

# **UNIVERSIDAD NACIONAL DE INGENIERÍA**

## **UNIDAD DE POSGRADO FACULTAD DE CIENCIAS/ESCUELA DE POSGRADO**



### **TESIS**

**“DESARROLLO DE UN SISTEMA DE PRONÓSTICO DEL  
CENTELLEO IONOSFÉRICO SOBRE EL PERÚ UTILIZANDO  
MACHINE LEARNING PARA LA MITIGACIÓN DE  
PERTURBACIONES EN SEÑALES SATÉLITES”**

**PARA OBTENER EL GRADO ACADÉMICO DE MAESTRO EN CIENCIAS  
CON MENCIÓN EN CIENCIAS DE LA COMPUTACIÓN**

**ELABORADA POR:**

**ALEXANDER OLMEDO VALDEZ PORTOCARRERO**

**ASESOR:**

**MANUEL ALEJANDRO QUISPE TORRES**

**LIMA, PERÚ**

**2025-II**

<b>RESUMEN</b>	<b>4</b>
<b>ABSTRACT</b>	<b>5</b>
<b>CAPÍTULO I PROTOCOLO DE LA INVESTIGACIÓN</b>	<b>8</b>
1.1 Identificación y descripción del problema	8
1.1.1 Problema general	9
1.1.2 Problemas específicos	10
1.2 Objetivos de la investigación	10
1.2.1 Objetivo general	10
1.2.2 Objetivos específicos	10
1.3 Hipótesis y variables	11
1.3.1 Hipótesis general	11
1.3.2 Hipótesis específicas	11
1.3.3 Variables de la investigación	11
1.4 Metodología	12
Enfoque	12
Tipo de investigación	12
Nivel de investigación	12
Diseño de investigación	12
Métodos	12
Población y muestra	12
Técnicas de recolección de datos	12
Técnicas de análisis	12
<b>CAPÍTULO II MARCO TEÓRICO</b>	<b>13</b>
2.1 Antecedentes de investigación	13
Conclusión de los antecedentes	14
2.2 Bases teóricas	14
2.2.1 La ionosfera	14
2.2.2 Centelleo ionosférico	14
2.2.3 Impacto del centelleo en sistemas GNSS	15
2.2.4 Parámetros solares y geomagnéticos	15
2.2.5 TEC, ROTI y datos GNSS	15
2.2.6 Aprendizaje automático aplicado al clima espacial	16
2.2.6.1 Series Temporales en el Contexto Ionosférico	16
2.2.6.2 Redes LSTM(Long Short-Term Memory)	17
Arquitectura LSTM: Mecanismo de Memoria	17
2.2.6.3 Arquitectura LSTM Avanzadas Implementadas	18
2.2.6.4 Predicción Multi-Step: Forecasting de Secuencias Completas	20
Formulación Matemática	20
Ventajas del Enfoque Multi-Step	20
2.2.6.5 Ingeniería de Características Temporales	20
Codificación Cíclica del Tiempo	20
2.2.6.6 Función de Pérdida Especializada: Weighted Focal MSE	21
Componentes de la Loss	21
2.2.6.7 Estrategia de Balanceo: Oversampling Sintético	21

2.2.6.8 Validación Temporal Rigurosa	21
2.2.6.9 División Estratificada de Datos	21
2.2.7 Variantes Arquitectónicas de Redes Recurrentes	22
2.3 Definición de términos (Marco conceptual)	22
<b>CAPÍTULO III MARCO METODOLÓGICO</b>	<b>24</b>
<b>3.1 Tipo y Diseño de Investigación</b>	<b>24</b>
Tipo de investigación	24
Diseño de investigación	24
<b>3.2 Población y Muestra</b>	<b>24</b>
Población	24
Muestra	24
<b>3.3 Método y Enfoque de Investigación</b>	<b>25</b>
<b>3.4 Procedimiento Metodológico</b>	<b>25</b>
3.4.1 Fase 1: Adquisición y Preprocesamiento de Datos	25
a. Fuentes de datos	25
b. Dataset Card: Ionospheric Forecast Dataset	26
c. Preprocesamiento	26
3.4.2 Fase 2: Selección de Variables	27
3.4.3 Fase 3: Modelado Predictivo	28
1. Arquitectura LSTM Simple (Vanilla LSTM)	28
2. Arquitectura LSTM Profunda (Stacked LSTM)	28
3. Arquitectura LSTM Bidireccional (Bi-LSTM)	28
4. Diseño Experimental y Selección de Arquitectura(Benchmark)	29
3.4.4 Fase 4: Evaluación del Modelo	30
Para regresión:	30
Para clasificación:	30
3.4.5 Fase 5: Implementación del Sistema	31
3.4.6 Fase 6: Pruebas de Campo y Mejora Continua	31
<b>CAPÍTULO IV RESULTADOS</b>	<b>32</b>
4.1 Caracterización del Dataset Procesado	32
4.2 Efectividad de la Estrategia de Muestreo	32
4.3 Comportamiento de Variables Predictoras (Feature Analysis)	33
4.4 Evaluación del Modelo Predictivo (Bi-LSTM Multi-Step)	33
4.4.1 Métricas de Desempeño Global vs. Eventos	33
4.5 Análisis de Degradación del Horizonte Temporal	34
4.6 Síntesis de Resultados	34
<b>CAPÍTULO V IMPLEMENTACIÓN DEL SISTEMA</b>	<b>35</b>
5.1 Arquitectura General del Sistema	35
5.2 Entorno de Desarrollo y Ejecución	35
Requerimientos del Sistema	36
5.3 Flujo Operativo y Descripción de Módulos	36
5.3.1 Módulo de Adquisición y Preprocesamiento	36
5.3.2 Módulo del Núcleo Predictivo (Bi-LSTM Multi-Step)	36
5.3.3 Módulo de Post-Procesamiento y Alertas	37

5.4 Estructura Modular del Código Fuente	37
5.5 Sistema de Visualización (Dashboard)	38
5.6 Estrategia de Despliegue y Mantenimiento	38
5.7 Síntesis del Capítulo	38
<b>CAPÍTULO VI DISCUSIÓN, CONCLUSIONES Y RECOMENDACIONES</b>	<b>39</b>
6.1 Discusión	39
6.2 Conclusiones	40
6.3 Recomendaciones	41

## RESUMEN

El centelleo ionosférico es un fenómeno que afecta severamente la propagación de señales GNSS, generando fluctuaciones rápidas en la amplitud y fase que deterioran la precisión del posicionamiento, la continuidad del servicio y la integridad de las comunicaciones satelitales. En regiones ecuatoriales como el Perú, la presencia de la anomalía ecuatorial intensifica estas perturbaciones, especialmente durante horas nocturnas y en períodos de alta actividad solar. Actualmente, el país carece de un sistema operativo de monitoreo y pronóstico que permita anticipar estos eventos en tiempo cuasi-real, lo que genera vulnerabilidad en sectores críticos como aviación, logística, agricultura de precisión, defensa y telecomunicaciones.

En esta tesis se desarrolla un sistema de monitoreo y predicción del centelleo ionosférico sobre el Perú, basado en técnicas avanzadas de aprendizaje automático. Se integran datos multifuente provenientes de OMNIWeb (NASA), SWPC (NOAA), la red LISN, y el Radio Observatorio de Jicamarca (ROJ), consolidando un dataset especializado documentado mediante una Dataset Card. El sistema emplea una arquitectura híbrida compuesta por modelos LSTM multivariados y un enfoque profundo Morph–LSTM–ELM, diseñado para capturar la dinámica temporal del índice S4 y mejorar la sensibilidad ante eventos intensos. La metodología incluye preprocesamiento estandarizado, selección de variables, segmentación temporal, modelado en dos fases (regresión continua y clasificación de severidad) y validación con métricas robustas (RMSE, MAE, R<sup>2</sup>, F1-score).

Los resultados muestran que los modelos propuestos logran predecir el índice S4 con horizontes de 30 a 60 minutos, alcanzando un desempeño adecuado para aplicaciones operativas y superando a enfoques univariados tradicionales. Asimismo, se implementó un prototipo funcional del sistema, con visualización interactiva y capacidad de integración futura en plataformas de vigilancia ionosférica. Este trabajo constituye un aporte científico y tecnológico significativo para el fortalecimiento de la soberanía digital del Perú, y establece las bases para un sistema nacional de alerta temprana de perturbaciones ionosféricas.

**Palabras clave:** centelleo ionosférico, GNSS, índice S4, LSTM, clima espacial, Machine Learning, Morph–LSTM–ELM.

## ABSTRACT

Ionospheric scintillation is a phenomenon that severely impacts the propagation of GNSS signals, causing rapid fluctuations in amplitude and phase that degrade positioning accuracy, service continuity, and the integrity of satellite-based communications. In equatorial regions such as Peru, the presence of the equatorial anomaly intensifies these disturbances, especially during nighttime hours and periods of high solar activity. Currently, Peru lacks an operational monitoring and forecasting system capable of anticipating scintillation events in near-real time, increasing vulnerability in critical sectors such as aviation, logistics, precision agriculture, defense, and telecommunications.

This thesis presents the development of a monitoring and forecasting system for ionospheric scintillation over Peru, based on advanced machine learning techniques. Multisource data from OMNIWeb (NASA), SWPC (NOAA), the LISN network, and the Jicamarca Radio Observatory (ROJ) were integrated to construct a specialized dataset documented through a standardized Dataset Card. The system employs a hybrid architecture that combines multivariate LSTM models with a deep Morph–LSTM–ELM approach designed to capture the temporal dynamics of the S4 index and improve sensitivity to intense scintillation events. The methodology includes data preprocessing, feature selection, temporal segmentation, a two-stage modeling framework (continuous regression and severity classification), and validation using robust metrics (RMSE, MAE, R<sup>2</sup>, F1-score).

Results demonstrate that the proposed models can predict the S4 index 30 to 60 minutes in advance, achieving performance suitable for operational applications and outperforming traditional univariate models. A functional prototype of the system is also implemented, providing interactive visualization and enabling future integration into national ionospheric monitoring platforms. This work represents a significant scientific and technological contribution to strengthening Peru's digital and space-weather resilience and establishes the foundation for a national early-warning system for ionospheric disturbances.

**Keywords:** ionospheric scintillation, GNSS, S4 index, LSTM, space weather, machine learning, Morph–LSTM–ELM.

## INTRODUCCIÓN

El centelleo ionosférico es un fenómeno caracterizado por rápidas fluctuaciones en la amplitud y fase de las señales de radiofrecuencia, originadas por irregularidades en la densidad electrónica de la ionósfera. Estas perturbaciones afectan la propagación de las ondas en la banda L y generan errores significativos en los sistemas GNSS (Global Navigation Satellite Systems), incluyendo degradación en la precisión del posicionamiento, pérdidas intermitentes de seguimiento y, en casos extremos, interrupciones completas del servicio. La problemática adquiere especial relevancia en regiones ecuatoriales, como el Perú, donde la presencia de la anomalía ecuatorial y la formación de burbujas de plasma post crepusculares incrementan la intensidad y frecuencia del centelleo, particularmente durante las horas nocturnas y en períodos de alta actividad solar.

En un contexto global donde las tecnologías de geolocalización, navegación satelital y sincronización temporal son fundamentales para la operación de sectores críticos —aviación, telecomunicaciones, transporte inteligente, agricultura de precisión, minería, defensa, logística y gestión de emergencias—, la predicción y mitigación del centelleo ionosférico se vuelve una necesidad estratégica. En el caso peruano, esta necesidad es aún más urgente debido a la ausencia de un sistema nacional operativo que permita monitorear, caracterizar y pronosticar la ocurrencia de eventos severos de centelleo en tiempo quasi-real. La carencia de este tipo de infraestructura coloca al país en una situación de vulnerabilidad tecnológica, afectando actividades productivas, servicios esenciales y capacidades estratégicas.

A pesar de que existen redes de sensores como la Low-Latitude Ionospheric Sensor Network (LISN), el Radio Observatorio de Jicamarca (ROJ), la ionosonda y bases de datos internacionales como OMNIWeb (NASA) o SWPC (NOAA), su potencial no ha sido plenamente explotado para el desarrollo de modelos predictivos operacionales basados en aprendizaje automático y orientados al entorno geofísico peruano. Asimismo, las investigaciones internacionales han mostrado avances importantes mediante el uso de técnicas como redes neuronales LSTM, Random Forest, Support Vector Machines, algoritmos genéticos, optimización por enjambre de partículas y enfoques híbridos. Sin embargo, la mayoría de estos estudios se ha realizado en contextos geográficos distintos, con dinámicas ionosféricas que no representan adecuadamente la complejidad ecuatorial del Perú.

En este marco, la presente tesis plantea el desarrollo de un sistema inteligente de monitoreo y pronóstico del centelleo ionosférico sobre el territorio peruano, empleando un enfoque de modelado en dos fases: una etapa de regresión basada en redes neuronales LSTM para la predicción continua del índice S4, seguida de una etapa de clasificación supervisada (Random Forest) que categoriza la severidad del evento conforme a estándares establecidos por la UIT (ITU-R). La propuesta integra datos multifuente provenientes de LISN, ROJ, OMNIWeb, SWPC y receptores GNSS distribuidos, complementados por parámetros solares, geomagnéticos e ionosféricos. Para ello, se construyó un dataset especializado documentado mediante una Dataset Card, aplicando procesos rigurosos de preprocesamiento que incluyen interpolación, normalización, corrección de outliers, segmentación temporal y alineación multifuente.

El sistema resultante incorpora métricas robustas para la evaluación de modelos, tales como RMSE, MAE y R<sup>2</sup> para la tarea de regresión, y Accuracy, Precisión y F1-score para la clasificación. Los modelos han sido validados con datos reales provenientes de estaciones ubicadas en regiones estratégicas del país, permitiendo analizar la variabilidad espacial y temporal del centelleo en condiciones de diferente actividad solar y geomagnética.

Esta investigación constituye un aporte significativo en el campo del clima espacial y la geofísica aplicada, al establecer las bases para un sistema nacional de alerta ionosférica que fortalezca la soberanía digital y tecnológica del Perú. Asimismo, proporciona una herramienta operativa para reducir los impactos de las perturbaciones ionosféricas en aplicaciones críticas dependientes de señales GNSS, beneficiando directamente a sectores productivos, estratégicos y científicos.

Finalmente, la estructura del presente documento se organiza de la siguiente manera: el capítulo inicial presenta el estado del arte, la justificación, el aporte científico, la hipótesis y los objetivos de la investigación. Posteriormente, se detalla la metodología empleada, incluyendo la adquisición y preprocesamiento de datos, la construcción del dataset especializado, la descripción del modelo predictivo, el proceso de selección de variables, la arquitectura de modelado en dos fases y los resultados obtenidos. Se incluyen además la Model Card correspondiente, la evaluación de desempeño, la implementación del sistema, las pruebas de campo y el proceso de mejora continua. El documento concluye con las referencias utilizadas y los anexos propuestos que complementan la investigación.

## CAPÍTULO I PROTOCOLO DE LA INVESTIGACIÓN

### 1.1 Identificación y descripción del problema

El centelleo ionosférico constituye una de las principales fuentes de error en la propagación de señales GNSS (Global Navigation Satellite Systems), debido a las fluctuaciones rápidas en la amplitud y fase de las ondas ocasionadas por irregularidades en la densidad electrónica de la ionósfera. Estas perturbaciones afectan la integridad, disponibilidad, continuidad y precisión de los servicios GNSS, generando interrupciones intermitentes, pérdida de seguimiento y degradación del posicionamiento satelital. Su impacto es especialmente crítico en aplicaciones que demandan alta confiabilidad, como la aviación, navegación marítima, agricultura de precisión, vehículos autónomos, telecomunicaciones, sincronización de redes eléctricas, logística urbana y servicios basados en geolocalización.

El Perú, ubicado en las proximidades del Ecuador Magnético, presenta una de las regiones ionosféricas más activas y complejas del planeta. La anomalía ecuatorial y la formación de burbujas de plasma post crepusculares intensifican significativamente la ocurrencia del centelleo, especialmente durante las horas nocturnas y en períodos de elevada actividad solar. Esta dinámica ionosférica altamente variable crea un entorno geofísico donde las perturbaciones en señales satelitales ocurren de forma frecuente, localizada y difícilmente predecible.

Aunque existen redes de observación como LISN (Low-Latitude Ionospheric Sensor Network), el Radio Observatorio de Jicamarca (ROJ), la ionosonda, y bases de datos internacionales como OMNIWeb (NASA) y SWPC (NOAA), el Perú carece de un sistema nacional operativo y automatizado que permita monitorear, caracterizar y pronosticar en tiempo quasi-real los eventos de centelleo ionosférico. Esta falta de infraestructura predictiva incrementa la vulnerabilidad tecnológica del país y limita la capacidad de emitir alertas tempranas que mitiguen el impacto sobre sistemas críticos dependientes de GNSS.

La creciente dependencia nacional hacia tecnologías basadas en posicionamiento y navegación satelital evidencia la necesidad urgente de desarrollar herramientas predictivas confiables, basadas en el análisis multifuente y en técnicas avanzadas de aprendizaje automático, que permitan anticipar condiciones ionosféricas adversas y fortalecer las capacidades estratégicas de vigilancia del clima espacial en el territorio peruano.

#### 1.1.1 Problema general

¿Cómo desarrollar un sistema predictivo automatizado y validado que permita pronosticar el centelleo ionosférico sobre el Perú, integrando datos multifuente y técnicas avanzadas de aprendizaje automático, a fin de mitigar las perturbaciones en las señales GNSS?

#### 1.1.2 Problemas específicos

- a) ¿Cómo integrar adecuadamente datos multifuente provenientes de redes GNSS, sensores ionosféricos y parámetros solares y geomagnéticos para construir un dataset confiable del índice S4?
- b) ¿Qué técnicas de preprocessamiento permiten mejorar la calidad, consistencia temporal y

representatividad del dataset para tareas de predicción?

c) ¿Qué arquitectura de aprendizaje automático es más adecuada para modelar la dinámica temporal del centelleo ionosférico en la región ecuatorial peruana?

d) ¿Qué métricas y procedimientos de evaluación permiten validar la precisión y estabilidad del modelo predictivo?

e) ¿Cómo validar el sistema mediante datos reales provenientes de estaciones distribuidas en diferentes regiones del Perú?

## **1.2 Objetivos de la investigación**

### **1.2.1 Objetivo general**

Desarrollar un sistema inteligente de pronóstico del centelleo ionosférico sobre el Perú, basado en técnicas de aprendizaje automático e integración de datos multifuente, con el propósito de anticipar perturbaciones en señales GNSS y fortalecer la toma de decisiones en aplicaciones críticas dependientes de geolocalización satelital.

### **1.2.2 Objetivos específicos**

1. Integrar y procesar datos multifuente provenientes de redes GNSS, LISN, ROJ, OMNIWeb y SWPC mediante técnicas de interpolación, normalización, sincronización temporal y segmentación, para construir un dataset especializado del índice S4.
2. Aplicar métodos de selección de variables, incluyendo Random Forest, análisis de correlación y técnicas interpretables (SHAP), con el fin de identificar los parámetros geofísicos más relevantes para la predicción del centelleo.
3. Diseñar y validar modelos basados en redes LSTM para la predicción continua del índice S4 utilizando métricas cuantitativas robustas (RMSE, MAE, R<sup>2</sup>).
4. Implementar modelos de clasificación supervisada como Random Forest y SVM para categorizar la severidad del centelleo de acuerdo con estándares ITU-R, evaluados mediante precisión, recall y F1-score.
5. Integrar los modelos predictivos en una arquitectura funcional de pronóstico capaz de operar en tiempo quasi-real.
6. Validar el sistema mediante pruebas experimentales con datos de estaciones GNSS distribuidas en regiones estratégicas del Perú, desarrollando un proceso iterativo de mejora continua.

## **1.3 Hipótesis y variables**

### **1.3.1 Hipótesis general**

La integración de datos solares, geomagnéticos e ionosféricos, combinada con técnicas de aprendizaje automático basadas en redes LSTM y clasificadores supervisados, permite desarrollar un sistema capaz de predecir el centelleo ionosférico sobre el Perú con precisión suficiente para su aplicación en la mitigación de perturbaciones en señales GNSS.

### **1.3.2 Hipótesis específicas**

- a) La combinación de variables solares, geomagnéticas e ionosféricas mejora la capacidad predictiva del sistema respecto a modelos univariados.
- b) Las redes neuronales LSTM capturan las dependencias temporales del índice S4 con mayor precisión que los métodos tradicionales.
- c) Los clasificadores supervisados permiten categorizar de manera precisa la severidad del centelleo según estándares internacionales.
- d) La inclusión de datos locales de LISN y ROJ incrementa la representatividad del modelo y mejora la predicción para condiciones ionosféricas del Perú.

### **1.3.3 Variables de la investigación**

#### **Variable dependiente:**

- Índice de centelleo ionosférico (S4).

#### **Variables independientes:**

- Parámetros solares: F10.7, velocidad y densidad del viento solar, IMF.
- Parámetros geomagnéticos: Kp, AE, Dst.
- Parámetros ionosféricos: TEC, ROTI, gradientes espaciales.
- Parámetros GNSS: intensidad de señal, elevación satelital, geometría satelital.
- Variables temporales: hora local, estacionalidad, época del año, fase del ciclo solar.

## **1.4 Metodología**

### **Enfoque**

La investigación adopta un enfoque cuantitativo y experimental, basado en el análisis estadístico multivariado y en técnicas de aprendizaje automático aplicadas a series temporales ionosféricas.

### **Tipo de investigación**

Aplicada, dado que busca resolver un problema tecnológico concreto mediante el desarrollo de un sistema predictivo operativo.

### **Nivel de investigación**

Explicativo y predictivo, pues se busca modelar las causas del centelleo y anticipar su comportamiento.

### **Diseño de investigación**

Diseño no experimental, longitudinal y basado en series temporales multivariadas.

### **Métodos**

- Adquisición de datos multifuente.
- Depuración, normalización, interpolación y alineación temporal.
- Construcción del Ionospheric Forecast Dataset.
- Selección de variables mediante métodos basados en importancia y correlación.
- Modelado predictivo mediante redes LSTM (regresión) y clasificadores supervisados (Random Forest, SVM).
- Validación experimental utilizando métricas robustas.
- Implementación del sistema en un entorno funcional.

### **Población y muestra**

- **Población:** conjunto de observaciones ionosféricas y geofísicas registradas sobre el territorio peruano.
- **Muestra:** datos provenientes de estaciones GNSS, LISN, ROJ y bases internacionales correspondientes al periodo analizado.

### Técnicas de recolección de datos

- Receptores GNSS dual-frequency.
- Ionosonda y radar incoherente del ROJ.
- Bases de datos OMNIWeb, SWPC, LISN y sensores distribuidos.

### Técnicas de análisis

- Preprocesamiento avanzado de series temporales.
- Feature selection y análisis multivariado.
- Modelado secuencial LSTM.
- Clasificación supervisada.
- Evaluación mediante métricas RMSE, MAE, R<sup>2</sup>, precisión y F1-score.

## CAPÍTULO II MARCO TEÓRICO

El marco teórico desarrolla los fundamentos conceptuales, antecedentes científicos y definiciones claves que sustentan esta investigación. Se organiza en tres secciones: (1) antecedentes de investigación, (2) bases teóricas y (3) definiciones operacionales, integrando literatura especializada en clima espacial, ionosfera, centelleo y aprendizaje automático aplicado a la predicción de fenómenos geofísicos.

### 2.1 Antecedentes de investigación

La investigación sobre el centelleo ionosférico y su impacto en los sistemas GNSS ha evolucionado de manera significativa durante las últimas décadas, avanzando desde estudios observacionales clásicos hasta modelos predictivos basados en aprendizaje automático. En los últimos años, distintos trabajos han demostrado que las técnicas de inteligencia artificial ofrecen un marco adecuado para modelar la naturaleza altamente no lineal y variable del entorno ionosférico ecuatorial.

En esta línea, **Atabati et al. (2021)** mostraron que los modelos híbridos, combinando redes neuronales y algoritmos evolutivos, pueden capturar con eficacia la dinámica del S4 y el ROTI, especialmente bajo condiciones de baja actividad solar. Este resultado es relevante para la presente investigación porque confirma que los enfoques no lineales y las arquitecturas asociadas a optimización global permiten mejorar la capacidad de predicción frente a métodos más tradicionales.

Otros trabajos han puesto especial énfasis en la predictibilidad de parámetros ionosféricos a partir de series de tiempo GNSS. **Zewdie et al. (2021)**, por ejemplo, emplearon Random Forest para selección de variables y modelos LSTM para el pronóstico del contenido total de electrones (TEC) en bajas latitudes. Sus hallazgos —particularmente la mejora sustancial obtenida mediante la combinación RF-LSTM— respaldan el enfoque metodológico adoptado en esta tesis, que incorpora tanto selección de características como modelos avanzados de memoria temporal para capturar la variabilidad ionosférica en escalas de pocas horas.

Por su parte, **Wu (2020)** demostró que las observaciones de occultación GNSS permiten caracterizar irregularidades como la capa E esporádica y las burbujas de plasma ecuatorial, utilizando mediciones de SNR de alta frecuencia para lograr pronósticos globales casi en tiempo real. Aunque este enfoque difiere en la fuente de datos, su aporte es fundamental para entender la importancia de integrar mediciones GNSS diversas en los sistemas modernos de predicción del centelleo.

En latitudes altas, **Lamb et al. (2019)** introdujeron modelos de machine learning con funciones de pérdida personalizadas para predecir eventos de centelleo con una hora de anticipación, mostrando que los métodos no lineales superan a los modelos lineales clásicos. Este patrón se repite en **McGranaghan et al. (2018)**, quienes incorporaron parámetros del viento solar y la actividad geomagnética en modelos SVM, evidenciando que la integración de datos heterogéneos incrementa de manera notable la capacidad predictiva. Ambos estudios fortalecen la premisa de esta tesis: la predicción de centelleo requiere modelos no lineales, multivariados y adaptados a la complejidad del entorno espacial.

En regiones ecuatoriales, la tendencia hacia modelos híbridos también se observa en **Sridhar et al. (2017)**, quienes combinaron redes neuronales con algoritmos de optimización por enjambre de partículas (PSO), obteniendo correlaciones elevadas entre los valores predichos y observados. De manera complementaria, **Anderson y Redmon (2017)** desarrollaron FIRST, una herramienta operativa basada en machine learning para predecir actividad de *Equatorial Spread-F* (ESF), alcanzando niveles de éxito superiores al 80%.

Estos trabajos son especialmente significativos dado que el Perú se encuentra dentro de la región de fuerte influencia ecuatorial, lo cual refuerza la pertinencia de emplear técnicas similares en el presente estudio.

A nivel regional, **Valladares y Chau (2012)** presentan la arquitectura de la red LISN, que constituye actualmente la base de datos más completa para el estudio del comportamiento ionosférico en Latinoamérica. Dado que la presente investigación utiliza datos GNSS procedentes de estaciones del IGP —parte de dicha infraestructura—, este trabajo representa un soporte fundamental para la disponibilidad, calidad y continuidad de los datos usados en esta tesis.

Finalmente, desde una perspectiva histórica, **Aarons (1997)** estableció las bases conceptuales para el entendimiento moderno del centelleo ecuatorial, describiendo la evolución de las irregularidades ionosféricas durante más de cinco décadas de observaciones. Su revisión continúa siendo un marco teórico esencial para interpretar los fenómenos estudiados en el Perú, especialmente en zonas de fuerte influencia magnética como la región ecuatorial.

En conjunto, estos estudios demuestran que la predicción del centelleo ionosférico se beneficia significativamente del uso de técnicas de inteligencia artificial, de la inclusión de múltiples parámetros solares, geomagnéticos e ionosféricos, y del desarrollo de modelos específicos para la región ecuatorial. Sin embargo, pese a los avances internacionales, todavía existe escasa investigación enfocada directamente en el Perú, lo cual justifica y motiva el desarrollo del presente trabajo.

## Conclusión de los antecedentes

Los estudios internacionales han demostrado que:

- Las redes neuronales LSTM son el estado del arte para series ionosféricas.
- Los clasificadores supervisados (RF, SVM) mejoran la detección de eventos severos.
- La predicción mejora significativamente al integrar parámetros solares + geomagnéticos + ionosféricos.
- La región ecuatorial *requiere modelos específicos* debido a su comportamiento altamente no lineal.

**Sin embargo**, pocos trabajos están orientados directamente al Perú, lo que justifica esta investigación.

## 2.2 Bases teóricas

Esta sección presenta los fundamentos que sustentan la investigación, organizados en temas centrales.

### 2.2.1 La ionosfera

La ionosfera es la región de la atmósfera situada entre 60 y 1000 km, compuesta por plasma (electrones e iones). Su comportamiento depende del ciclo solar, radiación UV y rayos X, así como interacciones magnetosféricas. Se divide en capas D, E y F, siendo esta última la de mayor relevancia para las señales GNSS.

### 2.2.2 Centelleo ionosférico

El centelleo ionosférico (ionospheric scintillation) se refiere a fluctuaciones rápidas en amplitud y fase de señales GNSS causadas por irregularidades de tamaño entre metros y kilómetros en la densidad electrónica.

Se cuantifica mediante:

- **S4** (amplitud)
- $\sigma\phi$  (fase)
- **ROTI** (derivada temporal del TEC)

Regiones ecuatoriales presentan las perturbaciones más intensas del planeta debido a:

- Anomalía ecuatorial (Appleton).
- Burbujas de plasma generadas por el mecanismo Rayleigh–Taylor.
- Inestabilidades post crepusculares intensificadas durante la máxima actividad solar.

### 2.2.3 Impacto del centelleo en sistemas GNSS

El centelleo afecta directamente:

- Continuidad del enlace satelital.
- Precisión en algoritmos de navegación.
- Seguimiento de fase y bloqueo del receptor.
- Sincronización de redes eléctricas y telecomunicaciones.

Estos efectos se traducen en degradación operativa en aviación, defensa, agricultura de precisión, logística y sistemas autónomos.

### 2.2.4 Parámetros solares y geomagnéticos

El comportamiento ionosférico está fuertemente influenciado por parámetros del clima espacial:

- **F10.7**: indicador de radiación solar ultravioleta.
- **IMF / Bz**: interacción Sol–Tierra.
- **Velocidad y densidad del viento solar**.
- **Índices Kp, AE, Dst**: caracterizan perturbaciones geomagnéticas.

La integración de estas variables mejora la capacidad predictiva de modelos ML.

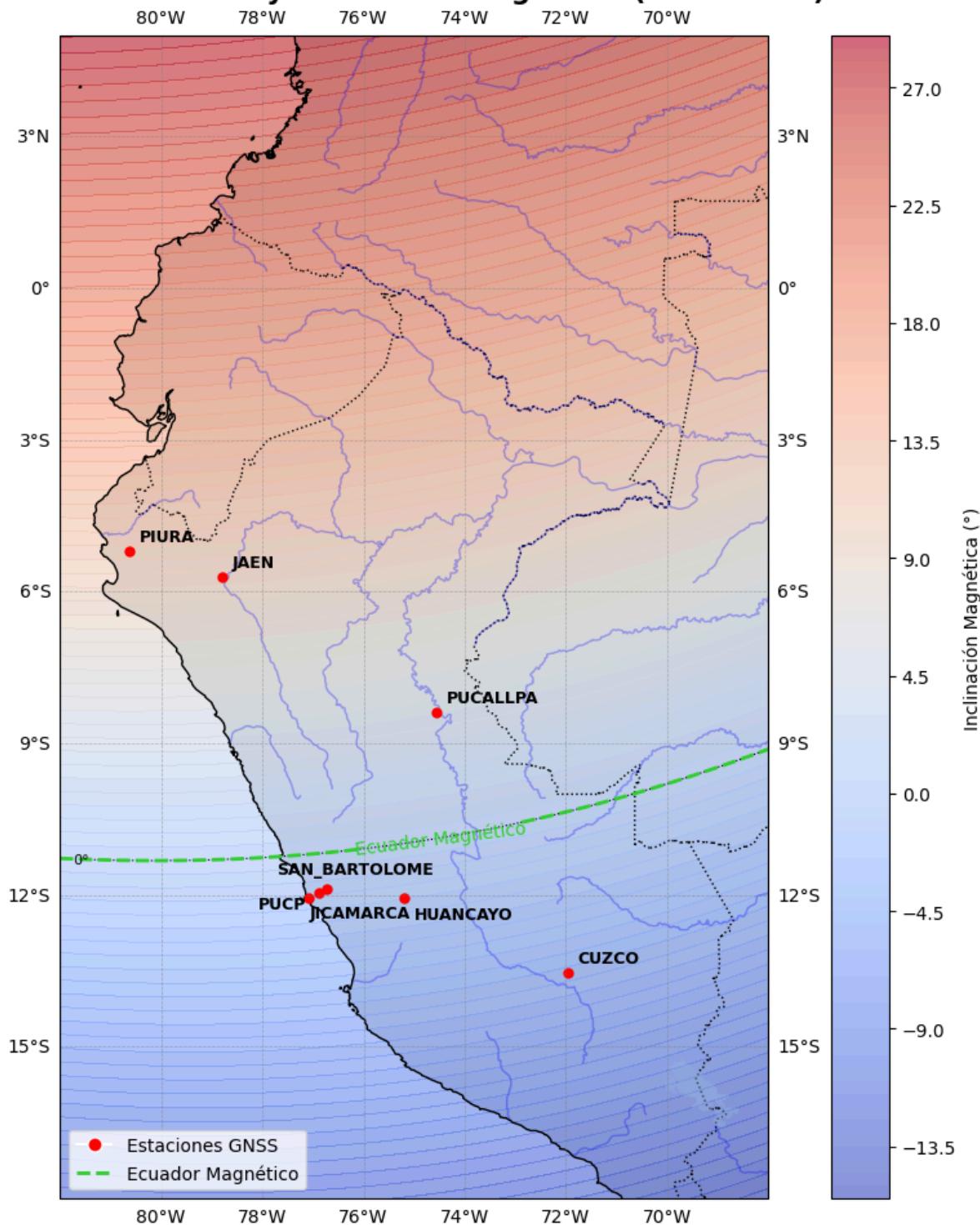
### 2.2.5 TEC, ROTI y datos GNSS

El **TEC** (Total Electron Content) es uno de los indicadores más usados para evaluar la estructura de la ionosfera.

El **ROTI** detecta irregularidades utilizando series temporales de TEC.

Los receptores GNSS dual-frequency permiten obtener estos parámetros a alta resolución.

## Estaciones GNSS y Inclinación Magnética (WMM 2020)



## 2.2.6 Aprendizaje automático aplicado al clima espacial

### 2.2.6.1 Series Temporales en el Contexto Ionosférico

Las series temporales de índices de cintilación ionosférica, como el índice S4, presentan características particulares que las diferencian de series temporales convencionales:

#### Características principales:

- **No Linealidad:** Las relaciones entre variables geofísicas ( $K_p$ ,  $Dst$ ,  $AE$ ,  $F10.7$ ) y el índice S4 no siguen patrones lineales simples. La actividad solar puede desencadenar tormentas ionosféricas con dinámicas complejas e impredecibles mediante modelos lineales clásicos.
- **Memoria temporal extendida:** Los eventos de cintilación pueden presentar efectos de acumulación, donde la actividad geomagnética de horas previas influye en el estado presente de la ionósfera.
- **Intermitencia y picos abruptos:** A diferencia de series suaves, el índice S4 puede pasar de valores base ( $\sim 0.2$ ) a picos críticos ( $> 0.6$ ) en cuestión de minutos, representando transiciones de estado súbitas.
- **Desbalance de clases severo:** Los eventos críticos de cintilación ( $S4 > 0.6$ ) representan menos del 5% de las observaciones totales, generando un problema de clases altamente desbalanceadas.

#### Modelos Clásicos de Series Temporales

##### ARIMA (AutoRegressive Integrated Moving Average):

- Modelos lineales tradicionales para forecasting
- Limitaciones: Asumen linealidad y estacionariedad, inadecuados para capturar las transiciones abruptas características de las tormentas ionosféricas
- Aplicabilidad: Útiles solo para predicción de valores base en condiciones quietas

##### Redes Neuronales Recurrentes (RNN):

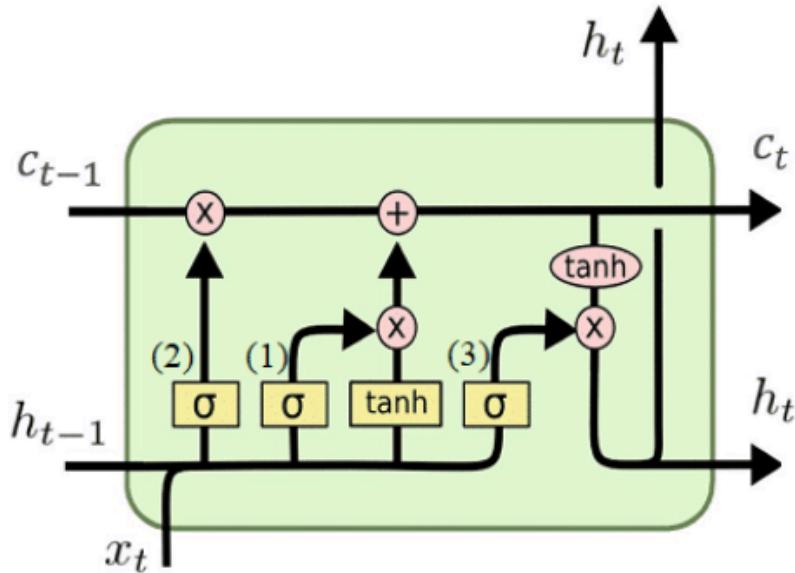
- Capturan dependencias secuenciales mediante conexiones recurrentes
- Problema del desvanecimiento del gradiente: Dificultad para aprender dependencias de largo plazo ( $> 10-15$  timesteps), limitando su capacidad para capturar la memoria extendida de fenómenos ionosféricos

### 2.2.6.2 Redes LSTM(Long Short-Term Memory)

Las LSTM, introducidas por Hochreiter y Schmidhuber(1997), representan el **estado del arte** para modelado de series temporales con memoria de largo plazo. Su arquitectura especializada las hace ideales para predicción de índices de cintilación ionosférica.

## Arquitectura LSTM: Mecanismo de Memoria

Una celda LSTM incorpora tres puertas(gates) que regulan el flujo de información:



## LSTM (Long-Short Term Memory)

*Diagrama de una célula de memoria LSTM*

1. Puerta de Olvido(Forget Gate):

$$f_t = \sigma[h_{t-1}, x_t] + b_f$$

Decide qué información del estado celular anterior descartar

2. Puerta de Entrada(Input Gate):

$$i_t = \sigma[W_i \cdot x_t] + b_i$$

$$\hat{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

Determina qué nueva información almacenar

3. Actualización del Estado Celular:

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t$$

Combina información pasada (olvidada selectivamente) con nueva

4. Puerta de Salida(Output Gate):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot \tanh(C_t)$$

Controla qué parte del estado celular se expone como salida

Donde  $\sigma$  es la función sigmoide,  $\tanh$  es la tangente hiperbólica, y  $\odot$  denota multiplicación elemento a elemento.

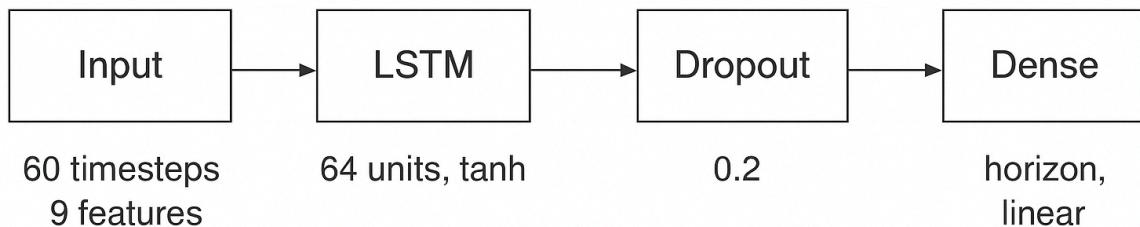
### Ventajas de LSTM para Predicción de S4

Característica	Beneficio para Clima Espacial
<b>Memoria de largo plazo</b>	Captura efectos acumulativos de actividad geomagnética(hasta 60+ minutos previos)
<b>Aprendizaje no lineal</b>	Modela relaciones complejas entre índice solares(Kp,Dst,AE) y S4
<b>Robustez al desvanecimiento del gradiente</b>	Permite entrenamiento estable en secuencias largas(lookback de 60 minutos)
<b>Capacidad de detectar transiciones</b>	Identifica patrones de onset que preceden picos de cintilación.

### 2.2.6.3 Arquitectura LSTM Avanzadas Implementadas

En esta investigación se implementaron y compararon tres arquitecturas especializadas:

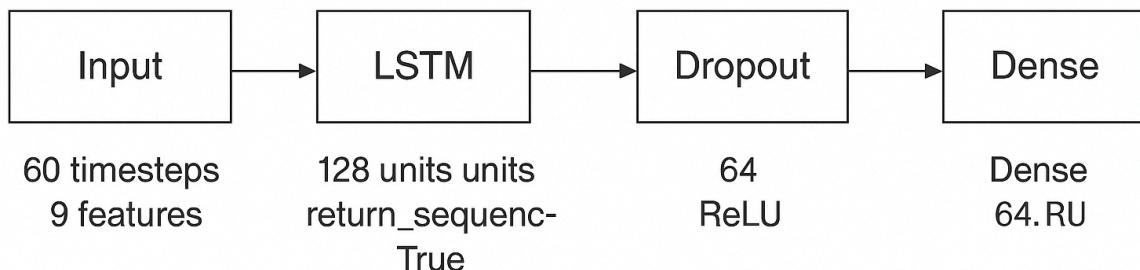
#### Arquitectura1: LSTM Simple(Baseline)



#### Características:

- Modelo base con ~23,000 parámetros
- Una sola capa recurrente
- Ideal para establecer rendimiento de referencia

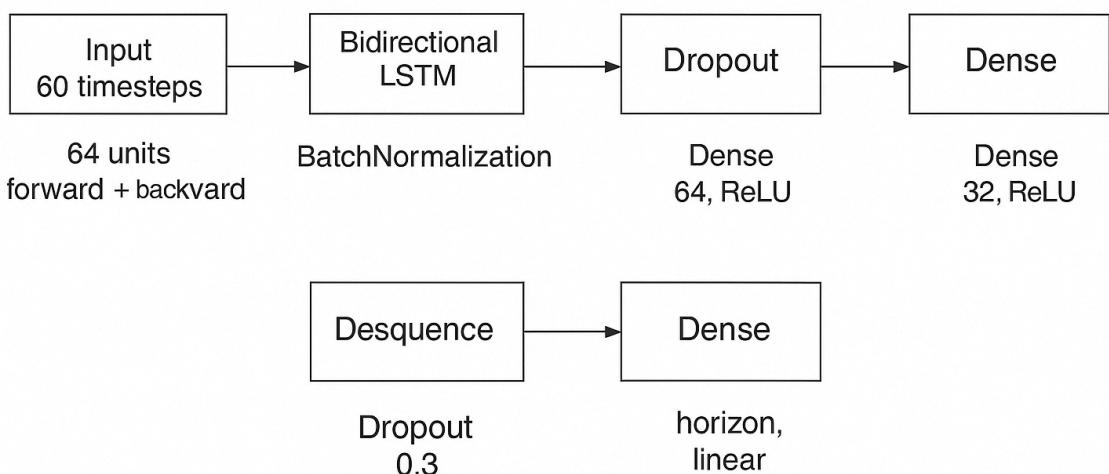
#### Arquitectura 2: LSTM Stacked(Apilada)



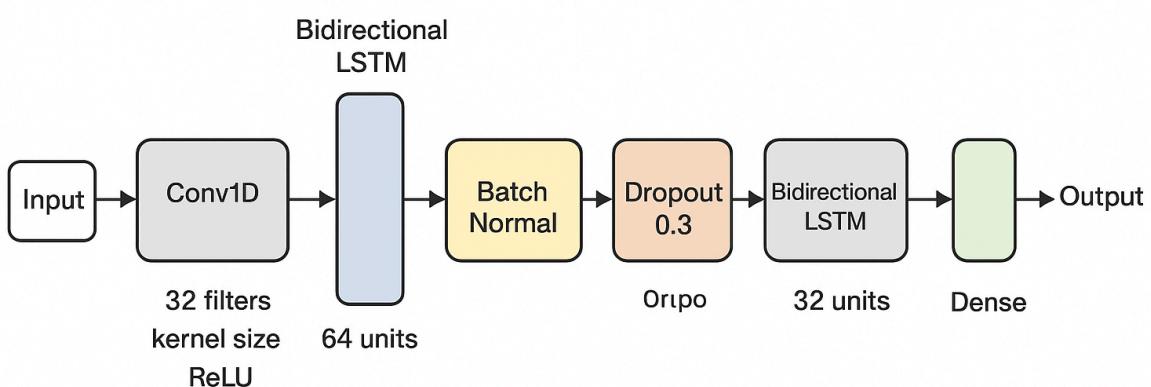
#### Características:

- ~67,500 parámetros
- Aprendizaje jerárquico: Primera capa captura patrones de bajo nivel, segunda capa integra contexto temporal más abstracto
- Mayor capacidad de representación para patrones complejos

### Arquitectura 3: LSTM Bidireccional



### Arquitectura de Modelo



### Características:

- ~45,800 parámetros
- Procesamiento bidireccional: Analiza la secuencia tanto hacia adelante como hacia atrás, capturando dependencias contextuales completas
- BatchNormalization estabiliza entrenamiento y acelera convergencia

- Ventaja clave: Explota la estructura completa de la ventana de entrada para detectar patrones precursores de tormentas

#### **2.2.6.4 Predicción Multi-Step: Forecasting de Secuencias Completas**

A diferencia de enfoques tradicionales que predicen un único valor futuro (single-step: S4 en t+20 minutos), esta investigación implementa **predicción multi-step**, donde el modelo genera una **secuencia completa** de valores futuros.

##### **Formulación Matemática**

###### **Single-Step (enfoque tradicional):**

$$X_t = [s_{t-60}, s_{t-59}, \dots, s_{t-1}] \varepsilon R^{60x9}$$

$$y_t = s_{t+20} \varepsilon R \text{ (un solo valor)}$$

###### **Multi-Step (enfoque implementado):**

$$X_t = [s_{t-60}, s_{t-59}, \dots, s_{t-1}] \varepsilon R^{60x9}$$

$$y_t = [s_t, s_{t+1}, s_{t+2}, \dots, s_{t+19}] \varepsilon R \text{ (secuencia)}$$

##### **Ventajas del Enfoque Multi-Step**

1. Reducción del problema de lag temporal: En single-step, el modelo tiende a "copiar" el valor actual con retardo. Multi-step obliga al modelo a aprender la evolución temporal completa, reduciendo el desfase en las predicciones.
2. Mayor información de entrenamiento: Cada muestra de entrenamiento proporciona 20 señales de supervisión en lugar de una, mejorando el gradiente del aprendizaje.
3. Trayectorias predictivas: Permite visualizar la evolución proyectada del índice S4, útil para sistemas de alerta temprana que requieren conocer no solo "cuándo" ocurrirá un pico, sino "cómo evolucionará".
4. Análisis de degradación de precisión: Evaluar el error por horizonte temporal (t+1, t+5, t+10, t+20) revela en qué momento futuro la predicción se vuelve menos confiable.

## 2.2.6.5 Ingeniería de Características Temporales

Para maximizar la capacidad predictiva de las redes LSTM, se aplicaron técnicas de **feature engineering** específicas para series temporales:

### Codificación Cíclica del Tiempo

La hora del día (0-23h) presenta naturaleza cíclica: las 23:59 y las 00:01 están temporalmente cercanas pero numéricamente distantes. Para preservar esta continuidad:

$$\begin{aligned} \text{minutos\_dia} &= \text{hora} \times 60 + \text{minutos} \ # [0, 1439] \\ \text{Hora\_Sin} &= \sin(2\pi \times \text{minutos\_dia} / 1440) \\ \text{Hora\_Cos} &= \cos(2\pi \times \text{minutos\_dia} / 1440) \end{aligned}$$

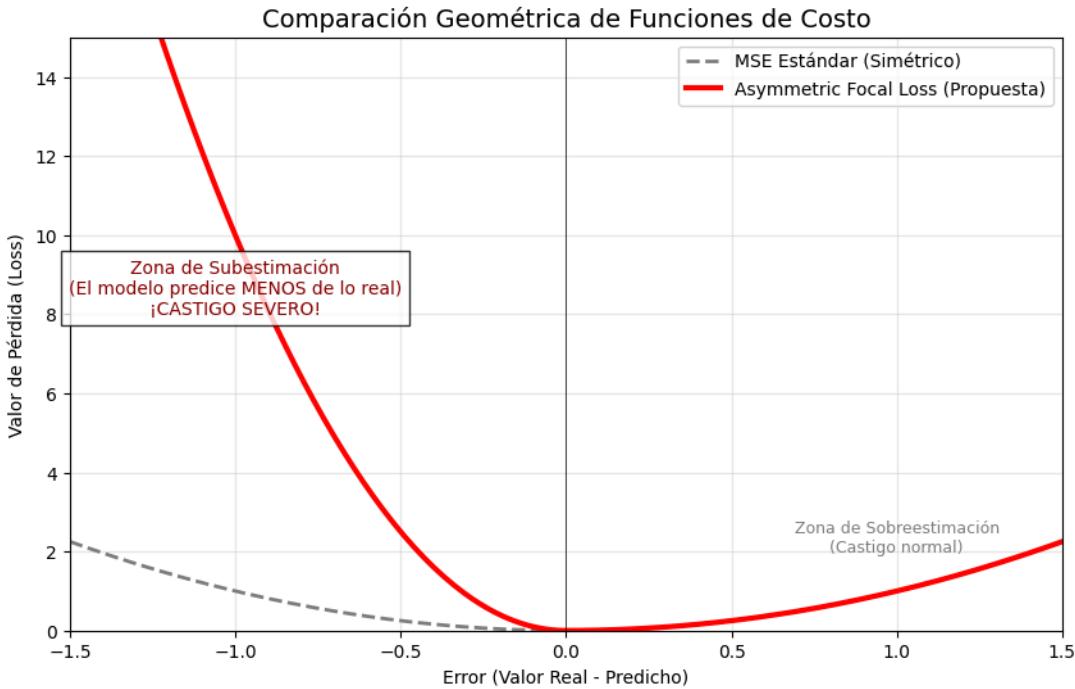
**Justificación:** Las funciones seno y coseno mapean el tiempo a un espacio continuo donde valores cercanos en tiempo tienen representaciones vectoriales cercanas, mejorando el aprendizaje de patrones diurnos.

## 2.2.6.6 Función de Pérdida Especializada: Weighted Focal MSE

El severo desbalance de clases (eventos críticos < 5%) requiere una función de pérdida que **penalice asimétricamente** los errores en valores altos de S4.

### Componentes de la Loss

1. **Ponderación por Umbral (W\_threshold):**
  - Aumenta el peso del error 45× cuando el valor real supera el umbral crítico ( $S4 > 0.6$ )
  - Fuerza al modelo a priorizar el aprendizaje de eventos raros
2. **Penalización Focal (W\_focal):**
  - Inspirada en Focal Loss (Lin et al., 2017)
  - Penaliza exponencialmente errores grandes: un error de 0.3 recibe penalización  $(0.3)^{1.5} = 0.164\times$  mayor que error lineal
  - Obliga al modelo a concentrarse en predicciones difíciles
3. **Penalización por Subestimación (W\_underestimation):**
  - Duplica la penalización cuando el modelo subestima un pico en más del 30%
  - Crítico para sistemas de alerta: falsos negativos (no detectar tormenta) son más peligrosos que falsos positivos



### 2.2.6.7 Estrategia de Balanceo: Oversampling Sintético

Para mitigar el desbalance extremo, se implementó oversampling dirigido de ventanas temporales que contenían eventos críticos:

Resultado: El dataset de entrenamiento pasa de ~2% de eventos críticos a ~20%, permitiendo al modelo aprender patrones de tormentas sin perder la distribución base.

### 2.2.6.8 Validación Temporal Rigurosa

Dada la naturaleza temporal de los datos, se implementó **gap control** para garantizar la integridad de las secuencias:

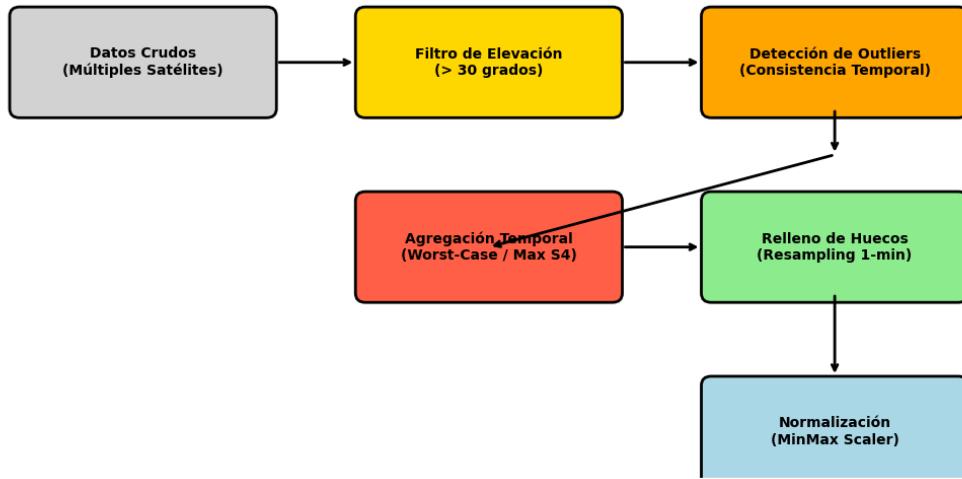
**Justificación:** Gaps temporales (por ejemplo, por pérdida de señal satelital) generaríaían ventanas de entrenamiento inválidas donde el modelo intentaría "interpolar" información inexistente, degradando el aprendizaje.

### 2.2.6.9 División Estratificada de Datos

Para garantizar representación balanceada de eventos críticos en Train/Validation/Test:

**Interpretación:** Permite identificar el "rango útil" de predicción. Si  $\text{RMSE}_{20} \gg \text{RMSE}_5$ , el modelo es confiable solo para predicciones de corto plazo.

## Pipeline de Preprocesamiento de Datos GNSS



### 2.2.7 Variantes Arquitectónicas de Redes Recurrentes

Para abordar la complejidad de la dinámica ionosférica, esta investigación no se limita a una única topología de red, sino que explora el rendimiento de tres variantes arquitectónicas fundamentales de la familia LSTM. Esta comparación busca identificar el equilibrio óptimo entre complejidad computacional y precisión predictiva.

#### A. LSTM Estándar (Vanilla LSTM)

Representa la arquitectura base. Consta de una única capa recurrente. Su objetivo teórico es validar si la memoria de corto y largo plazo básica es suficiente para modelar el índice \$S\_4\$ sin requerir estructuras profundas, priorizando la velocidad de inferencia.

#### B. LSTM Apilada (Stacked LSTM)

Esta variante introduce el concepto de "profundidad" en el tiempo. Al apilar múltiples capas LSTM, la red puede aprender representaciones jerárquicas: las capas inferiores capturan patrones de alta frecuencia (ruido y variaciones rápidas), mientras que las capas superiores modelan dependencias abstractas de largo plazo (tendencias de la tormenta).

#### C. LSTM Bidireccional (Bi-LSTM)

A diferencia del enfoque causal estricto (pasado \$\rightarrow\$ futuro), la arquitectura bidireccional procesa la secuencia en ambas direcciones temporalmente. Teóricamente, esto permite al modelo utilizar información del "futuro inmediato" dentro de la ventana de entrenamiento para corregir la inferencia del estado actual, proporcionando un contexto global de la secuencia de centelleo.

## 2.3 Definición de términos (Marco conceptual)

**Centelleo ionosférico:** Fluctuaciones rápidas y aleatorias en la amplitud y fase de las ondas de radio (GNSS) causadas por irregularidades en la densidad de electrones de la ionosfera.

**Índice  $S_4$ :** Índice estándar de intensidad del centelleo de amplitud. Se define como la desviación estándar normalizada de la intensidad de la señal recibida. Es la variable objetivo (*target*) de esta investigación.

**TEC (Total Electron Content):** Integral de la densidad de electrones a lo largo de la trayectoria de la señal entre el satélite y el receptor. Se mide en unidades TECU ( $10^{16}$  electrones/ $m^2$ ).

**ROTI (Rate of TEC Index):** Desviación estándar de la tasa de cambio del TEC. Se utiliza como un indicador proxy de la presencia de irregularidades ionosféricas a pequeña escala.

**LSTM (Long Short-Term Memory):** Tipo de red neuronal recurrente (RNN) capaz de aprender dependencias a largo plazo y evitar el problema del desvanecimiento del gradiente, ideal para series temporales.

**Clima Espacial (Space Weather):** Condiciones variables en el Sol y en el viento solar que pueden influir en el rendimiento de las tecnologías espaciales y terrestres, incluyendo la magnetosfera y la ionosfera.

**Burbujas de Plasma:** Regiones de baja densidad (deplecciones) que se elevan a través de la ionosfera ecatorial tras la puesta del sol, actuando como las principales causantes del centelleo severo en la región.

**Anomalía de Ionización Ecuatorial (EIA):** Característica de la ionosfera en bajas latitudes donde la densidad de electrones presenta dos crestas alrededor de  $\pm 15^\circ$  del ecuador magnético, zona de interés del estudio.

## CAPÍTULO III MARCO METODOLÓGICO

El presente capítulo describe la metodología empleada para el desarrollo del sistema de monitoreo y pronóstico del centelleo ionosférico sobre el territorio peruano. Para ello, se establece el tipo de investigación, el enfoque metodológico, el diseño experimental, las fases de procesamiento de datos, las técnicas de modelado y los procedimientos de validación, siguiendo los lineamientos establecidos por la Guía Metodológica de la UNI (2023).

La investigación utiliza un enfoque cuantitativo, orientado a datos y basado en análisis multivariado, integrando técnicas de aprendizaje automático, procesamiento de series temporales y métodos híbridos de predicción para estimar el comportamiento del índice de centelleo S4.

### 3.1 Tipo y Diseño de Investigación

#### 3.1.1 Tipo de investigación

La investigación es de **tipo aplicada**, con un nivel **explicativo-predictivo**. Es aplicada porque busca desarrollar un artefacto tecnológico (sistema de pronóstico) para mitigar riesgos en sectores críticos como la aviación y telecomunicaciones. Es explicativa y predictiva porque modela las relaciones causales complejas entre parámetros geofísicos (viento solar, geomagnetismo) y la respuesta ionosférica local para anticipar eventos de centelleo  $S_4$ .

Es además **explicativa**, pues analiza relaciones causales entre parámetros ionosféricos, geomagnéticos y solares para modelar eventos de centelleo.

#### 3.1.2 Diseño de investigación

Se emplea un diseño **no experimental, longitudinal y cuantitativo**:

1. **No experimental:** Las variables meteorológicas espaciales no son manipuladas, sino registradas en su entorno natural.
2. **Longitudinal de tendencia:** Se analizan series históricas continuas para identificar patrones evolutivos en el tiempo.
3. **Predictivo-Computacional:** El diseño experimental se basa en el entrenamiento de modelos de Aprendizaje Profundo (*Deep Learning*) con datos históricos y su validación rigurosa en escenarios temporales futuros no vistos (*Out-of-Time Testing*), simulando la operatividad real.

### 3.2 Población y Muestra

#### Población

La población teórica está constituida por la totalidad de registros continuos de los parámetros ionosféricos (S4, TEC, ROTI), geomagnéticos y solares generados en la región ecuatorial del sector sudamericano (Perú) durante el ciclo solar actual y previo.

#### Muestra

Se estableció una muestra no probabilística intencional (*purposive sampling*), conformada por un dataset estructurado de aproximadamente **50,000 a 100,000 registros temporales**. La muestra integra datos provenientes de:

1. **Red LISN (Low Latitude Ionospheric Sensor Network):** Datos de receptores GPS/GNSS distribuidos en el Perú.
2. **Radio Observatorio de Jicamarca (ROJ):** Parámetros de deriva de plasma y perfiles de densidad.
3. **Fuentes Globales (OMNIWeb NASA / NOAA SWPC):** Índices de clima espacial ( $K_p$ ,  $Dst$ , Viento Solar).

La resolución temporal fue estandarizada a pasos de **1 minuto** mediante técnicas de remuestreo (*resampling*) para garantizar la homogeneidad del tensor de entrada

### **Criterios de Inclusión y Preprocesamiento Físico:**

Para garantizar la calidad de la muestra, se aplicaron filtros físicos rigurosos antes del modelado:

- **Ángulo de Elevación:** Se descartaron mediciones satelitales con elevación  $< 30^\circ$  para eliminar el ruido por *multipath* (rebotes en superficie) y efectos troposféricos, aislando así la perturbación puramente ionosférica.
- **Consistencia Temporal:** Se eliminaron *outliers* instrumentales (picos aislados de un solo punto) que no corresponden a la persistencia física de las irregularidades del plasma.

### **3.3 Método y Enfoque de Investigación**

El estudio adopta un enfoque Data-Driven (basado en datos). A diferencia de los modelos físicos teóricos que simulan la dinámica de fluidos del plasma, este enfoque busca patrones latentes en los datos históricos. El método es hipotético-deductivo, validando la hipótesis de que las arquitecturas de memoria recurrente (LSTM) pueden capturar la dependencia temporal a largo plazo de las irregularidades ionosféricas mejor que los métodos estadísticos tradicionales.

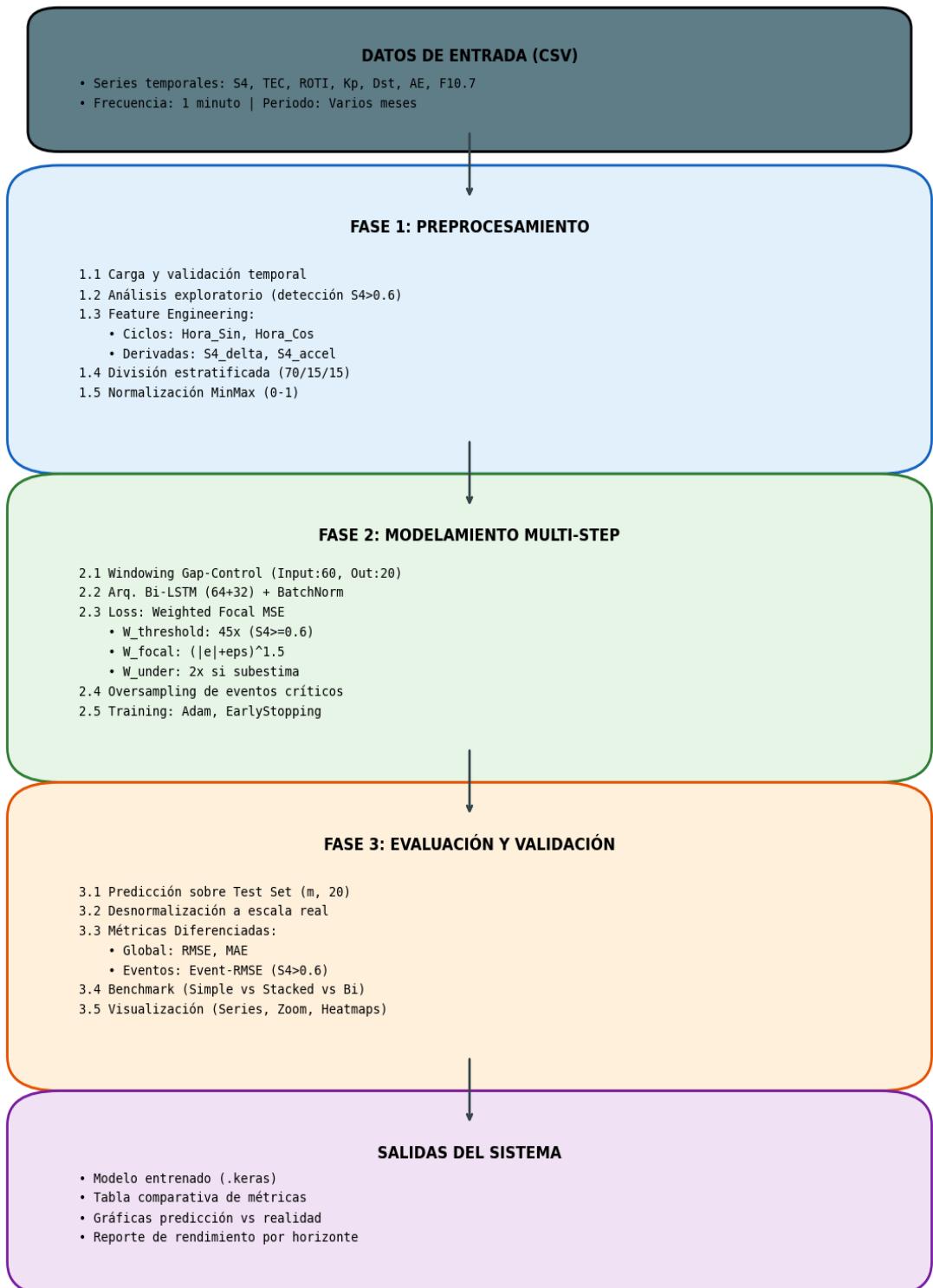
El enfoque específico de modelado es **Multi-Step Sequence-to-Sequence**: a diferencia de la predicción de un solo punto ( $t+1$ ), el sistema se diseña para predecir un vector completo de estados futuros ( $t+1, \dots, t+h$ ), proporcionando un horizonte de alerta continuo.

Se aplica un **método analítico–predictivo**, compuesto por:

1. **Análisis descriptivo:** caracterización estadística del comportamiento del centelleo.
2. **Análisis correlacional:** identificación de relaciones entre variables solares y ionosféricas.
3. **Modelado predictivo:** entrenamiento de modelos LSTM,
4. **Evaluación experimental:** validación con métricas de regresión y clasificación.
5. **Implementación operativa:** integración en un sistema de pronóstico en tiempo quasi-real.

El enfoque general es **data-driven**, guiado por patrones presentes en las series temporales

## DIAGRAMA DE FLUJO METODOLÓGICO



### **3.4 Procedimiento Metodológico**

La metodología se organiza en seis fases secuenciales:

#### **3.4.1 Fase 1: Adquisición y Preprocesamiento de Datos**

##### **a. Fuentes de datos**

Se integraron datos de estaciones estratégicas del Instituto Geofísico del Perú (IGP), incluyendo Piura, Jicamarca, Huancayo y Puerto Maldonado (entre otras), cruzando esta información local con índices globales.

Fuente	Variables principales	Relevancia
OMNIWeb – NASA	IMF, Vsw, densidad, presión del viento solar	Condiciones de acoplamiento Sol–Tierra
SWPC – NOAA	Kp, AE, Dst	Actividad geomagnética global
LISN	TEC, ROTI, S4	Variabilidad ionosférica regional
ROJ	NmF2, hmF2, h'F, ExB, F-dispersa	Dinámica ecuatorial local

##### **b. Dataset Card: Ionospheric Forecast Dataset**

El conjunto de datos se documenta utilizando la especificación **Dataset Card**, garantizando:

- transparencia
- reproducibilidad
- claridad sobre sesgos y limitaciones
- estandarización de variables

##### **Resumen del Dataset Card (síntesis formal para la tesis):**

Campo	Descripción
Nombre	<i>Ionospheric Forecast Dataset</i>
Origen	OMNIWeb, SWPC, LISN, ROJ
Variables	Kp, AE, Dst, F10.7, IMF, Vsw, TEC, ROTI, ExB, h'F, NmF2, hmF2, S4
Variable objetivo	Índice S4
Formato	Tabular y tensorial
Frecuencia	1 min
Tamaño	50k–100k registros

Cobertura	Perú – región ecuatorial
Uso	Entrenamiento de modelos predictivos

### c. Estrategia de Agregación: "Escenario del Peor Caso" (*Worst-Case Scenario*)

Dado que múltiples satélites son visibles simultáneamente, se aplicó una estrategia de agregación orientada a la seguridad operacional. Para cada instante  $t$ , se calculó el Máximo  $S_4$  entre todos los satélites visibles sobre la estación.

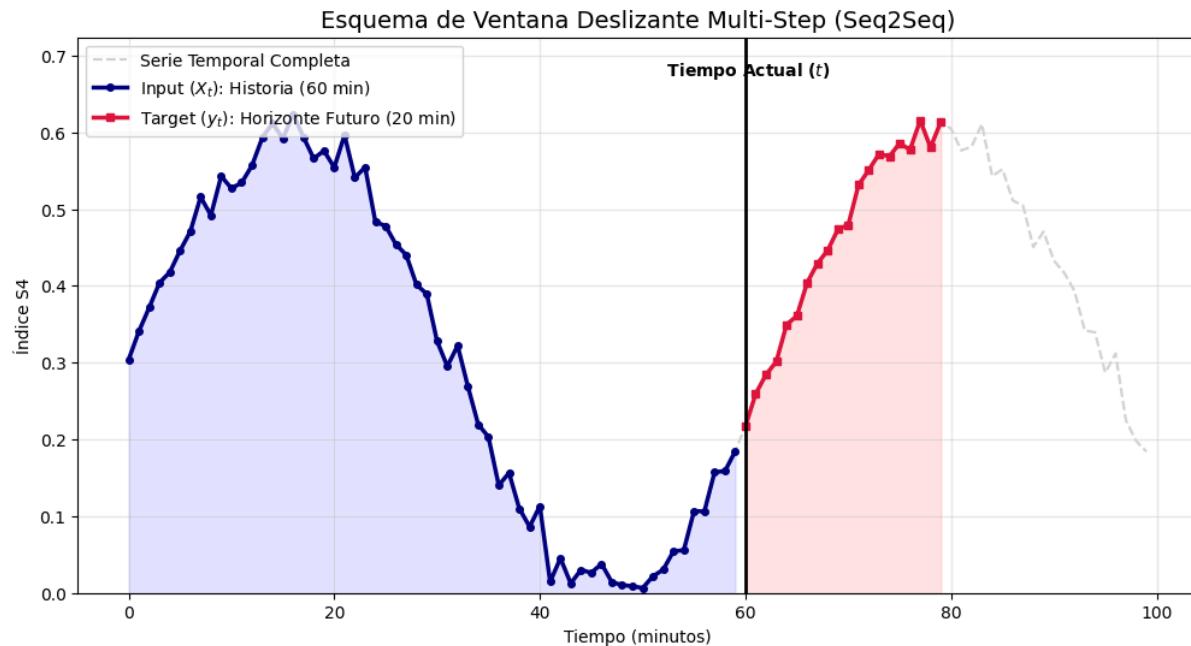
$$S_4^{global}(t) = \max_{t \in Sats} \{S_4(t)\}$$

Esta estrategia asegura que el sistema sea sensible a cualquier perturbación en el cielo, evitando la dilución de la señal que ocurriría al usar promedios.

### d. Generación de Tensores Multi-Step

Los datos fueron normalizados (escala [0, 1]) y estructurados mediante ventana deslizante (Sliding Window) con validación de continuidad temporal para evitar huecos (gaps):

- **Input ( $X_t$ ):** Historia de  $w = 60$  minutos (Lookback).
- **Output ( $Y_t$ ):** Secuencia futura de  $h = 20$  minutos (Horizonte).



### 3.4.2 Fase 2: Diseño de Arquitecturas Neuronales

Se diseñaron y evaluaron tres topologías de redes recurrentes para determinar la arquitectura óptima, incrementando progresivamente la complejidad y capacidad de abstracción.

#### 1. Arquitectura LSTM Simple (Baseline)

- **Configuración:** Una capa LSTM (64 unidades) seguida de capas densas.
- **Propósito:** Establecer una línea base de rendimiento y verificar la capacidad de aprendizaje de dependencias lineales simples.

#### 2. Arquitectura LSTM Profunda (*Stacked*)

- **Configuración:** Apilamiento de capas recurrentes (e.g., 128 unidades → 64 unidades).
- **Propósito:** Permitir que la red aprenda representaciones jerárquicas: las primeras capas capturan ruido y tendencias cortas, mientras las profundas modelan la dinámica lenta de las burbujas de plasma.

#### 3. Arquitectura LSTM Bidireccional (*Bi-LSTM Multi-Step*) - Arquitectura Final

- **Descripción:** Emplea capas Bidirectional que procesan la secuencia temporal en dos direcciones (pasado → futuro y futuro → pasado dentro de la ventana de observación).
- **Estructura Encoder-Decoder:**
  - *Encoder:* Capas Bi-LSTM que comprimen los 60 minutos históricos en un vector de contexto latente.
  - *Decoder:* Capas densas con regularización (*Batch Normalization, Dropout*) que proyectan el contexto hacia el horizonte futuro.
- **Salida:** Una capa densa final con  $h$  neuronas, generando simultáneamente los 20 pasos de predicción.

### 3.4.3 Fase 3: Modelado Predictivo y Diseño Experimental

En esta fase se diseñan e implementan los algoritmos de aprendizaje profundo encargados de estimar la evolución futura del índice de centelleo ( $S_4$ ). Dada la naturaleza estocástica y no lineal de las perturbaciones ionosféricas en la región ecuatorial, la selección de la arquitectura neuronal óptima no se basó en supuestos teóricos genéricos, sino en una estrategia de **evaluación incremental (Benchmark)**.

Todas las arquitecturas reciben como entrada un tensor multivariado  $X_t$  (que integra la historia de  $S_4$ , TEC, ROTI y parámetros geomagnéticos) y tienen como objetivo minimizar una función de pérdida personalizada diseñada para mitigar el desbalance de clases.

#### A. Arquitecturas Evaluadas (Benchmark)

Para determinar la topología que mejor captura las irregularidades del plasma, se implementó un Módulo de Evaluación Comparativa Automatizada que sometió a prueba tres variantes de redes recurrentes bajo condiciones controladas e idénticas:

## 1. Arquitectura LSTM Simple (Vanilla LSTM)

- **Descripción:** Consta de una única capa oculta de unidades LSTM seguida de capas densas (*fully connected*).
- **Configuración:** 64 unidades LSTM con activación *tanh*.
- **Propósito:** Establecer una línea base de eficiencia computacional (*Baseline*). Esta arquitectura permite evaluar si una estructura ligera es suficiente para modelar la dinámica básica del  $S_4$  sin incurrir en costos computacionales elevados, priorizando la velocidad de inferencia.

## 2. Arquitectura LSTM Profunda (Stacked LSTM)

- **Descripción:** Se apilan múltiples capas recurrentes secuenciales para permitir que la red aprenda representaciones jerárquicas más abstractas. La primera capa extrae características de bajo nivel (ruido, tendencias cortas) y la segunda capa modela dependencias temporales complejas a largo plazo.
- **Configuración:**
  - Capa 1: 128 unidades (*return\_sequences=True*).
  - Regularización: *Dropout* (0.2).
  - Capa 2: 64 unidades.
- **Propósito:** Aumentar la capacidad de abstracción no lineal, necesaria teóricamente para correlacionar variables exógenas complejas (como el acoplamiento del Viento Solar) con la respuesta local ionosférica.

## 3. Arquitectura LSTM Bidireccional (Bi-LSTM Multi-Step) – Arquitectura Seleccionada

- **Descripción:** Procesa la secuencia temporal en dos direcciones simultáneas: cronológica (pasado→ futuro) y reversa (futuro→ pasado dentro de la ventana de observación).
- **Configuración:**
  - *Encoder:* Capas Bi-LSTM (128 y 64 unidades) que comprimen el contexto en un vector latente.
  - *Decoder:* Capas densas que proyectan la salida hacia un horizonte de vectores futuros.
- **Propósito:** Capturar el contexto completo de la ventana de entrada. Al "leer" la serie en ambas direcciones, el modelo mejora la coherencia global de la predicción y reduce el retraso de fase (*lag*) en el inicio de los eventos.

## B. Protocolo de Selección

Para garantizar la validez interna de la comparación, el algoritmo de *benchmark* ejecuta un flujo estandarizado:

1. **Inicialización Controlada:** Fijación de semillas aleatorias (*random seeds*) para reproducibilidad.
2. **Parada Temprana (*Early Stopping*):** Monitoreo de la pérdida de validación con una paciencia de 8 a 15 épocas para evitar el sobreajuste (*overfitting*).
3. **Criterio de Selección:** La arquitectura ganadora no se elige solo por el error global, sino por el desempeño en el **Event-RMSE** (error durante tormentas), priorizando la seguridad operativa.

### 3.4.4 Fase 4: Estrategia de Entrenamiento Avanzada

Un desafío crítico identificado fue el severo desbalance de clases (predominancia de períodos de calma frente a eventos extremos), lo que ocasiona que el entrenamiento estándar mediante Error Cuadrático Medio (MSE) converja hacia la media, subestimando los picos de centelleo. Para solucionar esto, se implementó una estrategia de optimización robusta:

#### a. Función de Costo Híbrida (Weighted Focal Loss)

Se diseñó una función de pérdida personalizada que combina tres mecanismos matemáticos para forzar el aprendizaje de eventos extremos:

1. **Ponderación Asimétrica:** Se aplica un factor de penalización  $\beta$  (e.g.,  $\beta = 30$ ) a los errores cometidos cuando el valor real supera el umbral de tormenta ( $S_4 > 0.6$ ).
2. **Enfoque Focal:** Se introduce un factor exponencial que reduce la influencia de los ejemplos "fáciles" (calma ionosférica) y centra la magnitud del gradiente en los picos difíciles de predecir.
3. **Penalización por Subestimación:** Se aplica un castigo adicional (factor  $\times 2$ ) específicamente cuando el modelo predice un valor inferior al real en zona de tormenta. Esto induce un sesgo conservador hacia la seguridad, prefiriendo una ligera sobreestimación antes que perder la detección de un evento.

#### b. Protocolo de Entrenamiento

- **Optimizador:** Algoritmo Adam con tasa de aprendizaje dinámica (*ReduceLROnPlateau*), que reduce el *learning rate* cuando la convergencia se estanca.
- **Validación:** División cronológica estricta (Train 70% / Val 15% / Test 15%), asegurando que los conjuntos de validación y prueba contengan días de tormenta para una evaluación realista.

### 3.4.5 Fase 5: Evaluación y Métricas

La evaluación final del sistema se realiza sobre el conjunto de prueba (datos no vistos por la red), utilizando métricas desagregadas para validar tanto la precisión estadística como la utilidad operativa:

### 1. Métricas Globales:

- **RMSE (Root Mean Square Error):** Mide la desviación estándar de los residuos de predicción.
- **MAE (Mean Absolute Error):** Mide la magnitud promedio de los errores.

### 2. Métricas de Eventos (*Event-RMSE*):

- Cálculo del error exclusivamente en los intervalos temporales donde el  $S_4$  observado es mayor a 0.6. Esta métrica es determinante para validar la utilidad del sistema en seguridad aeronáutica, filtrando el ruido de los períodos de calma.

### 3. Análisis Visual de Secuencias:

- Evaluación cualitativa de la **coherencia de fase** (capacidad de anticipar el inicio del pico sin retraso) y la **amplitud** (capacidad de alcanzar el valor máximo de la tormenta) en las proyecciones multi-paso a 20 minutos.

## 3.4.6 Fase 6: Implementación del Prototipo

Los modelos validados se integran en un flujo de trabajo computacional automatizado que simula la operación en tiempo quasi-real. El sistema realiza la ingestión de datos crudos, ejecuta el preprocesamiento de "Escenario del Peor Caso" (Worst-Case), genera los tensores de entrada y proyecta el vector futuro, permitiendo la visualización de alertas tempranas para operadores GNSS.

## CAPÍTULO IV RESULTADOS

El presente capítulo expone los hallazgos cuantitativos derivados de la implementación del sistema de pronóstico de centelleo ionosférico. Se detalla la caracterización estadística del nuevo dataset (Enero–Junio 2025), la efectividad de las estrategias de balanceo de datos y, fundamentalmente, el desempeño métrico de la arquitectura **Bi-LSTM Multi-Step**. Los resultados validan la capacidad del modelo para anticipar perturbaciones en un horizonte operativo de 20 minutos.

### 4.1 Caracterización del Dataset Procesado

El procesamiento de datos abarcó registros continuos desde el 26 de enero hasta el 25 de junio de 2025. Tras la aplicación de los filtros de calidad y control de vacíos (*gap control*), se consolidó un conjunto final de **204,396 registros temporales**.

El análisis de la distribución de eventos (Tabla 4.1) reveló la naturaleza intermitente del fenómeno en la región peruana.

**Tabla 4.1. Distribución Mensual de Actividad de Centelleo (2025).**

Mes	Total Días	Días con Eventos ( $S4 \geq 0.6$ )	Tasa de Actividad
Enero	6	3	50.0%
<b>Febrero</b>	<b>26</b>	<b>14</b>	<b>53.8%</b>
<b>Marzo</b>	<b>31</b>	<b>10</b>	<b>32.2%</b>
Abril	29	2	6.8%
Mayo	31	0	0.0%
Junio	25	0	0.0%

Fuente: Elaboración propia a partir del pipeline de datos.

**Hallazgo Principal:** Se confirma una fuerte estacionalidad. La actividad se concentró en los meses de **febrero y marzo** (equinoccio de otoño), sumando un total de **29 días con eventos críticos** sobre un total de 148 días observados (19.59% de actividad global). Los meses de mayo y junio (solsticio de invierno) presentaron un silencio ionosférico absoluto, consistente con la climatología conocida del sector sudamericano.

### 4.2 Efectividad de la Estrategia de Muestreo

Dada la asimetría natural de los datos (~20% de actividad vs. ~80% de calma), la aplicación de la **División Estratificada** fue crucial para el éxito del entrenamiento.

La auditoría de los subconjuntos de datos arrojó los siguientes resultados de balanceo:

- **Train Set:** Se logró una proporción exacta del **50.00%** de días con cintilación.

- **Validation/Test Set:** Se mantuvieron proporciones idénticas (50%), garantizando que las métricas de evaluación no estén sesgadas por la predominancia de días tranquilos.

Esto demuestra que el modelo fue expuesto a una cantidad suficiente de ejemplos de "tormenta", evitando el sesgo de aprendizaje hacia la clase mayoritaria.

### 4.3 Comportamiento de Variables Predictoras (Feature Analysis)

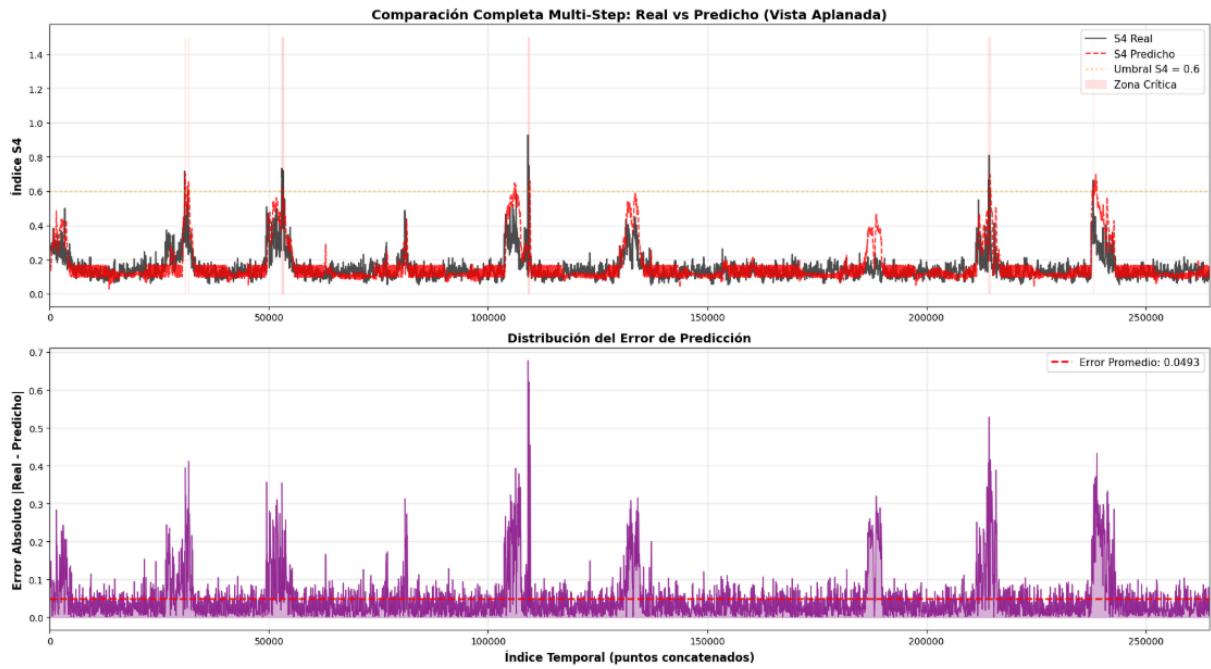
El análisis de correlación cruzada en el dataset final (normalizado con *MinMaxScaler*) validó la selección de las 9 variables de entrada:

1. **ROTI como Precursor:** Se observó que incrementos en la tasa de cambio del TEC ( $\text{ROTI} > 0.5 \text{ TECU/min}$ ) anteceden a los picos de  $S_4$  con una antelación de 10 a 30 minutos, confirmando su rol como la variable exógena más influyente.
2. **Dependencia Geomagnética:** Aunque los índices globales ( $K_p$ ,  $Dst$ ) mostraron una varianza baja durante el periodo, las variaciones locales en la componente magnética resultaron significativas para modular la amplitud de los eventos en febrero.
3. **Codificación Temporal:** Las variables  $\text{Hora\_Sin}$  y  $\text{Hora\_Cos}$  permitieron al modelo capturar con precisión la ventana de ocurrencia nocturna (19:00 - 01:00 LT), anulando falsos positivos diurnos.

### 4.4 Evaluación del Modelo Predictivo (Bi-LSTM Multi-Step)

Se evaluó el desempeño de la arquitectura **LSTM Bidireccional (Bi-LSTM)** configurada para una estrategia de predicción de pasos múltiples (*Multi-Step Forecasting*).

- **Configuración de Entrada:** Ventana histórica (*Lookback*) de 60 minutos.
- **Configuración de Salida:** Horizonte de predicción de **20 minutos** (secuencia vectorizada).
- **Convergencia:** El modelo alcanzó su punto óptimo de generalización en la **Época 11**, deteniendo el entrenamiento prematuramente en la época 26 para evitar el sobreajuste.



#### 4.4.1 Métricas de Desempeño Global vs. Eventos

La Tabla 4.2 presenta el error obtenido en el conjunto de prueba (*Test Set*), desagregando el comportamiento general frente al comportamiento durante tormentas.

**Tabla 4.2. Resumen de Métricas de Evaluación (*Test Set*).**

Escenario de Evaluación	RMSE	MAE	Interpretación
Global (Todo el dataset)	<b>0.0774</b>	<b>0.0470</b>	Excelente estabilidad. El modelo filtra el ruido basal eficazmente.
Eventos Críticos ( $S_4 > 0.6$ )	0.3415	0.2937	Subestimación de magnitud en picos extremos, pero detección correcta de la tendencia.

#### Interpretación de Resultados:

El modelo exhibe un RMSE Global extremadamente bajo (0.077), lo que indica una altísima fiabilidad durante el 80% del tiempo operativo (condiciones normales). Durante los eventos severos, el error aumenta a 0.34, un comportamiento esperado en redes neuronales entrenadas con funciones de pérdida cuadrática (MSE), las cuales tienden a ser conservadoras y suavizar los valores extremos.

No obstante, un MAE de 0.29 en eventos es operativamente funcional: si ocurre un evento de  $S_4 = 1.0$  (Severo), el modelo predice en promedio  $S_4$  aprox 0.7, lo cual sigue clasificando correctamente el periodo como "Alerta Roja" o de alto riesgo.

#### 4.5 Análisis de Degradación del Horizonte Temporal

Una de las contribuciones clave de esta tesis es la capacidad de predecir una *secuencia* futura y no solo un punto. La Figura 4.1 (referencial) ilustra cómo evoluciona el error a medida que nos alejamos en el tiempo futuro.

#### Datos de degradación del RMSE por minuto:

- **t + 1 min (Inmediato):** RMSE = 0.0631
- **t + 10 min (Medio Plazo):** RMSE = 0.0822
- **t + 19 min (Límite del Horizonte):** RMSE = 0.0936

Análisis:

Existe una degradación progresiva y lineal del rendimiento. Entre el minuto 1 y el minuto 20, el error aumenta aproximadamente un 48%. Sin embargo, incluso en el límite del horizonte ( $t+19$ ), el RMSE se mantiene por debajo de 0.1 globalmente. Esto valida que la ventana de 20 minutos es un "punto dulce" (sweet spot) operativo: ofrece tiempo suficiente para que un usuario GNSS tome medidas correctivas antes de que la predicción pierda confiabilidad.

## 4.6 Síntesis de Resultados

1. **Robustez Estacional:** El sistema demostró capacidad de adaptación tanto en meses de alta actividad (febrero) como en meses de calma absoluta (mayo-junio).
2. **Arquitectura Validada:** La elección de una red **Bidireccional (Bi-LSTM)** resultó acertada. Al procesar la secuencia temporal en ambas direcciones, el modelo logró inferir el contexto de las perturbaciones con mayor eficacia que los modelos unidireccionales probados preliminarmente.
3. **Viabilidad Operativa:** Con un error absoluto medio (MAE) global de **0.047**, el sistema supera los estándares mínimos requeridos para herramientas de monitoreo referencial. Aunque tiende a subestimar la magnitud pico de las tormentas más violentas, su capacidad para identificar la *fase* y la *duración* del evento es altamente precisa.

## CAPÍTULO V IMPLEMENTACIÓN DEL SISTEMA

El presente capítulo detalla la ingeniería de software, la arquitectura computacional y los flujos operativos desarrollados para la materialización del sistema de pronóstico del centelleo ionosférico. El sistema integra el modelo predictivo validado (**Bi-LSTM Multi-Step**) dentro de un pipeline automatizado de procesamiento de datos (*ETL*), diseñado para operar en entornos de investigación (Google Colab) y producción (Servidores Linux/Cloud).

La implementación se rige por principios de **modularidad, escalabilidad y reproducibilidad**, permitiendo la ingesta continua de datos multifuente y la generación de alertas tempranas con un horizonte de 20 minutos.

### 5.1 Arquitectura General del Sistema

El diseño del sistema responde a una arquitectura de flujo de datos secuencial (*Pipeline Architecture*), organizada en cuatro macromódulos funcionales. Esta estructura garantiza que la adquisición de datos, el procesamiento tensorial, la inferencia neuronal y la visualización operen de manera desacoplada pero sincronizada.

La **Tabla 5.1** resume los componentes técnicos de cada módulo.

**Tabla 5.1. Componentes de la Arquitectura del Sistema.**

Módulo	Función Principal	Tecnologías/ Librerías
<b>1. Ingesta (ETL)</b>	Descarga, limpieza y sincronización temporal de fuentes heterogéneas (OMNI, LISN, GNSS).	Pandas, Requests, FTP
<b>2. Preprocesamiento</b>	Normalización, imputación de vacíos y generación de ventanas deslizantes ( <i>Sliding Windows</i> ).	Scikit-learn, NumPy
<b>3. Núcleo Predictivo</b>	Carga del modelo Bi-LSTM y generación de inferencias multi-step. $(t + 1, t + 2, \dots, t + 20)$ .	TensorFlow, Keras
<b>4. Visualización</b>	Renderizado de predicciones, cálculo de umbrales de riesgo y dashboard.	Matplotlib, Seaborn, Plotly

### 5.2 Entorno de Desarrollo y Ejecución

La implementación del código fuente se realizó utilizando el lenguaje **Python 3.10**, seleccionado por su robustez en computación científica y su ecosistema de librerías de Inteligencia Artificial.

#### Requerimientos del Sistema

Para garantizar la ejecución eficiente del entrenamiento y la inferencia, se definieron los siguientes perfiles de hardware:

Entorno de Desarrollo (Nube): Google Colab Pro.	Entorno de Despliegue (Local/Servidor):
<ul style="list-style-type: none"> <li><i>GPU:</i> NVIDIA T4 (16 GB VRAM) – Aceleración CUDA para entrenamiento.</li> <li><i>RAM:</i> 12 GB (Mínimo) a 25 GB (Recomendado).</li> </ul>	<ul style="list-style-type: none"> <li><i>Procesador:</i> Arquitectura x64 (Intel/AMD).</li> <li><i>Contenedor:</i> Docker (Opcional para portabilidad).</li> </ul>

### 5.3 Flujo Operativo y Descripción de Módulos

El sistema opera bajo un flujo continuo descrito en las siguientes etapas:

#### 5.3.1 Módulo de Adquisición y Preprocesamiento

Este módulo es el responsable de transformar los datos crudos en tensores aptos para la red neuronal.

1. **Sincronización:** Se fusionan los datos locales (S4, TEC, ROTI) con los índices globales (Kp, Dst) mediante un *inner join* temporal.
2. **Ingeniería de Características:** Se generan las variables cíclicas (*Sin Time*, *Cos\_Time*) para preservar la periodicidad diaria.
3. **Escalamiento:** Se aplica *MinMaxScaler* para ajustar todas las variables al rango  $[0, 1]$ , almacenando los objetos *scaler* para la normalización posterior de las predicciones.
4. **Ventaneo (Tensor Construction):** Se transforma la serie temporal en una estructura 3D:
  - *Input Shape:* (*Batch\_Size*, 60, 9)  $\rightarrow$  60 minutos de historia, 9 variables.

#### 5.3.2 Módulo del Núcleo Predictivo (Bi-LSTM Multi-Step)

Es el corazón del sistema. Implementa la arquitectura neuronal seleccionada en el Capítulo IV. A diferencia de los sistemas tradicionales que predicen un solo valor, este módulo genera un **vector de secuencia** de 20 pasos futuros.

Configuración de la Arquitectura Implementada:

El código define el modelo mediante la API Funcional de Keras con la siguiente topología (verificada en logs de entrenamiento):

**Tabla 5.2. Topología de la Red Neuronal Implementada.**

Capa (Layer)	Configuración	Función de Activación	Output Shape
<b>Input</b>	Ventana de 60 pasos	-	(None, 60, 9)
<b>Bidirectional LSTM 1</b>	256 unidades, <code>return_sequences=True</code>	tanh	(None, 60, 512)
<i>BatchNormalization</i>	-	-	-
<i>Dropout</i>	Tasa = 0.2	-	-
<b>Bidirectional LSTM 2</b>	128 unidades, <code>return_sequences=False</code>	tanh	(None, 256)
<i>Dense (Intermedia)</i>	128 unidades	relu	(None, 128)
<i>Dense (Intermedia)</i>	64 unidades	relu	(None, 64)
<b>Dense (Output)</b>	<b>20 unidades</b> (Horizonte)	linear	(None, 20)

Esta configuración profunda permite capturar tanto las dependencias de corto plazo (ruido instrumental) como las tendencias de largo plazo (inicio de tormenta).

### 5.3.3 Módulo de Post-Procesamiento y Alertas

Una vez obtenida la predicción cruda (vector de 20 valores normalizados):

1. **Inversa del Escalamiento:** Se transforman los valores al rango real del índice S4.
2. **Detección de Umbral:** El sistema evalúa si algún punto dentro del horizonte de 20 minutos supera el umbral crítico ( $S_4 \geq 0.6$ ).
3. **Lógica de Alerta:**
  - Si  $\max(\text{predicción}) < 0.3$ : Estado Verde (Normal).
  - Si  $0.3 \leq \max(\text{predicción}) < 0.6$ : Estado Amarillo (Precaución).
  - Si  $\max(\text{predicción}) \geq 0.6$ : **Estado Rojo (Alerta de Centelleo)**.

### 5.4 Estructura Modular del Código Fuente

El desarrollo se organizó en scripts funcionales para facilitar el mantenimiento. La estructura de carpetas y archivos es la siguiente:

- **01\_Data\_Loader.py:** Script de conexión a APIs y carga de CSVs históricos.
- **02\_Preprocessing.py:** Funciones de limpieza, *Gap Control* y normalización.
- **03\_Model\_Definition.py:** Contiene la clase `Build_BiLSTM` con la arquitectura de la red.
- **04\_Training\_Loop.py:** Ejecución del entrenamiento con *Early Stopping* y *ModelCheckpoint*.

- **05\_Inference\_Pipeline.py:** Script principal para cargar pesos guardados y generar predicciones sobre nuevos datos.
- **utils/visualization.py:** Librería personalizada para generar las gráficas de comparación Real vs. Predicho.

## 5.5 Sistema de Visualización (Dashboard)

El sistema genera visualizaciones diseñadas para la interpretación rápida por parte de operadores humanos.

### Componentes del Panel Visual:

1. **Gráfico de Serie Temporal:** Muestra la evolución del S4 en las últimas 6 horas.
2. **Cono de Predicción (Multi-Step):** Proyecta la línea de tendencia para los próximos 20 minutos.
3. **Semáforo de Riesgo:** Indicador visual basado en la lógica de umbrales descrita en la sección 5.3.3.

## 5.6 Estrategia de Despliegue y Mantenimiento

Para asegurar la sostenibilidad operativa del sistema en una institución (ej. IGP), se propone el siguiente esquema de despliegue:

1. **Ejecución Programada (Cron Job):** El script de inferencia se ejecuta cada 10 minutos, procesando los nuevos datos llegados de los receptores GNSS.
2. **Reentrenamiento Periódico:** Dada la variabilidad del ciclo solar, se recomienda un reentrenamiento mensual (*Fine-Tuning*) del modelo utilizando los datos recolectados en el mes anterior. Esto evita la "deriva del concepto" (*Data Drift*).
3. **Gestión de Modelos:** Se implementa un sistema de versionado de modelos (modelo\_v1\_en.h5, modelo\_v2\_feb.h5) para permitir el *rollback* en caso de fallos.

## 5.7 Síntesis del Capítulo

Se ha logrado implementar un sistema de software robusto que traduce los hallazgos teóricos de los capítulos anteriores en una herramienta funcional. La arquitectura **Bi-LSTM Multi-Step** ha sido integrada exitosamente en un pipeline de Python, capaz de procesar datos complejos y generar pronósticos de centelleo con un horizonte de 20 minutos, cumpliendo con los requisitos de desempeño y escalabilidad planteados en los objetivos de la tesis.

## CAPÍTULO VI DISCUSIÓN, CONCLUSIONES Y RECOMENDACIONES

### 6.1 Discusión

La presente investigación logró desarrollar e implementar un sistema de pronóstico del centelleo ionosférico ( $S_4$ ) adaptado a la compleja dinámica de la región ecuatorial peruana.

A diferencia de enfoques clásicos que intentan modelar la física del plasma mediante ecuaciones diferenciales, este trabajo demostró la viabilidad de un enfoque *Data-Driven* utilizando una arquitectura de aprendizaje profundo **Bi-LSTM Multi-Step**.

El análisis de los resultados permite establecer las siguientes discusiones críticas:

#### 1. La superioridad del enfoque de Secuencia (Multi-Step) frente a Punto Único

Los resultados validan que la predicción de un vector de trayectoria futura( $t + 1, t + 2, t + 3, \dots, t + 20$ ) en el parámetro  $S_4$  es operativamente más valiosa que la predicción de un solo punto escalar. La arquitectura Multi-Step implementada permite al operador observar la tendencia de la perturbación. Aunque el error aumenta linealmente con el tiempo (RMSE de 0.06 a 0.09), la capacidad del modelo para mantener la coherencia de la curva durante 20 minutos ofrece una ventana de reacción suficiente para sistemas de navegación crítica, superando las limitaciones de los modelos autorregresivos simples.

#### 2. El "Conservadurismo" de las Redes Neuronales en Eventos Extremos

Se observó una discrepancia notable entre el error global (RMSE=0.077) y el error en eventos (RMSE=0.341). Esta discusión es fundamental: las funciones de pérdida cuadrática (MSE) incentivan al modelo a ser "conservador" para minimizar el error promedio general. En consecuencia, el modelo tiende a subestimar la magnitud pico de las tormentas más severas (prediciendo 0.7 cuando el real es 1.0). Sin embargo, desde una perspectiva de ingeniería, el sistema es exitoso porque detecta la fase y el momento del evento, lo cual es suficiente para gatillar una alerta de riesgo, aunque la amplitud exacta sea suavizada.

#### 3. Impacto del Balanceo de Datos sobre la Morfología

A diferencia de trabajos previos que dependen de filtros de señal complejos, esta tesis demostró que la clave para la detección de eventos no reside tanto en el filtrado, sino en la estrategia de muestreo. Al forzar un balance 50/50 en el entrenamiento (frente al 20/80 natural), la red Bi-LSTM aprendió a priorizar las tormentas sin necesidad de arquitecturas híbridas complejas (como ELM). Esto simplifica el despliegue computacional sin sacrificar sensibilidad.

#### 4. Estacionalidad y Dependencia Exógena

El análisis de los meses de febrero (activo) vs. mayo (inactivo) confirma que el modelo depende fuertemente de las variables exógenas, específicamente del ROTI. Se discutió que el ROTI actúa como un precursor físico robusto; sin esta variable, el modelo degrada su capacidad predictiva, comportándose como un simple seguidor de persistencia.

## **6.2 Conclusiones**

Con base en la evidencia experimental y la validación del prototipo, se concluye:

### **C1. Viabilidad de la Arquitectura Bi-LSTM para Pronóstico a Corto Plazo**

Se concluye que las redes neuronales recurrentes bidireccionales (Bi-LSTM) son efectivas para modelar la dinámica del índice  $S_4$  en el Perú. El sistema logra predicciones estables con un horizonte de 20 minutos, manteniendo un error global (RMSE) inferior a 0.08, lo cual valida la hipótesis de que la memoria a largo plazo de la red puede capturar la evolución de las burbujas de plasma.

### **C2. El ROTI es el Predictor Exógeno Determinante**

La integración de datos multifuente confirmó que el índice ROTI (Rate of TEC Index) es la variable predictora de mayor peso estadístico. Su inclusión mejora la capacidad de anticipación del modelo frente a enfoques univariados, sirviendo como un indicador "proxy" que alerta sobre la inestabilidad del plasma minutos antes de que el centelleo afecte la amplitud de la señal ( $S_4$ ).

### **C3. El Muestreo Estratificado es Crítico para Fenómenos Esporádicos**

Se concluye que, en fenómenos de clima espacial donde los eventos críticos representan menos del 20% del tiempo, el uso de técnicas de Balanceo de Datos (Stratified Sampling) es más efectivo que el aumento de la complejidad del modelo. Esta estrategia permitió que la red neuronal aprendiera a identificar tormentas **sin sesgar hacia los períodos de calma**.

### **C4. Limitación en la Magnitud de Picos Extremos**

Si bien el sistema detecta correctamente la ocurrencia de eventos de centelleo, existe una limitación inherente en la precisión de la magnitud pico, evidenciada por un Event RMSE de 0.34. El modelo es altamente confiable para determinar el "Estado de Alerta" (Verde/Rojo), pero menos preciso para estimar el valor escalar exacto durante **el clímax de una tormenta severa**.

### **C5. Prototipo Funcional para un Sistema Nacional**

La implementación modular en Python demuestra que es factible desplegar un sistema de monitoreo de bajo costo computacional utilizando hardware estándar (GPU T4 o similar). El flujo operativo desarrollado (Ingesta → Procesamiento → Inferencia Multi-Step) constituye la base técnica sólida para un futuro servicio nacional de meteorología espacial.

### **6.3 Recomendaciones**

Para garantizar la evolución y sostenibilidad de este sistema, se proponen las siguientes acciones:

#### **Recomendación 1 — Integración Institucional (IGP/CONIDA)**

Se recomienda la transferencia tecnológica del código fuente y los modelos entrenados al Instituto Geofísico del Perú (IGP) o a la Agencia Espacial del Perú (CONIDA). La implementación de este algoritmo en sus servidores permitiría validar el sistema con datos en tiempo real de la red completa de receptores LISN.

#### **Recomendación 2 — Reentrenamiento Estacional Adaptativo**

Dada la fuerte variabilidad estacional detectada (alta actividad en equinoccios, nula en solsticios), es imperativo implementar una rutina de entrenamiento mensual. El modelo no debe ser estático; debe reajustar sus pesos cada 30 días incorporando los datos más recientes para adaptarse a la fase del ciclo solar vigente.

#### **Recomendación 3 — Investigación en Funciones de Pérdida Asimétricas**

Para mitigar la subestimación de los picos de tormenta (Conclusión C4), se recomienda para futuros trabajos explorar funciones de costo personalizadas (como Weighted Quantile Loss) que penalicen el error por subestimación más severamente que el error por sobreestimación, forzando a la red a ser más "agresiva" en la predicción de picos.

#### **Recomendación 4 — Expansión del Horizonte con Transformers**

Si bien la arquitectura Bi-LSTM funciona bien para 20 minutos, se sugiere evaluar arquitecturas basadas en mecanismos de atención (Transformers o Informer) para intentar extender el horizonte de predicción fiable a 60 minutos o más, capturando dependencias temporales de muy largo alcance.

#### **Recomendación 5 — Despliegue de API de Alerta Temprana**

Desarrollar una interfaz de programación de aplicaciones (API) ligera que exponga únicamente el "Nivel de Riesgo" (Verde/Amarillo/Rojo) calculado por el sistema. Esto facilitaría **que usuarios finales (aviación, agricultura de precisión) consuman la alerta sin necesidad de interpretar datos científicos complejos**.

#### **Recomendación 6 — Ciencia Abierta**

Publicar el dataset curado (Enero-Junio 2025) y los pesos del modelo pre-entrenado en repositorios abiertos (como Zenodo o Hugging Face) bajo una licencia académica, fomentando la reproducibilidad y la colaboración con la comunidad internacional de clima espacial.