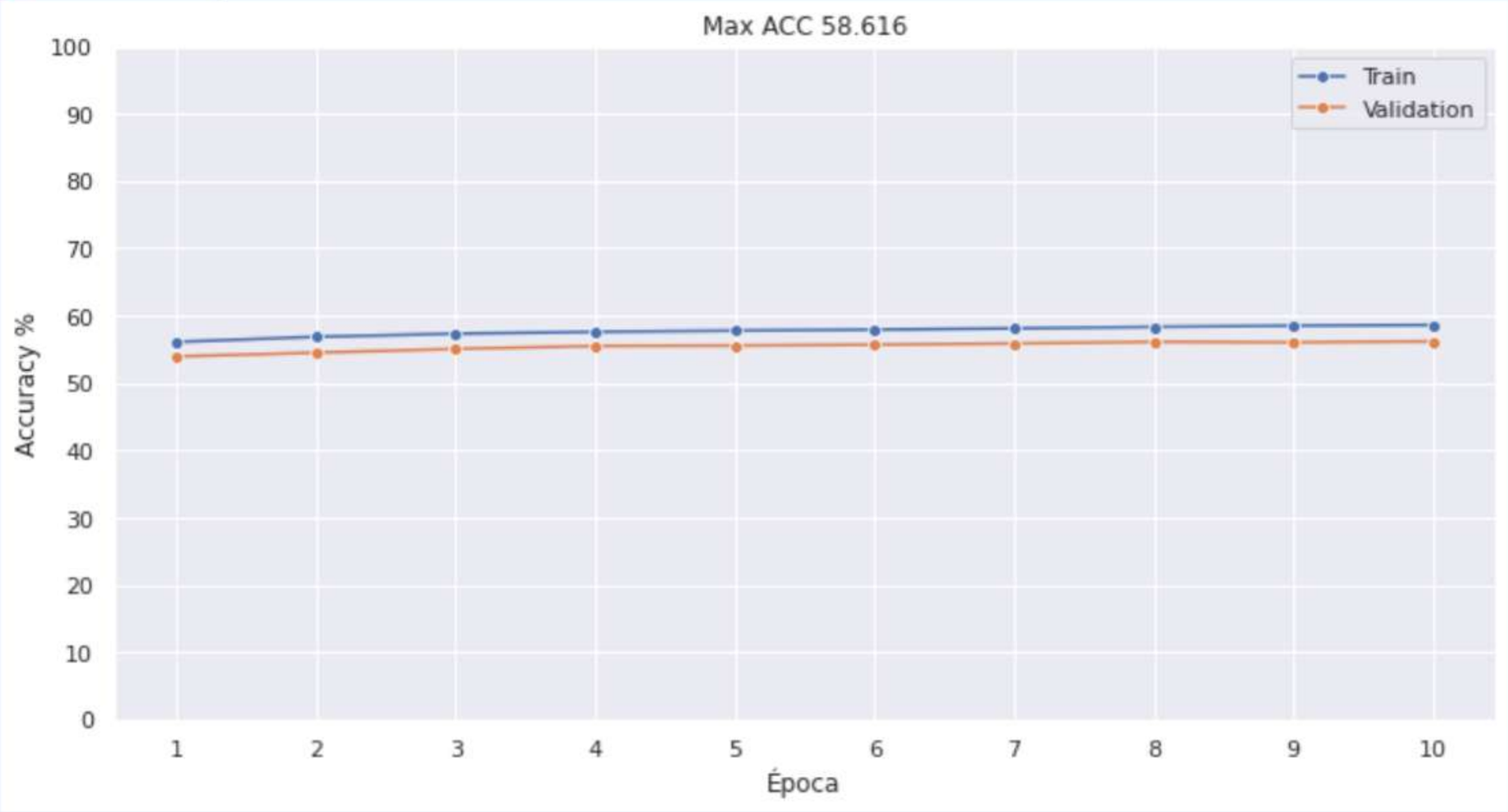


# Arquitectura 3

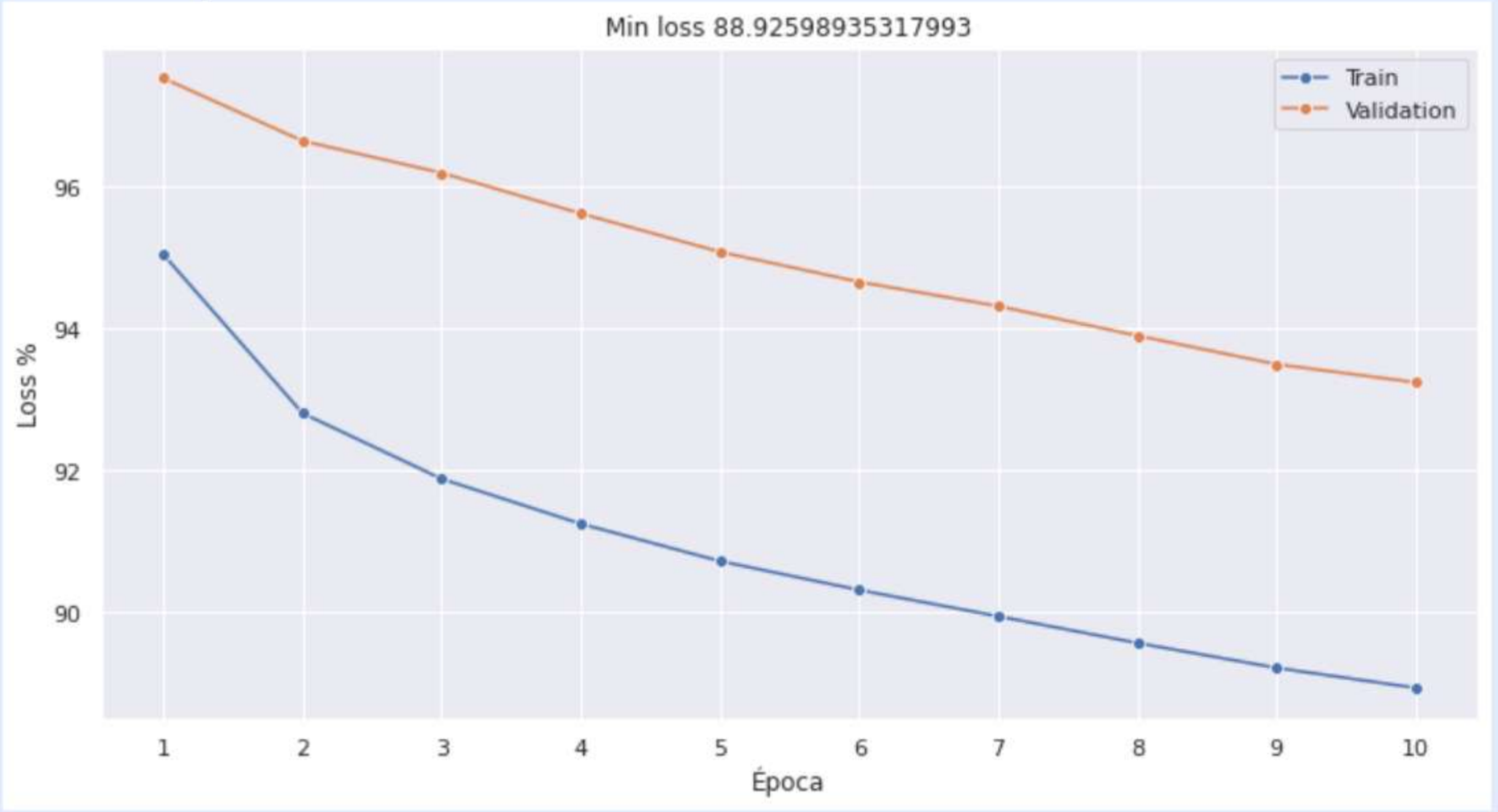



- Consecuencia del contraste en resultados anteriores.
- 1 DropOuts
- Resultados cercanos a la segunda arquitectura
- Accuracy: 58.616
- Loss: 88.92

# Arquitectura 3. ACCURACY



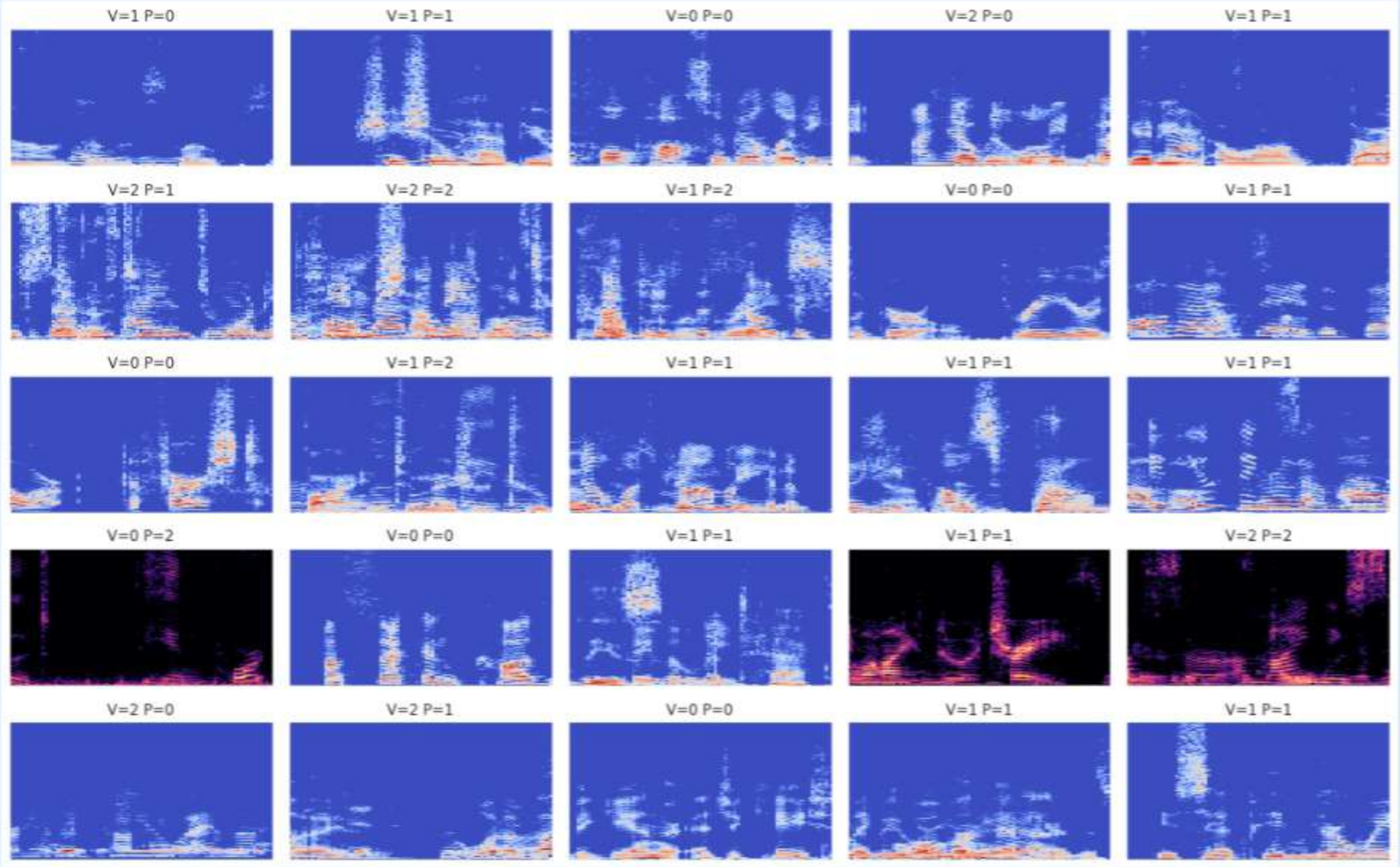
# Arquitectura 3. LOSS



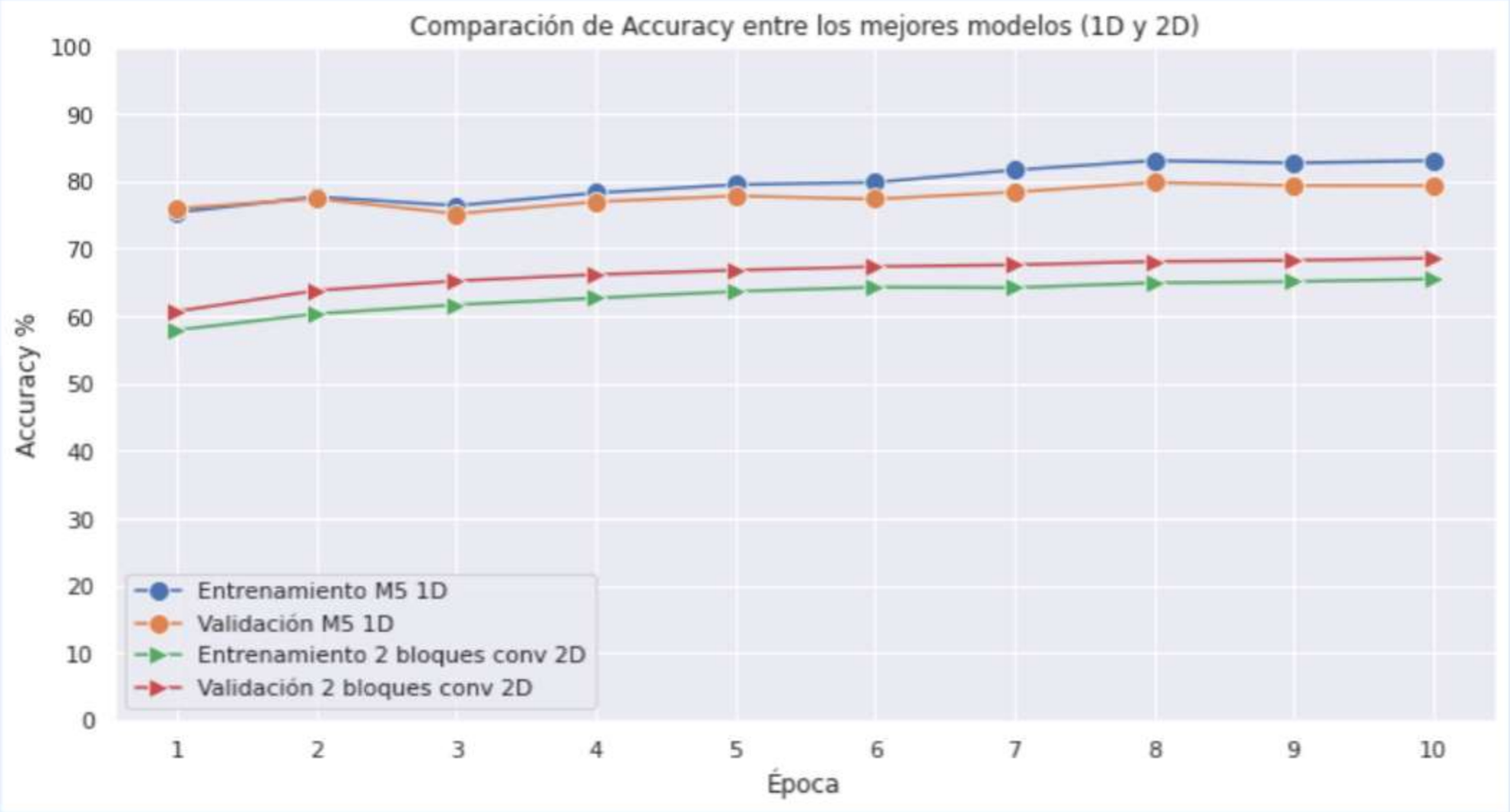
# Resumen

| Arquitectura | Acc | Loss | Time | Parámetros |
|---|---|---|---|---|
| 1 | 68.574 | 69.949 | 28.58 | 9,795 |
| 2 | 54.20 | 94.84 | ... | 65,571 |
| 3 | 58.616 | 88.92 | ... | 483 |

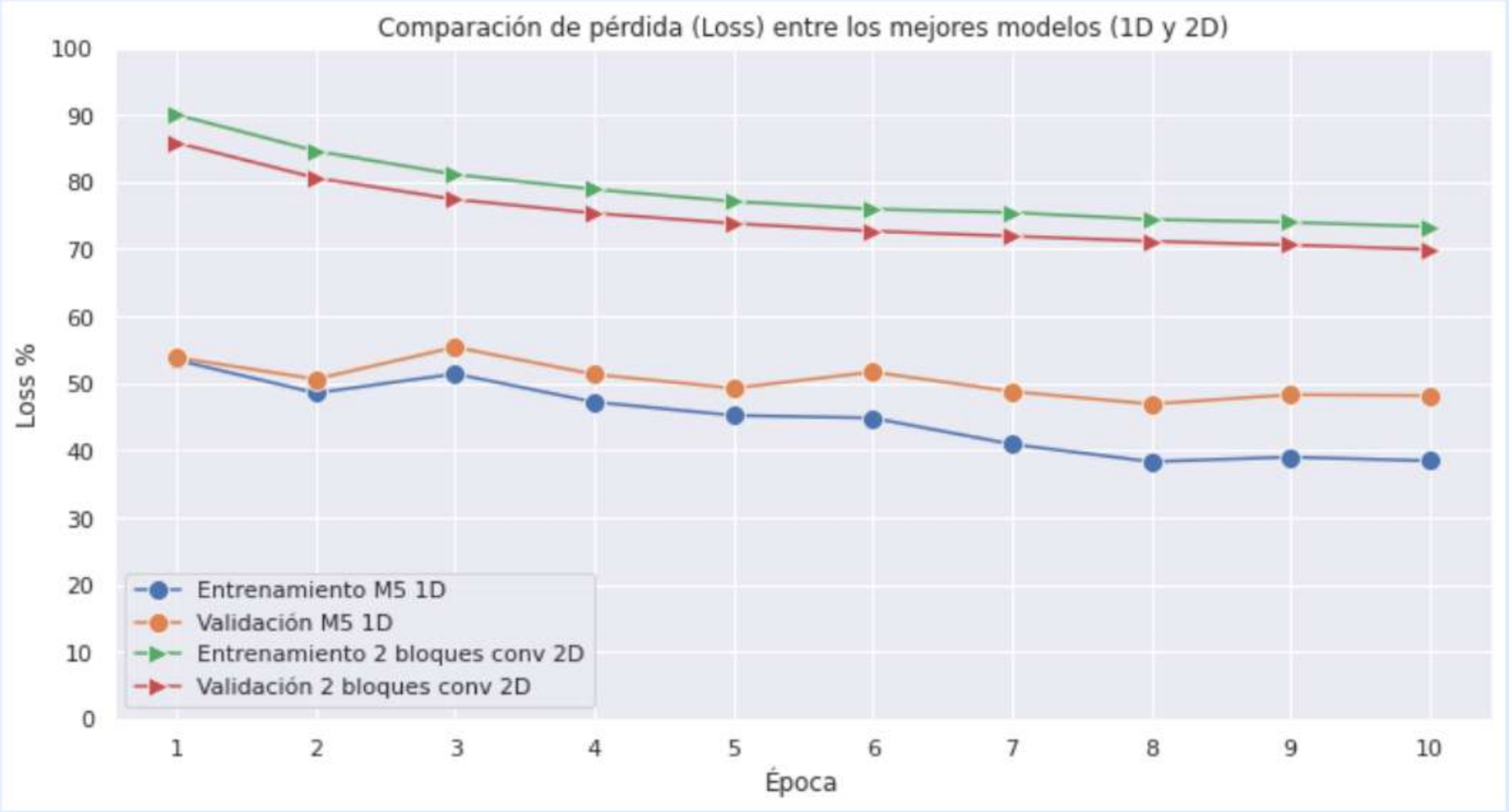
# Evaluación (Arq. 1)



# Conclusiones Accuracy



# Conclusiones Loss



# Conclusiones

- Esta tarea nos resultó complicada desde el momento en que no contábamos con los datos (audios) que nos permitieran atacarla. Afortunadamente, conseguimos generarlos por medio de otros conjuntos de datos que se encontraban en internet y logramos que estuvieran balanceados, lo cual ayudó bastante.
- De los análisis pasados, parece ser que esta problemática se resuelve de una mejor manera  utilizando redes neuronales convolucionales 1D, específicamente con la arquitectura M5 propuesta en el atrículo con las pocas adaptaciones que mencionamos. Es importante decir que decidimos no usar la M34-Res debido a que sus parámetros eran demasiados y creíamos que presentaría dificultades similares a la M11 y M18.
- Por lo tanto, podemos concluir que la resolución parece ser adecuada, pero en un futuro no descartamos atacarla basándonos, por ejemplo, en redes recurrentes. Asimismo, próximamente pretendemos darle más complejidad a esta tarea tratando de predecir el sexo de el, la, los o las participantes en el audio.

# ¿Alguna duda?