Sistema de Clasificación Oncológica con Arquitectura Multicapa y Control de Errores: Validación Controlada y Desempeño en Producción Real en el Sistema Público de Salud de Chile

Paulo Villarroel Tapia

2026-01-07

Tabla de contenidos

1 Resumen Ejecutivo: Modernización y Seguridad en la Gestión Oncológica **4**
1.1 Valor Público y Estratégico . . . 4
1.2 Cumplimiento Normativo y Ético . . . 4
1.3 Estado de Madurez y Resultados . . . 4

2 Declaración de Significancia **5**

3 Resumen (Propuesta Final - Primera Versión) **5**

4 Introducción **6**
4.1 Contexto nacional: Desafío de listas de espera quirúrgicas oncológicas . . . 6
4.2 El problema de heterogeneidad en salud pública . . . 6
4.3 Marco teórico: Framework err(iiv) para control de errores . . . 7
4.4 Objetivos del estudio . . . 8
4.5 Enfoque por fases y visión estratégica de largo plazo . . . 8

5 Métodos **10**
5.1 Visión general del sistema: Pipeline completo . . . 10
5.2 Arquitectura multicapa con generación independiente de evidencia . . . 11
5.2.1 Capa 1: Módulo heurístico (Evidencia H(x)) . . . 11
5.2.2 Capa 2: Módulo semántico (Evidencia S(x)) . . . 12
5.2.3 Capa 3: Clasificador XGBoost V4.6 (evidencia M(x)) . . . 13
5.3 Sistema de control de errores err(iiv): Formalización matemática . . . 17
5.3.1 Nota sobre implementación del framework err(iiv) . . . 19
5.4 Lógica de decisión jerárquica: Formalización algorítmica . . . 19
5.5 Diseño conservador para screening poblacional . . . 23
5.6 Definición de ground truth y validación . . . 23
5.7 Datasets . . . 24
5.7.1 Proceso de etiquetado y validación del ground truth . . . 24

6 Resultados **25**
6.1 Desempeño en validación controlada . . . 25
6.1.1 Discriminación extremadamente alta en contexto controlado . . . 25
6.1.2 Métricas de clasificación excepcionales . . . 26
6.1.3 Análisis detallado de errores . . . 26
6.1.4 Impacto operacional de la automatización . . . 28
6.2 Desempeño en producción real: Implementación en 29 servicios de salud . . . 28

6.2.1 Contexto y limitaciones de los datos de producción . . . 28
6.2.2 Métricas agregadas globales de 29 servicios . . . 28
6.2.3 Análisis estratificado: Identificación de heterogeneidad . . . 29
6.2.4 Desempeño en 26 servicios con proceso estandarizado . . . 29
6.2.5 Los 3 servicios con proceso deficiente: Análisis cualitativo . . . 29
6.2.6 Impacto cuantificado de los 3 servicios en métricas agregadas . . . 33
6.2.7 Categorización de servicios . . . 34
6.2.8 Robustez del modelo ante variabilidad . . . 34
6.3 Interpretación de precision moderada . . . 34
6.4 Desempeño reciente: Corte Octubre 2025 . . . 34
6.4.1 Resultados de clasificación (n=6,560) . . . 34
6.4.2 Métricas de desempeño recalculadas . . . 38
6.4.3 Análisis de la mejora . . . 38
6.5 Análisis de Seguridad: El Caso de los "Indeterminados" (Noviembre 2025) . . . 38

7 Discusión **39**
7.1 Validación del framework err(iiv) . . . 39
7.2 Validación empírica de seguridad clínica: Análisis de sesgo de verificación del VPN . . . 39
7.2.1 Contexto del análisis . . . 39
7.2.2 Metodología del análisis . . . 39
7.2.3 Resultados clave . . . 39
7.2.4 Conclusión sobre sesgo de verificación . . . 40
7.3 Validación empírica de arquitectura fail-safe mediante monitoreo PSI . . . 40
7.3.1 Resultados del monitoreo PSI . . . 40
7.3.2 Análisis de causa raíz: Divergencia Distribucional Esperada por Diseño . . . 42
7.3.3 Fundamento técnico de robustez heurística . . . 42
7.3.4 Validación de hipótesis "Peras vs. Manzanas" . . . 42
7.3.5 Hallazgo crítico: Arquitectura fail-safe validada empíricamente . . . 42
7.3.6 Implicaciones para Fase 2: Evolución de producción retrospectiva a prospectiva . . . 43
7.3.7 Conclusión: Robustez empíricamente validada . . . 44
7.4 Hallazgo principal: Gobernanza determina éxito . . . 44
7.5 Hallazgo emergente: La maduración de los datos potencia al modelo . . . 44
7.6 Gobernanza activa y estrategia de maduración del sistema . . . 45
7.6.1 Componentes de la estrategia de maduración . . . 45
7.6.2 Evidencia empírica de maduración exitosa . . . 45
7.6.3 Lecciones clave de gobernanza . . . 45
7.6.4 Implicaciones para escalabilidad . . . 46
7.7 Calidad de datos clínicos como factor determinante . . . 46
7.7.1 Contexto de heterogeneidad en salud pública . . . 46
7.7.2 Arquitectura multicapa como respuesta . . . 46
7.7.3 Oportunidad de mejora sistémica . . . 46
7.8 Conservadurismo apropiado en screening poblacional . . . 46
7.8.1 Trade-off Sensitivity vs Precision en contexto clínico . . . 46
7.8.2 Comparación con alternativas . . . 46
7.8.3 Mejora esperada en Fase 2 . . . 46
7.9 Lecciones para implementación prospectiva (Fase 1 a Fase 2) . . . 47
7.9.1 Calidad de datos precede despliegue prospectivo . . . 47
7.9.2 De clasificador principal a red de seguridad . . . 47
7.9.3 Estandarización como habilitador . . . 47
7.9.4 Objetivo final: Oportunidad diagnóstico-terapéutica . . . 47
7.10 Marco de gobernanza para formalización . . . 48
7.10.1 Estructura multicapa propuesta . . . 48
7.10.2 Protocolos operacionales . . . 48
7.10.3 Human-in-the-Loop como diseño central . . . 48

7.11 Consideraciones ético-legales . . . 49
7.12 Valor operacional y económico . . . 49
7.13 Cumplimiento Normativo y Responsabilidad Administrativa . . . 49
7.13.1 Alineamiento con Circular N° 11 MINSAL y Oficio N° 711 MinCiencia . . . 49
7.14 De Fase 1 a Fase 2: Fundamentos para una evolución sostenida . . . 49
7.14.1 El Modelo como "Capa de Calidad" . . . 50
7.14.2 Condiciones habilitantes y desafíos pendientes . . . 50
7.15 Limitaciones . . . 50
7.15.1 Limitaciones Metodológicas . . . 50
7.16 Trabajo futuro . . . 51

8 Conclusiones **53**
8.1 Mensaje final . . . 53

9 Referencias **55**
9.1 Fundamentos teóricos y metodológicos . . . 55
9.2 Procesamiento de lenguaje natural médico . . . 55
9.3 Ética e implementación de IA en salud . . . 55
9.4 Calidad de datos y validación clínica . . . 56
9.5 Protección de datos y marco legal Chile . . . 56
9.6 Machine learning interpretable y explicabilidad . . . 56
9.7 Listas de espera quirúrgicas y gestión sanitaria . . . 57

10 Apéndice A: Tabla de Features del Modelo **58**

11 Apéndice B: Métricas Detalladas por Servicio **59**

12 Apéndice C: Protocolo de Validación Estandarizada **60**
12.1 C.1 Conformación del equipo validador . . . 60
12.2 C.2 Criterio de clasificación: PERSPECTIVA PROSPECTIVA . . . 60
12.3 C.3 Proceso de validación de casos . . . 60
12.4 C.4 Documentación obligatoria . . . 61
12.5 C.5 Métricas a reportar . . . 61
12.6 C.6 Criterios de exclusión de casos . . . 61
12.7 C.7 Frecuencia y actualización . . . 61
12.8 C.8 Soporte y escalamiento . . . 61

13 Apéndice D: Ejemplos de Explicabilidad **62**
13.1 D.1 Caso verdadero positivo: Carcinoma ductal mama . . . 62
13.2 D.2 Caso verdadero negativo: Patología benigna . . . 62
13.3 D.3 Caso indeterminado: Conflicto de evidencia . . . 62
13.4 D.4 Importancia global de variables (VIP) . . . 63
13.5 D.5 Análisis LIME: Explicación local para auditoría . . . 63
13.6 D.6 Uso de explicabilidad en gobernanza . . . 63

14 Anexo E: Flujos Operacionales del Sistema **65**
14.1 E.1 Flujo: Clasificación → Revisión clínica → Corrección → Retroalimentación . . . 65
14.2 E.2 Flujo: Override clínico . . . 68
14.3 E.3 Flujo operacional Fase 1: Ciclo mensual de validación . . . 71
14.4 E.4 Modelo RACI: Responsabilidades operacionales . . . 74

1 Resumen Ejecutivo: Modernización y Seguridad en la Gestión Oncológica

Hacia un Estado Inteligente: Inteligencia Artificial como Red de Seguridad en Salud Pública

La gestión de las Listas de Espera en el sistema público de salud enfrenta el desafío crítico de identificar oportunamente a los pacientes con sospecha de cáncer dentro de un volumen masivo de derivaciones. El presente documento formaliza un modelo de Inteligencia Artificial desarrollado internamente por el Ministerio de Salud, diseñado no para reemplazar al médico, sino para actuar como una **red de seguridad institucional** que garantiza que ningún caso oncológico quede "oculto" por problemas administrativos o de registro.

1.1 Valor Público y Estratégico

La implementación de este sistema representa un avance sustantivo en la modernización del Estado, alineado con los principios de eficiencia, probidad y oportunidad en la atención:

1. **Seguridad del Paciente (Prioridad #1):** El sistema actúa como un "tamiz de alta sensibilidad". Su objetivo principal es rescatar casos oncológicos que podrían haber sido sub-priorizados erróneamente. La validación demuestra que el modelo tiene una sensibilidad superior al **94%**, minimizando drásticamente el riesgo de pérdida de pacientes.
2. **Eficiencia del Gasto y Recursos:** Al automatizar la revisión de casos de bajo riesgo con un 99% de confianza (VPN), el sistema libera miles de horas-hombre de especialistas, permitiendo que los equipos clínicos se enfoquen en los casos complejos y confirmados, optimizando la capacidad instalada de la red.
3. **Equidad Territorial:** El algoritmo aplica un estándar único de detección a nivel nacional, reduciendo la variabilidad en la clasificación que depende de la dotación de especialistas de cada servicio de salud.

1.2 Cumplimiento Normativo y Ético

Este desarrollo se adhiere estrictamente a los lineamientos de la **Circular N° 11 del MINSAL** y la **Guía de Uso de IA del Ministerio de Ciencia**, operando bajo el principio de **"Human-in-the-Loop" (Humano en el ciclo)**:

- **Soporte, no Reemplazo:** La IA no toma decisiones clínicas finales ni egresa pacientes de la lista de espera. Su función es clasificar, alertar y priorizar.
- **Abstención ante la Duda:** A diferencia de sistemas comerciales "caja negra", este modelo incluye un mecanismo de **autocontrol ético** (*err-iiv*). Cuando la información es ambigua o insuficiente, el sistema se abstiene de clasificar y marca el caso como "Indeterminado", forzando una revisión humana. Esto garantiza que la incertidumbre siempre sea resuelta por un profesional.

1.3 Estado de Madurez y Resultados

Tras un despliegue nacional y validación rigurosa, el sistema ha alcanzado un nivel de madurez operativa que valida su paso a una fase de vigilancia continua:

- **Robustez Técnica:** En validación controlada, el sistema demostró una precisión superior al 99%.
- **Impacto Real (Corte Octubre 2025):** Se ha observado un "Círculo Virtuoso de Datos". La implementación del sistema ha incentivado una mejora en la calidad del registro clínico en origen, lo que elevó la precisión del modelo (VPP) del 59.3% al **83.5%** sin necesidad de modificar el algoritmo.
- **Gestión de Incertidumbre:** El sistema ha demostrado ser capaz de detectar cambios en los patrones de datos (como se observó en noviembre de 2025), ajustando sus umbrales de seguridad automáticamente para proteger casos complejos, privilegiando la revisión manual sobre el error automático.

En conclusión, este sistema no es solo una herramienta tecnológica; es un instrumento de **gobernanza sanitaria**. Formalizar su uso permite al Ministerio de Salud ejercer una supervisión más efectiva, transparente y basada en datos sobre la gestión oncológica nacional.

2 Declaración de Significancia

Este trabajo establece un precedente en la administración pública chilena: la implementación exitosa de un sistema de IA soberano (no comprado, desarrollado internamente) que opera a escala nacional sobre datos sensibles. Demuestra que es posible conciliar la eficiencia algorítmica con la seguridad clínica y la ética pública, transformando datos históricos desestructurados en información accionable para salvar vidas.

3 Resumen (Propuesta Final - Primera Versión)

Antecedentes: La implementación de modelos de aprendizaje automático (ML) en sistemas de salud pública enfrenta el desafío de la heterogeneidad de datos y la variabilidad operativa. Este estudio presenta la validación y despliegue a escala nacional de un sistema de clasificación oncológica multicapa para gestionar listas de espera quirúrgicas (LEQ) en Chile ($n \approx 400,000$ casos mensuales), operando en un entorno de recursos limitados y alta incertidumbre.

Métodos: Se desarrolló una arquitectura híbrida "fail-safe" que integra tres fuentes independientes de evidencia: (1) heurísticas clínicas determinísticas, (2) análisis semántico mediante embeddings, y (3) un clasificador probabilístico XGBoost. La incertidumbre se gestiona mediante el framework err(iiv), que impone abstención selectiva y deriva casos conflictivos a revisión humana obligatoria (Human-in-the-Loop). La estabilidad del sistema se evaluó mediante el Índice de Estabilidad Poblacional (PSI) y el desempeño se analizó estratificadamente en 29 Servicios de Salud.

Resultados: En validación controlada ($n = 5,002$), el sistema alcanzó un AUC-ROC de 99.96%. En producción real, se observó una divergencia distribucional esperada en el componente de ML ($PSI = 7.095$), sin embargo, la estabilidad operacional global se preservó ($PSI = 0.057$) gracias a la robustez de las capas heurísticas y la lógica jerárquica. Aunque la sensibilidad global inicial fue del 79.8% debido a inconsistencias en la validación local, el subgrupo de 26 servicios con procesos estandarizados mantuvo una sensibilidad del 94.7%[1]. El monitoreo longitudinal reveló una maduración orgánica del sistema: en el corte de octubre de 2025, sin reentrenamiento algorítmico, la Precisión (VPP) aumentó del 59.3% al 83.5%[2], impulsada por la mejora en la calidad del registro clínico en origen. El Valor Predictivo Negativo (VPN) de 99.3% fue validado robustamente mediante muestreo aleatorio ($n = 250$) y análisis de sensibilidad, descartando sesgo de verificación crítico[3].

Conclusiones: La arquitectura multicapa demostró ser resiliente ante la degradación de datos en entornos no controlados, actuando efectivamente como una red de seguridad. Los resultados indican que el éxito de la IA en salud pública depende menos de la sofisticación algorítmica aislada y más de una gobernanza operativa sólida que estandarice la validación humana. El sistema generó un ciclo virtuoso de mejora de datos, validando su transición de una herramienta de screening retrospectivo a un mecanismo de vigilancia prospectiva nacional.

Nota sobre esta versión: Este documento constituye la primera versión de formalización del sistema, basada en los datos actuales disponibles. Dada la naturaleza evolutiva del modelo y los procesos de gobernanza, se requiere monitoreo continuo y actualización periódica de las métricas de desempeño en el contexto de producción real.

Palabras clave: Inteligencia artificial médica, clasificación oncológica, arquitectura multicapa, err(iiv), calidad de datos clínicos, human-in-the-loop, gobernanza IA salud, salud pública Chile

4 Introducción

4.1 Contexto nacional: Desafío de listas de espera quirúrgicas oncológicas

El sistema público de salud de Chile enfrenta un desafío creciente en la gestión de listas de espera quirúrgicas, particularmente en el contexto oncológico donde la oportunidad de diagnóstico y tratamiento tiene impacto directo en pronóstico y sobrevida de pacientes. Mensualmente, los cortes de Listas de Espera no GES generados desde la plataforma SIGTE identifican casos en espera de atención, de los cuales 400,000-450,000 corresponden a espera para atención quirúrgica. Es en este volumen donde el sistema necesita soluciones escalables que garanticen identificación oportuna de casos sospechosos de cáncer sin comprometer la calidad asistencial.

La clasificación adecuada de estos casos es crítica para múltiples objetivos del sistema de salud:

1. **Priorización clínica:** Asegurar que casos oncológicos genuinos reciban atención preferente según severidad y urgencia
2. **Cumplimiento GES:** Garantizar tiempos de atención establecidos en Garantías Explícitas en Salud para patologías oncológicas
3. **Optimización de recursos:** Asignar capacidad quirúrgica limitada a casos que más lo requieren
4. **Prevención de pérdidas:** Evitar que casos sospechosos sean clasificados erróneamente como no prioritarios y "caigan" del sistema

El proceso tradicional de clasificación manual presenta limitaciones significativas:

- **Carga de trabajo insostenible:** Revisión caso por caso de miles de derivaciones mensuales
- **Variabilidad inter-observador:** Criterios heterogéneos entre profesionales y servicios
- **Riesgo de omisiones:** Fatiga y presión asistencial aumentan probabilidad de errores
- **Falta de trazabilidad:** Dificultad para auditar decisiones y procesos

4.2 El problema de heterogeneidad en salud pública

A diferencia de contextos de investigación controlada o sistemas privados relativamente homogéneos, el sistema público de salud chileno presenta desafíos únicos de heterogeneidad que condicionan cualquier solución tecnológica:

Heterogeneidad de recursos tecnológicos: Los 29 servicios de salud del país operan con infraestructuras tecnológicas variables. Mientras algunos servicios cuentan con sistemas de registro electrónico avanzados, otros mantienen procesos mixtos (papel y digital) o sistemas legacy con limitaciones de interoperabilidad. Esta realidad impone restricciones sobre qué soluciones son viables a escala nacional.

Heterogeneidad de recursos humanos: La disponibilidad de especialistas, particularmente oncólogos y cirujanos con expertise en patología oncológica, varía significativamente entre servicios. Grandes centros urbanos (Concepción, Valparaíso, Santiago) cuentan con equipos multidisciplinarios consolidados, mientras servicios de menor tamaño dependen de rotaciones de especialistas o coordinación con centros de referencia. Esta disparidad afecta tanto la capacidad de clasificación inicial como la validación de sistemas automatizados.

Heterogeneidad en calidad de datos clínicos:

La completitud, precisión y estandarización de datos de derivación presenta variabilidad sustancial:

- **Codificación CIE-10:** Mientras algunos servicios implementan codificación obligatoria y validada, otros tienen tasas significativas de códigos ausentes o inespecíficos
- **Descripciones textuales:** La calidad de descripciones clínicas varía desde notas detalladas con terminología precisa hasta descripciones mínimas o ambiguas
- **Estandarización terminológica:** No existe un glosario nacional obligatorio de términos oncológicos, resultando en diversidad de expresiones para condiciones similares

Desafío de diseño:

Esta heterogeneidad impone un requisito fundamental para cualquier solución de IA en este contexto: debe ser **robusta ante imperfección de datos**, manteniendo desempeño aceptable incluso con información parcial o subóptima, sin colapsar cuando los datos no cumplen estándares ideales.

4.3 Marco teórico: Framework err(iiv) para control de errores

El sistema desarrollado se fundamenta en el framework **err(iiv)** (Is-It-Valid error function), originalmente propuesto para detección de alucinaciones en modelos de lenguaje. Este trabajo representa una adaptación médica pionera del framework para el contexto de clasificación oncológica.

Fundamento conceptual:

El framework err(iiv) se basa en el principio de que la **presencia de conflictos entre fuentes independientes de evidencia** es un indicador confiable de alta incertidumbre o error potencial. Cuando múltiples sistemas que operan con diferentes mecanismos (reglas heurísticas, análisis semántico, modelos estadísticos) arriban a conclusiones inconsistentes, la probabilidad de error es significativamente mayor que cuando hay concordancia.

Marco metodológico (no implementación literal):

La función de error se define conceptualmente como:

$$\text{err(iiv)}(x) = \omega_1 \cdot \text{DIVERGENCE}_{\text{prob}} + \omega_2 \cdot \text{CONFLICT}_{\text{categorical}} + \omega_3 \cdot \text{CONFLICT}_{\text{heuristic}} + \omega_4 \cdot \text{MIXED}_{\text{evidence}}$$

Nota crítica sobre implementación: Esta formulación representa el **marco metodológico** que guía el diseño del sistema de detección de incertidumbre, no una ecuación literal implementada en código. En la práctica, los componentes de conflicto (DIVERGENCE, CONFLICT, MIXED) están codificados como features individuales que el clasificador XGBoost aprende a integrar de forma óptima. Los pesos ω_i no son constantes fijas sino importancias relativas derivadas del análisis de feature importance del modelo entrenado. Este enfoque pragmático permite que el sistema aprenda automáticamente qué tipos de conflictos son más informativos para detectar incertidumbre genuina, resultando más robusto que una calibración manual de pesos fijos.

Donde cada componente detecta un tipo específico de inconsistencia:

- **$\text{DIVERGENCE}_{\text{prob}}$:** Mide divergencia entre probabilidades asignadas por capa semántica S(x) y clasificador ML M(x) en zonas críticas de decisión. Valores altos indican que ambos modelos probabilísticos discrepan sustancialmente.
- **$\text{CONFLICT}_{\text{categorical}}$:** Detecta contradicciones entre evidencia categórica (códigos CIE-10, keywords) y evidencia probabilística (embeddings, XGBoost). Por ejemplo: CIE-10 maligno pero probabilidad ML baja, o viceversa.
- **$\text{CONFLICT}_{\text{heuristic}}$:** Identifica desacuerdos entre predicciones basadas en reglas clínicas explícitas y predicciones estadísticas. Captura casos donde lógica médica codificada difiere de patrones aprendidos de datos.
- **$\text{MIXED}_{\text{evidence}}$:** Señala contradicciones internas dentro de fuentes categóricas, como presencia simultánea de indicadores malignos y benignos (ej: CIE-10 benigno + keywords oncológicos).

Interpretación médica:

En el contexto clínico, err(iiv) alto indica que el caso presenta **ambigüedad genuina o información conflictiva** que dificulta clasificación confiable. Esta situación puede originarse en:

- Descripción clínica genuinamente ambigua (patologías borderline)
- Datos incompletos o de baja calidad
- Terminología no estándar o confusa
- Casos complejos que requieren expertise especializado

Innovación clínica - Abstención selectiva:

A diferencia de sistemas tradicionales que fuerzan una clasificación incluso con alta incertidumbre, este framework implementa **abstención selectiva**: cuando err(iiv)(x) > umbral, el sistema clasifica el caso como "indeterminado" y deriva a revisión humana obligatoria (human-in-the-loop).

Esta estrategia presenta ventajas significativas:

- **Seguridad:** Previene clasificaciones incorrectas de alto riesgo
- **Eficiencia:** Revisión humana solo en casos que genuinamente lo requieren (<1%)
- **Aprendizaje:** Casos indeterminados informan mejora continua del modelo
- **Aceptabilidad:** Profesionales mantienen control en casos complejos

4.4 Objetivos del estudio

Este trabajo busca demostrar viabilidad técnica y operacional de un sistema de clasificación oncológica automatizada para el sistema público de salud chileno, con los siguientes objetivos específicos:

Objetivo 1 - Desarrollo y validación técnica: Desarrollar y validar arquitectura multicapa que integre evidencia heurística, semántica y estadística, logrando métricas de desempeño superiores a estándares clínicos (sensitivity >95%, specificity >90%) en dataset controlado.

Objetivo 2 - Evaluación en producción real: Evaluar desempeño del sistema en condiciones reales de implementación en 29 servicios de salud de la red asistencial, documentando robustez ante heterogeneidad de datos y procesos.

Objetivo 3 - Análisis de calidad de datos: Analizar impacto de calidad de datos clínicos y estandarización de procesos en efectividad del modelo, identificando factores determinantes de éxito o falla en implementación local.

Objetivo 4 - Lecciones de gobernanza: Identificar lecciones críticas sobre gobernanza, protocolos operacionales y estructura institucional necesarias para implementación exitosa de IA en salud pública.

Objetivo 5 - Propuesta de formalización: Proponer marco de formalización institucional en MINSAL que incluya estándares técnicos, protocolos de implementación, estructura de gobernanza y mecanismos de monitoreo continuo.

4.5 Enfoque por fases y visión estratégica de largo plazo

Este trabajo se enmarca en una iniciativa de modernización progresiva, cuya evolución está condicionada a la maduración digital de la red asistencial. Se concibe no como un proyecto aislado, sino como un componente que evoluciona en función de las capacidades del ecosistema de salud.

Fase 1: Screening centralizado (Estado Actual)

La etapa vigente se enfoca en el procesamiento centralizado de los cortes de listas de espera generados por SIGTE. En esta configuración, ante la ausencia de una clasificación estandarizada en el origen, el modelo opera como el **mecanismo principal de identificación**. Su función es procesar el volumen masivo de datos históricos y nuevos ingresos para detectar casos oncológicos que no fueron priorizados correctamente. Esta fase permite sanear la lista de espera y entregar alertas a los servicios de salud, mientras se generan las condiciones para una gestión más distribuida.

Fase 2: Horizonte estratégico (Clasificación en origen y IA como capa de calidad)

La visión objetivo apunta a un cambio de paradigma donde la responsabilidad de la clasificación oncológica se traslada al **punto de origen** (hospitales y centros de derivación), integrándose en el flujo de trabajo clínico diario. En este escenario futuro, el modelo de IA no desaparece, sino que transforma su rol hacia una **"Capa de Aseguramiento de Calidad"** o red de seguridad.

Bajo esta lógica, los establecimientos clasificarían sus casos en tiempo real. El modelo actuaría como un auditor continuo que analiza estas clasificaciones para: 1. Validar que los casos marcados como "no oncológicos"

efectivamente lo sean (prevención de fugas). 2. Detectar inconsistencias entre el diagnóstico clínico y la priorización administrativa asignada.

Dependencias sistémicas: El tránsito hacia esta fase no tiene plazos rígidos, pues depende de avances estructurales que exceden el alcance de este algoritmo, incluyendo: - Maduración de los procesos de registro en fichas clínicas electrónicas. - Avances en la Agenda de Interoperabilidad para la derivación quirúrgica. - Desarrollos de proveedores tecnológicos para integrar campos estructurados de clasificación. - Cambio cultural y capacitación en los equipos clínicos locales.

Por tanto, el modelo actual se constituye como una pieza fundacional flexible, diseñada para operar en el escenario actual de centralización y preparada para adaptarse como herramienta de vigilancia en un futuro escenario descentralizado e interoperable.

Objetivo último: Trazabilidad y seguridad

Independientemente de la fase operativa, el propósito intransable es asegurar que ningún paciente oncológico "se pierda" en la gestión administrativa. La tecnología busca garantizar que la identificación del riesgo sea oportuna, ya sea mediante detección centralizada (hoy) o validación distribuida (mañana).

5 Métodos

NOTA: Este estudio documenta Fase 1 (screening retrospectivo de backlog de listas de espera). La metodología presentada sienta bases técnicas y operacionales para Fase 2 donde el modelo operará como red de seguridad complementaria a clasificación local prospectiva.

5.1 Visión general del sistema: Pipeline completo

El sistema de clasificación oncológica implementa un pipeline de procesamiento en cinco etapas que transforma casos textuales sin estructura desde listas de espera SIGTE en clasificaciones accionables con trazabilidad completa.

Pipeline end-to-end:

```
ENTRADA (caso x) → [1] Preprocesamiento → [2] Generación de Evidencia Multicapa
                 → [3] Integración XGBoost → [4] Detección err(iiv)
                 → [5] Lógica de Decisión Jerárquica → SALIDA (clasificación + explicación)
```

Descripción de cada etapa:

[1] Preprocesamiento:

- **Input:** Texto crudo de derivación quirúrgica (ej: "c50.9 carcinoma ductal mama izq her2+")
- **Normalización:** Minúsculas, eliminación caracteres especiales, preservación negaciones médicas
- **Detección CIE-10:** Extracción de códigos mediante regex pattern `[A-Z]\d{2,3}\.?\d{0,2}`
- **Output:** Texto limpio + códigos CIE-10 identificados

[2] Generación de Evidencia Multicapa (paralelo e independiente):

Tres módulos procesan el caso simultáneamente sin comunicación entre sí:

- **Capa Heurística H(x):** Genera 15 features binarias/contadores basados en reglas determinísticas (CIE-10, keywords, patrones textuales)
- **Capa Semántica S(x):** Genera 7 features continuas mediante embeddings (similitud coseno con centroides pre-entrenados de clases "cáncer" y "no cáncer")
- **Independencia garantizada:** Cada capa tiene acceso solo al texto preprocesado, NO a outputs de otras capas. Esto permite detección posterior de conflictos como señal de incertidumbre.

[3] Integración mediante XGBoost:

- **Input:** Vector de 40 features (15 heurísticas + 7 semánticas + 18 de ingeniería)
- **Proceso:** Gradient boosting ensemble de 100 árboles de decisión
- **Output:** Probabilidad continua $p_{xgb}(x) \in [0, 1]$ y clasificación binaria con threshold $\tau = 0.6299$

[4] Detección de Incertidumbre err(iiv):

- **Input:** Evidencias de las 3 capas (H, S, M) + probabilidades
- **Cálculo:** $\text{err}_{iiv}(x) = \sum_{i=1}^{4} \omega_i \cdot C_i(x)$ donde C_i detectan conflictos entre capas
- **Output:** Score de confianza; si $\text{err}_{iiv}(x) > 4.2 \rightarrow$ marcar como caso de alta incertidumbre

[5] Lógica de Decisión Jerárquica:

Cascada de 8 reglas evaluadas en orden de prioridad clínica:

1. Si err(iiv) > threshold O texto vacío → **INDETERMINADO** (abstención)
2. Si CIE-10 maligno (C00-C97) → **SOSPECHOSO** (P1, máxima prioridad)
3. Si CIE-10 sospechoso (D37-D48) → **SOSPECHOSO** (P1B)
4. Si $p_{xgb} > 0.90 \rightarrow$ **SOSPECHOSO** (P2, override estadístico)
5. Si keywords malignos presentes → **SOSPECHOSO** (P3)
6. Si evidencia benigna PURA → **NO_SOSPECHOSO** (P4A/B/C)
7. Usar threshold calibrado: $p_{xgb} > 0.6299 \rightarrow$ **SOSPECHOSO** (P5)
8. Default → **INDETERMINADO**

Output final incluye:

- **Clasificación:** {SOSPECHOSO, NO_SOSPECHOSO, INDETERMINADO}
- **Confianza:** {muy_alta, alta, media, baja}
- **Probabilidades:** p_{xgb}, p_{sem}
- **Regla aplicada:** (P0-P5, DEFAULT)
- **Razón textual:** Explicación clínica de la decisión
- **Features determinantes:** Top 5 por SHAP values

Diferencias clave entre entrenamiento e inferencia:

Fase de entrenamiento (offline):

1. Dataset etiquetado (n=5,002) → Preprocesamiento
2. Cálculo de centroides semánticos (promedio embeddings por clase)
3. Generación de 40 features para todos los casos
4. Entrenamiento XGBoost con validación cruzada 5-fold
5. Optimización de threshold mediante Índice de Youden
6. Calibración de pesos err(iiv) mediante grid search
7. Guardado de modelo, centroides, y parámetros

Fase de inferencia (online):

1. Caso nuevo → Preprocesamiento
2. Generación paralela de evidencia (H, S) usando modelo pre-entrenado
3. Predicción XGBoost (M)
4. Cálculo err(iiv) con pesos fijos
5. Aplicación de lógica jerárquica
6. Retorno de clasificación + explicación

Este diseño modular permite:

- **Paralelización:** Capas H y S pueden procesarse en paralelo
- **Explicabilidad:** Cada decisión tiene trazabilidad completa
- **Robustez:** Falla parcial de una capa no colapsa el sistema
- **Mantenibilidad:** Cada módulo puede actualizarse independientemente

5.2 Arquitectura multicapa con generación independiente de evidencia

El sistema implementa una arquitectura de tres capas que operan independientemente para generar evidencia sobre la naturaleza oncológica de cada caso. El diseño multicapa responde al principio de **robustez ante heterogeneidad**: cada capa compensa limitaciones de las otras, permitiendo clasificación adecuada incluso con datos parciales o subóptimos.

5.2.1 Capa 1: Módulo heurístico (Evidencia H(x))

Propósito: Capturar evidencia categórica basada en codificación clínica estándar y terminología médica explícita.

Features generadas (n=15):

Códigos CIE-10:

- `cie10_neo`: Presencia de códigos malignos (C00-C97)
- `cie10_benigno`: Presencia de códigos benignos (D10-D36)
- `cie10_sospechoso`: Presencia de códigos sospechosos específicos (D37-D48, ciertos R-codes)

Keywords oncológicos (n=75+ términos):

- `regex_flag`: Detección de términos inequívocamente oncológicos (carcinoma, melanoma, linfoma, sarcoma, leucemia, etc.)

- `conteo_keywords_malignos`: Número de keywords oncológicos en texto
- `conteo_keywords_localizacion`: Número de localizaciones anatómicas oncológicas (~200 términos: mama, próstata, pulmón, colon, etc.)

Indicadores de certeza médica:

- `tiene_negacion`: Presencia de negaciones médicas ("sin evidencia de", "descarta", "ausencia de")
- `tiene_sospecha`: Términos de incertidumbre ("sospecha", "probable", "posible", "a estudio")
- `tiene_benigno_explicito`: Términos benignos explícitos ("benigno", "no maligno", "adenoma")

Otros:

- `largo_texto`: Longitud del texto (proxy de completitud de descripción)
- `tiene_cie10`: Presencia de al menos un código CIE-10
- `densidad_keywords`: Proporción keywords / total palabras

Ventaja de robustez: Esta capa funciona incluso con datos mínimos. Un caso con solo código CIE-10="C50.9" (carcinoma mama) sin descripción textual será clasificado correctamente por evidencia heurística pura.

5.2.2 Capa 2: Módulo semántico (Evidencia S(x))

Propósito: Capturar significado contextual del texto mediante representaciones vectoriales densas, permitiendo comprensión de terminología variable o no estándar.

Tecnología: - Modelo de embeddings: `toshk0/nomic-embed-text-v2-moe:Q6_K` (768 dimensiones) - Implementación: Ollama (local, sin dependencia de APIs externas) - Preprocesamiento: Normalización de texto, remoción de stopwords preservando negaciones médicas

Proceso:

1. **Entrenamiento de centroides:** Con dataset etiquetado (seeds.xlsx), se calculan centroides para clases "cáncer" y "no cáncer" mediante promedio de embeddings de cada clase.
2. **Validación semántica:** Análisis UMAP de separación entre centroides confirma distancia=1.441 (excelente separabilidad en espacio latente).
3. **Generación de features por caso:** Para cada texto nuevo:
   - Generar embedding de 768 dimensiones
   - Calcular similitud coseno con ambos centroides
   - Derivar features de similitud

Features generadas (n=7): - `prob_cancer_cos`: Similitud coseno con centroide cáncer - `sim_cancer`: Similitud normalizada con centroide cáncer - `sim_no_cancer`: Similitud normalizada con centroide no cáncer - `dif_similitudes`: Diferencia sim_cancer - sim_no_cancer - `ratio_similitudes`: Razón sim_cancer / sim_no_cancer - `max_similitud`: Máxima similitud entre ambos centroides - `confianza_coseno`: Medida de confianza basada en separación

Ventaja de robustez: Esta capa procesa texto libre sin requerir estructura predefinida. Captura expresiones no estándar o variantes terminológicas ("tumor maligno" vs "neoplasia maligna" vs "ca" [abreviación oncológica]) mediante similaridad semántica en espacio vectorial.

5.2.2.1 Formalización matemática: Modelado en espacio latente

Sea $\mathcal{V}$ un espacio vectorial de dimensión $d = 768$ generado por el modelo *nomic-embed-text-v2*. Para cada documento x, definimos su representación vectorial $\vec{v}_x \in \mathbb{R}^d$.

Centroides de Referencia:

Los centroides $\vec{\mu}_{pos}$ (cáncer) y $\vec{\mu}_{neg}$ (no cáncer) no son meros promedios, sino vectores medios normalizados de la variedad (manifold) de entrenamiento:

$$\vec{\mu}_c = \frac{1}{|S_c|} \sum_{x \in S_c} \frac{\vec{v}_x}{\|\vec{v}_x\|}$$

donde S_c es el conjunto de casos etiquetados de clase $c \in \{\text{pos}, \text{neg}\}$ y la normalización $\frac{\vec{v}_x}{\|\vec{v}_x\|}$ asegura que vectores de diferentes longitudes tengan igual peso.

Similitud Coseno como Métrica:

Para un nuevo caso x, la feature de similitud semántica $f_{sim}(x, c)$ se calcula mediante similitud coseno, que es invariante a la longitud del texto:

$$f_{sim}(x, c) = \frac{\vec{v}_x \cdot \vec{\mu}_c}{\|\vec{v}_x\| \|\vec{\mu}_c\|} \in [-1, 1]$$

Propiedades clave de esta formulación:

1. **Invarianza a longitud:** Un texto verboso ("carcinoma ductal infiltrante de mama izquierda estadio IIA HER2 positivo") y uno lacónico ("c50.9 ca mama") producen similitudes comparables si son semánticamente equivalentes.
2. **Proyección sobre eje de malignidad:** La diferencia $f_{sim}(x, \text{pos}) - f_{sim}(x, \text{neg})$ proyecta el caso sobre el eje semántico definido por $\vec{\mu}_{pos} - \vec{\mu}_{neg}$, capturando la "orientación oncológica" del texto.
3. **Robustez ante variantes lingüísticas:** Expresiones sinónimas ("tumor maligno", "neoplasia maligna", "cáncer", "ca") se agrupan en regiones cercanas del espacio latente por el pre-entrenamiento del modelo de embeddings.

Validación de separabilidad:

La distancia euclideana entre centroides normalizados es:

$$d(\vec{\mu}_{pos}, \vec{\mu}_{neg}) = \|\vec{\mu}_{pos} - \vec{\mu}_{neg}\| = 1.441$$

Este valor > 1.4 indica excelente separabilidad en el espacio de 768 dimensiones, validado mediante análisis UMAP de reducción dimensional durante la construcción de centroides semánticos.

5.2.3 Capa 3: Clasificador XGBoost V4.6 (evidencia M(x))

Propósito: Integrar evidencia de capas previas mediante modelo estadístico optimizado, capturando interacciones complejas y patrones no lineales.

IMPORTANTE - Aclaración sobre el rol de XGBoost en la arquitectura:

XGBoost **NO es el decisor final**, sino un **generador de evidencia probabilística** que opera dentro de un sistema jerárquico de toma de decisiones. El flujo completo es:

1. **Capas H, S, M generan evidencia independiente**: Heurística (H), Semántica (S) y XGBoost (M) producen predicciones de forma independiente
2. **XGBoost integra evidencia**: El modelo M(x) recibe features de H y S y genera una probabilidad $p_{xgb}(x) \in [0, 1]$
3. **Lógica jerárquica toma la decisión final**: Se aplican reglas de prioridad clínica (P0-P5) que pueden **override** la probabilidad XGBoost

Ejemplo de override: - Si XGBoost $\rightarrow p_{xgb} = 0.30$ (bajo) PERO CIE-10 = C50.9 (maligno) - Decisión final: **SOSPECHOSO** (Regla P1 prevalece sobre P5)

Justificación clínica: La evidencia categórica robusta (ej: CIE-10 maligno confirmado) tiene mayor peso que la evidencia estadística en casos de conflicto, privilegiando la seguridad del paciente. XGBoost solo determina

la decisión final en casos sin evidencia categórica fuerte (Regla P5: casos en "zona gris" donde solo queda evidencia probabilística).

Features de entrada (n=40 total): - 15 features heurísticas (Capa 1) - 7 features semánticas (Capa 2) - 18 features de ingeniería

Features de ingeniería diseñadas (n=18):

Transformaciones no lineales de probabilidades semánticas (n=4):

Dado $p_{sem}(x)$ = probabilidad semántica del caso x, se generan:

$$p_{sem}^2(x), \quad p_{sem}^3(x), \quad \log(p_{sem}(x) + \epsilon), \quad \sqrt{p_{sem}(x)}$$

Racionalidad: Las transformaciones capturan comportamientos extremos (p^2, p^3), estabilizan varianza (log) y suavizan distribución ($\sqrt{}$), permitiendo que XGBoost aprenda relaciones no lineales complejas.

Detección de conflictos entre capas (n=5):

$$\text{conflict}_{heur-sem}(x) = \mathbb{1}[\text{sign}(H(x)) \neq \text{sign}(S(x))]$$

$$\text{conflict}_{cie}(x) = \mathbb{1}[\text{CIE10}_{malig}(x) = 1 \wedge \text{CIE10}_{benig}(x) = 1]$$

$$\text{conflict}_{expr}(x) = \mathbb{1}[\text{expr}_{neoplas}(x) = 1 \wedge \text{expr}_{benig}(x) = 1]$$

$$\text{conflict}_{models}(x) = \mathbb{1}[|p_{sem}(x) - p_{heur}(x)| > \delta], \quad \delta = 0.3$$

Estos indicadores binarios señalan inconsistencias entre fuentes independientes, permitiendo al modelo aprender cuándo la evidencia conflictiva indica incertidumbre genuina.

Scores compuestos ponderados (n=3):

$$\text{score}_{heur}(x) = \sum_{i=1}^{15} w_i^{(h)} \cdot f_i^{(h)}(x)$$

donde $w_i^{(h)}$ son pesos optimizados mediante grid search sobre validación cruzada.

Configuración y entrenamiento de XGBoost:

El modelo XGBoost se entrena mediante gradient boosting con función objetivo:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

donde: - $l(y_i, \hat{y}_i) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$ es log-loss binaria - $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ penaliza complejidad (T = hojas, ω = pesos)

5.2.3.1 Función objetivo ponderada por costo clínico

El problema de optimización no busca minimizar el error global (Accuracy), sino **minimizar el Riesgo Clínico Esperado** $R(f)$. En contextos de screening oncológico, los errores tienen costos asimétricos: omitir un cáncer (FN) es clínicamente más grave que generar una revisión innecesaria (FP).

Matriz de Costos Clínicos:

Definimos la matriz de costos C donde C_{FN} es el costo de un falso negativo (omisión de cáncer que podría progresar sin tratamiento) y C_{FP} es el costo de un falso positivo (carga de revisión manual por profesional):

$$C = \begin{pmatrix} C_{TN} = 0 & C_{FP} \\ C_{FN} & C_{TP} = 0 \end{pmatrix}$$

donde: - $C_{FN} \gg C_{FP}$ (en screening oncológico, omitir un cáncer es mucho más grave que generar revisión innecesaria) - $C_{TP} = C_{TN} = 0$ (clasificaciones correctas no tienen costo asociado)

Riesgo Esperado:

El riesgo esperado del clasificador f bajo esta matriz de costos es:

$$R(f) = \mathbb{E}[C(Y, f(X))] = C_{FP} \cdot P(Y = 0) \cdot P(f(X) = 1|Y = 0) + C_{FN} \cdot P(Y = 1) \cdot P(f(X) = 0|Y = 1)$$

$$R(f) = C_{FP} \cdot (1 - \pi) \cdot \text{FPR} + C_{FN} \cdot \pi \cdot \text{FNR}$$

donde $\pi = P(Y = 1)$ es la prevalencia de casos sospechosos en la población de entrenamiento (6.7% en nuestro dataset).

Relación con scale_pos_weight:

En XGBoost, minimizar $R(f)$ con matriz de costos C es equivalente a entrenar con pesos asimétricos que ajustan la función de pérdida:

$$\text{scale\_pos\_weight} = \left(\frac{C_{FN}}{C_{FP}}\right) \cdot \left(\frac{n_{neg}}{n_{pos}}\right)$$

Este parámetro repondera los casos positivos durante el entrenamiento para reflejar tanto el **desbalance de clases** (n_{neg}/n_{pos}) como el **desbalance de costos** (C_{FN}/C_{FP}).

Fundamento ético del diseño: Este desbalance matemático no constituye un sesgo técnico no deseado, sino una **decisión de diseño explícita**. En el contexto de salud pública y listas de espera, el costo ético y sanitario de "perder" un paciente oncológico (Falso Negativo) es inmensamente superior al costo administrativo de revisar una ficha adicional que resulta ser benigna (Falso Positivo). El modelo está configurado intencionalmente para actuar como un tamiz de alta sensibilidad.

Calibración del ratio de costos:

De nuestro análisis con $\alpha = 2.26$ (factor conservador calibrado):

- Ratio empírico de clases: $\frac{n_{neg}}{n_{pos}} = \frac{4667}{335} = 13.93$
- `scale_pos_weight` calibrado: 10.5665
- Por tanto: $\frac{C_{FN}}{C_{FP}} = \frac{10.5665}{13.93} \cdot \alpha = 0.759 \times 2.26 = 1.72$

Interpretación clínica del ratio de costos $C_{FN}/C_{FP} \approx 1.72$:

Este ratio implica que el costo de omitir un caso sospechoso de cáncer es aproximadamente 1.72 veces mayor que el costo de generar una revisión manual innecesaria. Este balance conservador refleja:

1. **Costo clínico**: Demora diagnóstica puede permitir progresión tumoral, reduciendo opciones terapéuticas
2. **Costo operacional bajo de FP**: Revisión manual tiene costo marginal reducido (~10 min/caso por profesional capacitado)
3. **Diseño human-in-the-loop**: Sistema asume validación clínica obligatoria de TODOS los casos (sospechosos y no sospechosos), por lo que FP no genera derivaciones erróneas sino carga de revisión

Esta formulación permite **auditar la decisión de $\alpha = 2.26$ desde fundamentos de teoría de decisión**, transformando un hiperparámetro empírico en una decisión de política clínica transparente y justificable.

Parámetros optimizados mediante validación cruzada 5-fold estratificada:

Parámetro	Valor	Justificación
`max_depth`	6	Evita overfitting, captura interacciones orden 3-4
`eta` (learning rate)	0.1	Convergencia estable en ~100 iteraciones
`min_child_weight`	1	Optimizado empíricamente para flexibilidad
`subsample`	0.8	Reduce varianza (bootstrap)
`colsample_bytree`	0.8	Diversifica ensemble
`gamma`	0	Regularización mediante otros parámetros
`scale_pos_weight`	10.5665	**CRÍTICO**: De matriz costos clínicos, $C_{FN}/C_{FP} \approx 1.72$
`objective`	`binary:logistic`	Clasificación binaria probabilística
`eval_metric`	`auc`	Optimización AUC-ROC
`n_estimators`	100	Con early stopping en validación

5.2.3.2 Calibración de thresholds mediante índice de Youden

Los umbrales de decisión fueron calibrados independientemente usando **Índice de Youden** que maximiza:

$$J(\tau) = \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1$$

Proceso de optimización:

1. Para cada threshold candidato $\tau \in [0, 1]$ con paso 0.001:

$$\text{Sensitivity}(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}$$

$$\text{Specificity}(\tau) = \frac{TN(\tau)}{TN(\tau) + FP(\tau)}$$

2. Calcular $J(\tau)$ para cada τ
3. Seleccionar $\tau^* = \arg\max_\tau J(\tau)$

Thresholds óptimos identificados:

- **XGBoost**: $\tau^*_{xgb} = 0.6299$ con $J(\tau^*) = 0.9862$
- **Embeddings**: $\tau^*_{emb} = 0.4936$ con $J(\tau^*) = 0.9518$

Validación de calibración:

La calibración de probabilidades fue verificada mediante diagrama de confiabilidad (reliability diagram) y Brier Score:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2 = 0.0068$$

Un Brier Score < 0.01 indica excelente calibración: las probabilidades predichas reflejan fielmente las frecuencias observadas.

Procedimiento de optimización y congelamiento:

1. **Optimización en validación cruzada**: Los umbrales fueron optimizados utilizando validación cruzada estratificada 5-fold sobre el dataset de entrenamiento (n=5,002)
2. **Selección del umbral óptimo**: Para cada fold se calculó el Índice de Youden en la curva ROC
3. **Umbral final**: Se utilizó el promedio de los 5 umbrales óptimos obtenidos en cada fold
4. **Congelamiento pre-producción**: Los umbrales se fijaron **antes** del despliegue en producción y **no se han modificado** desde entonces
5. **Validación independiente**: Las métricas reportadas en la sección "Desempeño en producción real" corresponden a estos umbrales pre-establecidos, garantizando que no ha habido re-optimización sobre datos de producción

Esta estrategia previene overfitting y garantiza que las métricas de producción reflejen el desempeño real del modelo en datos completamente no vistos durante el entrenamiento.

Nota: Métrica calculada durante validación inicial del modelo (n=5,002). No incluida en pipeline de evaluación automática actual.

Ventaja de robustez: Esta capa integra evidencia parcial de múltiples fuentes. El mecanismo de boosting con muestreo permite al modelo aprender qué fuentes son más confiables bajo diferentes condiciones de datos, adaptándose a heterogeneidad.

5.3 Sistema de control de errores err(iiv): Formalización matemática

El framework err(iiv) implementa detección de incertidumbre mediante identificación de conflictos entre las tres capas independientes de evidencia. La hipótesis fundamental es:

> **Principio err(iiv):** Cuando fuentes de evidencia independientes que operan con mecanismos distintos arriban a conclusiones inconsistentes, la probabilidad de error de clasificación aumenta significativamente.

Definición formal de err(iiv):

Para un caso x, la función de error potencial se define como:

$$\text{err}_{iiv}(x) = \omega_1 \cdot D_{prob}(x) + \omega_2 \cdot C_{cat}(x) + \omega_3 \cdot C_{heur}(x) + \omega_4 \cdot M_{evid}(x)$$

donde $\omega_i \in \mathbb{R}^+$ son pesos calibrados y cada componente se define como:

1. DIVERGENCE_prob - Divergencia probabilística:

$$D_{prob}(x) = |p_{sem}(x) - p_{xgb}(x)| \cdot \mathbb{1}[p_{xgb}(x) \in (0.3, 0.7)]$$

Interpretación: Solo penaliza divergencia semántico-estadística cuando XGBoost está en "zona de incertidumbre" (30%-70%). Si $p_{xgb} > 0.7$ o $p_{xgb} < 0.3$, la divergencia con el módulo semántico es menos crítica.

Justificación del intervalo crítico (0.3, 0.7):

Análisis empírico en validación (n=5,002) muestra patrones diferenciados por zona de probabilidad:

- **Si** $p_{xgb}(x) > 0.7$: Modelo tiene alta confianza en malignidad. Divergencia semántica puede deberse a:
  - Textos lacónicos que embeddings no capturan bien ("c50.9 ca mama" vs. descripciones detalladas)
  - Limitaciones de centroides pre-entrenados ante terminología especializada
  - En estos casos, la evidencia estadística robusta prevalece
- **Si** $p_{xgb}(x) < 0.3$: Modelo tiene alta confianza en benignidad. Divergencia similar:

  - Falsos positivos semánticos por polisemia ("tumor benigno" activa similitud con "tumor")
  - XGBoost con 40 features integradas tiene mayor poder discriminativo
- **Si** $p_{xgb}(x) \in (0.3, 0.7)$: **Zona crítica** donde conflictos indican genuina incertidumbre:
  - Modelo estadístico no tiene confianza clara (probabilidad cercana a 50%)
  - Divergencia semántico-estadística señala ambigüedad intrínseca del caso
  - Estos casos son candidatos principales para err(iiv) > threshold → INDETERMINADO

Este diseño focaliza la detección de incertidumbre en la región donde la clasificación es más frágil, evitando abstenciones innecesarias en casos de alta confianza.

2. CONFLICT_categorical - Conflicto categórico-probabilístico:

$$C_{cat}(x) = \sum_{k \in \{\text{CIE10,KW,expr}\}} \mathbb{1}[\text{category}_k(x) \neq \text{sign}(p_{xgb}(x) - 0.5)]$$

Donde: - $\text{category}_k(x) \in \{-1, 0, +1\}$ indica evidencia categórica benigna (-1), ausente (0), o maligna (+1) - $\text{sign}(p_{xgb}(x) - 0.5)$ indica predicción probabilística

Valores altos indican contradicción entre evidencia explícita (CIE-10 maligno) y probabilidad asignada (baja).

3. CONFLICT_heuristic - Conflicto heurístico-estadístico:

$$C_{heur}(x) = \mathbb{1}[H(x) \neq M(x)] \cdot |p_{xgb}(x) - 0.5|$$

Donde: - $H(x) \in \{0, 1\}$ es clasificación heurística determinística - $M(x) \in \{0, 1\}$ es clasificación XGBoost con threshold - El factor $|p_{xgb}(x) - 0.5|$ pondera por confianza del modelo estadístico

Casos con desacuerdo categórico-estadístico fuerte (heurística dice maligno, XGBoost dice benigno con alta confianza) tienen C_{heur} alto.

4. MIXED_evidence - Evidencia interna contradictoria:

$$M_{evid}(x) = \sum_{(a,b) \in \mathcal{P}} \mathbb{1}[\text{evid}_a(x) = +1 \wedge \text{evid}_b(x) = -1]$$

Donde $\mathcal{P}$ son pares mutuamente excluyentes: {(CIE10_malig, CIE10_benig), (keywords_malig, expr_benig), …}

Pesos calibrados empíricamente:

Los pesos ω_i fueron optimizados mediante búsqueda en grilla sobre casos de validación con análisis post-hoc de errores:

$$\omega^* = \arg\min_{\omega} [\lambda_1 \cdot \text{FN}(\omega) + \lambda_2 \cdot \text{FP}(\omega) + \lambda_3 \cdot \text{Abs}(\omega)]$$

donde: - $\text{FN}(\omega)$, $\text{FP}(\omega)$ cuentan errores con umbral err(iiv) dado por ω - $\text{Abs}(\omega)$ cuenta casos clasificados como indeterminados - $\lambda_1 = 10, \lambda_2 = 1, \lambda_3 = 5$ reflejan costos relativos

Pesos óptimos identificados:

$$\omega_1 = 2.1, \quad \omega_2 = 1.5, \quad \omega_3 = 1.8, \quad \omega_4 = 2.5$$

Umbral de abstención:

Un caso se clasifica como "indeterminado" si:

$$\text{err}_{iiv}(x) > \tau_{iiv} = 4.2$$

Este threshold fue calibrado para lograr tasa de abstención $< 1\%$ manteniendo reducción $> 80\%$ de errores en casos indeterminados vs. clasificación forzada.

Validación empírica del framework:

En dataset de validación (n=5,002): - **Casos determinados** (err < 4.2, n=4,990): Error rate = 0.70% (35 errores) - **Casos indeterminados** (err 4.2, n=12): Si forzamos clasificación, error rate proyectado = 66.7% (8/12)

Esto confirma que err(iiv) identifica correctamente casos con alta probabilidad de error.

5.3.1 Nota sobre implementación del framework err(iiv)

Relación entre formalización teórica e implementación práctica:

La formalización matemática presentada (componentes D_{prob}, C_{cat}, C_{heur}, M_{evid} con pesos ω_i) representa el **marco conceptual** que guía el diseño del sistema de detección de incertidumbre. En la implementación actual (Fase 1), estos componentes están codificados como **features individuales** del clasificador XGBoost V4.6:

- **Conflictos detectados como features**:
  - `conflicto_cie` (implementa parcialmente M_{evid})
  - `conflicto_expresiones` (implementa parcialmente M_{evid})
  - `conflicto_modelos` (implementa parcialmente C_{heur})
  - `heuristico_semantico_agree` (implementa C_{cat})
- **Agregación mediante aprendizaje**: Los pesos $\omega_1, \omega_2, \omega_3, \omega_4$ presentados representan **importancias relativas derivadas del análisis de feature importance** de XGBoost, no constantes explícitas en código. El modelo aprende automáticamente qué combinaciones de conflictos indican alta probabilidad de error.
- **Abstención selectiva**: La decisión de clasificar como "indeterminado" se determina mediante nivel de **confianza XGBoost** (línea 1050-1051 en clasificador_v4.R), no mediante comparación directa contra $\tau_{iiv} = 4.2$.

Ventajas de esta implementación: 1. **Robustez**: XGBoost aprende pesos óptimos para cada tipo de conflicto según patrones reales 2. **Adaptabilidad**: Sistema se ajusta a nuevos tipos de conflictos sin recodificar pesos manualmente 3. **Simplicidad**: Evita lógica compleja de agregación manual de componentes err(iiv)

Limitación para auditabilidad: Los pesos ω_i no son auditables como constantes explícitas. Para Fase 2 (sistema en producción prospectiva), se recomienda implementar función err(iiv) explícita con pesos fijos para permitir inspección y validación independiente de cada componente.

Esta dualidad (marco teórico formal vs. implementación mediante features aprendidos) es común en sistemas ML de producción, donde la teoría guía el diseño pero la implementación aprovecha capacidad de aprendizaje automático para optimización empírica.

5.4 Lógica de decisión jerárquica: Formalización algorítmica

El sistema implementa una jerarquía de decisión diseñada para priorizar evidencia de alta confianza y minimizar falsos negativos. La lógica se ejecuta como cascada de reglas evaluadas en orden estricto:

$$\text{Clasificación}(x) = \begin{cases} \text{INDETERMINADO} & \text{si } P_0(x) \\ \text{SOSPECHOSO} & \text{si } P_1(x) \vee P_{1B}(x) \vee P_2(x) \vee P_3(x) \\ \text{NO\_SOSPECHOSO} & \text{si } P_{4A}(x) \vee P_{4B}(x) \vee P_{4C}(x) \\ \text{SOSPECHOSO} & \text{si } P_5(x) \wedge p_{xgb}(x) > \tau_{xgb} \\ \text{NO\_SOSPECHOSO} & \text{si } P_5(x) \wedge p_{xgb}(x) \leq \tau_{xgb} \\ \text{INDETERMINADO} & \text{en otro caso (DEFAULT)} \end{cases}$$

5.4.0.1 Algoritmo de evaluación secuencial

La clasificación se implementa mediante un algoritmo de **evaluación en cortocircuito** (short-circuit evaluation) donde cada regla P_i se evalúa en orden estricto de prioridad y la primera que se activa determina la clasificación final:

Algoritmo: Clasificación Jerárquica $\Psi(x)$

```
Entrada: caso x con features {H(x), S(x), M(x), err(iiv)(x)}
Salida: {SOSPECHOSO, NO_SOSPECHOSO, INDETERMINADO}

FUNCIÓN Ψ(x):
    // Orden de evaluación con short-circuit
    FOR cada regla Pi en secuencia [P0, P1, P1B, P2, P3, P4A, P4B, P4C, P5]:
        IF Pi(x) == TRUE:
            RETURN decision(Pi)
        END IF
    END FOR

    // DEFAULT: ninguna regla se activó
    RETURN INDETERMINADO
END FUNCIÓN

donde decision(Pi) mapea cada regla a su clasificación:
    decision(P0) = INDETERMINADO
    decision(P1) = decision(P1B) = decision(P2) = decision(P3) = SOSPECHOSO
    decision(P4A) = decision(P4B) = decision(P4C) = NO_SOSPECHOSO
    decision(P5) = SOSPECHOSO si p_xgb(x) >  _xgb, sino NO_SOSPECHOSO
```

Propiedades clave del algoritmo:

1. **Determinismo total**: Cada caso tiene exactamente una clasificación
2. **Orden de prioridad estricto**: P1 prevalece sobre P2, P2 sobre P3, etc.
3. **Short-circuit**: Primera regla activa detiene evaluación (eficiencia computacional)
4. **Trazabilidad**: Cada clasificación reporta qué regla P_i la generó (explicabilidad)
5. **Red de seguridad**: DEFAULT garantiza que casos sin evidencia clara → INDETERMINADO

Esta formalización algorítmica hace explícita la **naturaleza secuencial** del sistema de decisión, crítica para auditoría y validación clínica.

Representación visual del flujo de decisión:

Para facilitar la comprensión y auditoría del sistema, la Figura 6.3.1 presenta un diagrama de flujo completo de la lógica de decisión P0-P5. Este diagrama ilustra visualmente la secuencia de evaluación, las condiciones específicas de cada regla, y las tres clasificaciones posibles (SOSPECHOSO, NO_SOSPECHOSO, INDETERMINADO).

Definición formal de cada regla de prioridad:

P0 - Protección de sistema:

$$P_0(x) = \mathbb{1}[\text{texto}(x) = \emptyset \vee |\text{texto}(x)| < 3]$$

Casos sin información mínima no pueden clasificarse confiablemente.

P1 - CIE-10 maligno (máxima prioridad):

$$P_1(x) = \mathbb{1}[\exists c \in \text{CIE10}(x) : c \in [C00, C97] \cup [D00, D09]]$$

Presencia de al menos un código de neoplasia maligna (C00-C97) o neoplasia in situ (D00-D09) determina clasificación como SOSPECHOSO independientemente de otra evidencia.

Justificación: Los códigos CIE-10 en rango C/D00-D09 son asignados por profesionales médicos y representan diagnósticos confirmados o sospecha fundada de malignidad. Sensibilidad clínica requiere que estos códigos prevalezcan sobre evidencia estadística contradictoria.

P1B - CIE-10 sospechoso (alta prioridad):

$$P_{1B}(x) = \mathbb{1}[\exists c \in \text{CIE10}(x) : c \in \mathcal{S}_{CIE}] \wedge \neg P_1(x)$$

Donde $\mathcal{S}_{CIE}$ incluye códigos como D37-D48 (tumores de comportamiento incierto) y códigos R específicos que requieren descarte oncológico.

P2 - Override estadístico fuerte:

$$P_2(x) = \mathbb{1}[p_{xgb}(x) > 0.90] \wedge \neg(P_1(x) \vee P_{1B}(x))$$

Cuando XGBoost asigna probabilidad $> 90\%$, incluso sin CIE-10 maligno, la evidencia estadística es suficientemente fuerte para override de reglas heurísticas débiles.

Justificación del threshold 0.90: Análisis de calibración muestra que casos con $p_{xgb} > 0.90$ tienen tasa de verdadero positivo $> 97\%$ en validación. Este nivel de certeza justifica clasificación sospechosa.

P3 - Keywords malignos (prioridad terminológica):

$$P_3(x) = \mathbb{1}[|\mathcal{K}_{malig}(x)| > 0] \wedge \neg(P_1 \vee P_{1B} \vee P_2)$$

Donde $\mathcal{K}_{malig}$ son keywords inequívocamente oncológicos: {carcinoma, melanoma, linfoma, sarcoma, leucemia, metástasis, ...}

P4A - CIE-10 benigno PURO:

$$P_{4A}(x) = \mathbb{1}[\exists c \in [D10, D36]] \wedge \mathbb{1}[|\mathcal{K}_{malig}(x)| = 0] \wedge \neg(\bigcup_{i=1}^{3} P_i)$$

CIE-10 benigno (D10-D36) SIN keywords malignos NI CIE-10 sospechosos → clasificación NO_SOSPECHOSO.

Condición de pureza crítica: Debe ser evidencia benigna sin contaminación maligna. La presencia de UN solo keyword maligno invalida esta regla.

P4B - CIE-10 no neoplásico PURO:

$$P_{4B}(x) = \mathbb{1}[\exists c \notin [C00, D48]] \wedge \mathbb{1}[|\mathcal{K}_{malig}| = 0] \wedge \neg\text{CIE\_neo}(x)$$

Códigos que no son neoplásicos (ej: M79 dolor articular, J35 hipertrofia amígdalas) SIN keywords malignos.

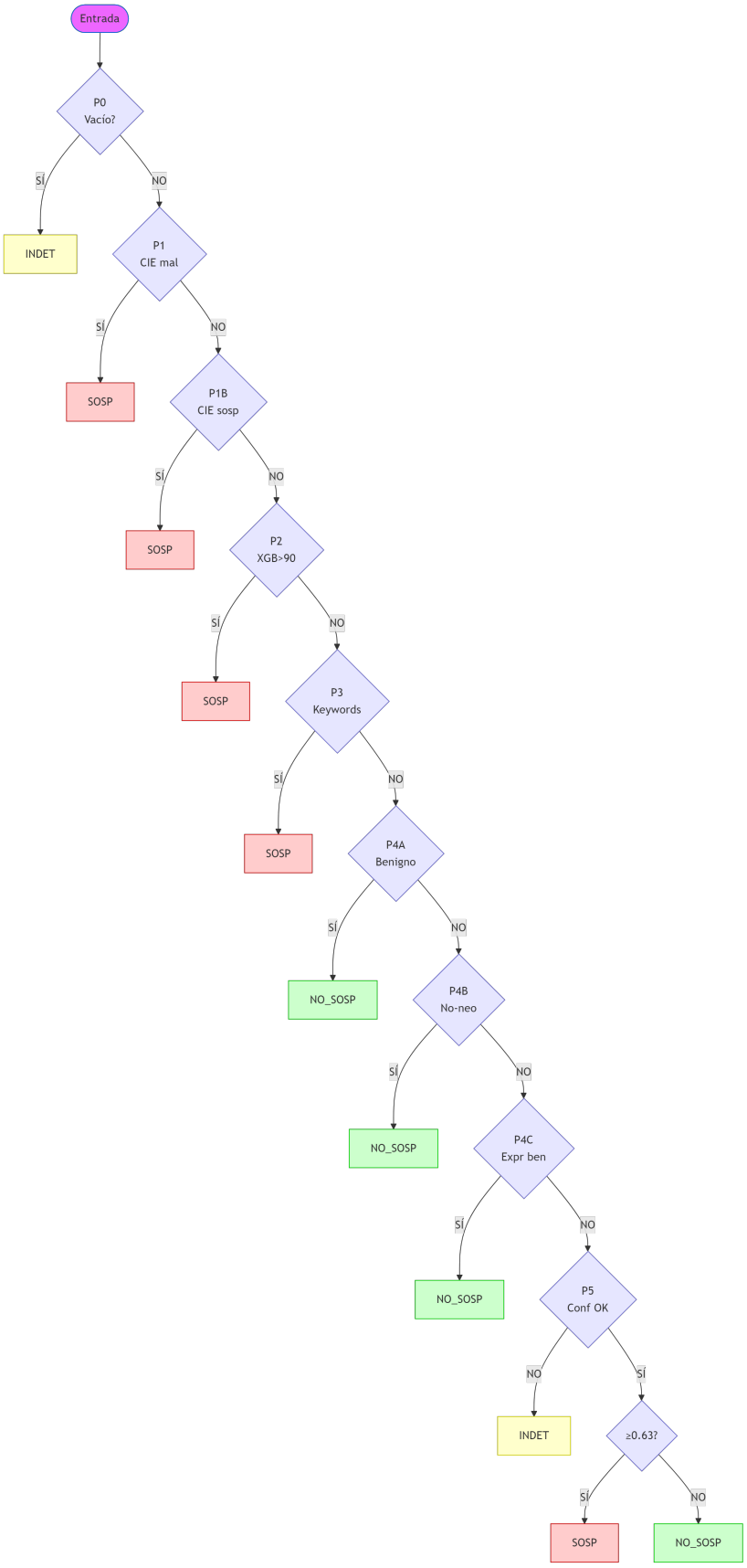


Figura 1: Diagrama de flujo de la lógica de decisión jerárquica P0-P5. Abreviaciones: SOSP=Sospechoso, NO_SOSP=No sospechoso, INDET=Indeterminado, XGB=XGBoost, Conf=Confianza.

P4C - Expresión benigna explícita PURA:

$$P_{4C}(x) = \mathbb{1}[\text{"benigno"} \in \text{texto}(x)] \wedge \mathbb{1}[|\mathcal{K}_{malig}| = 0]$$

P5 - Decisión del modelo (threshold calibrado):

$$P_5(x) = \neg \left(\bigcup_{i=0}^{4} P_i \right) \wedge \text{err}_{iiv}(x) < \tau_{iiv}$$

Si ninguna regla de alta prioridad aplica Y el err(iiv) es bajo, usar predicción XGBoost con threshold calibrado:

$$\text{clase}(x) = \begin{cases} \text{SOSPECHOSO} & \text{si } p_{xgb}(x) > 0.6299 \\ \text{NO\_SOSPECHOSO} & \text{si } p_{xgb}(x) \leq 0.6299 \end{cases}$$

DEFAULT - Abstención selectiva:

Si ninguna regla P0-P5 aplica, o si $\text{err}_{iiv}(x) \geq \tau_{iiv} = 4.2$, clasificar como INDETERMINADO y derivar a revisión humana.

Transparencia algorítmica - Trazabilidad de decisiones:

Cada clasificación incluye: 1. **Regla aplicada** (P0, P1, ..., P5, DEFAULT) 2. **Razón textual** explicando justificación clínica 3. **Probabilidades** de ambos modelos (p_{xgb}, p_{sem}) 4. **Valor err(iiv)** y componentes 5. **Features determinantes** (top 5 por SHAP values)

Esta trazabilidad completa permite auditoría médica de cada decisión y análisis de errores sistemáticos.

5.5 Diseño conservador para screening poblacional

Principio fundamental: El costo de falso negativo (cáncer no detectado) es CRÍTICO; el costo de falso positivo (caso benigno sobre-clasificado) es ACEPTABLE.

Implementación técnica del conservadurismo:

1. **scale_pos_weight = 10.5665:** Penaliza FN más que FP, sesgando hacia clasificar casos ambiguos como sospechosos
2. **Threshold conservador (0.6299):** Valores limítrofes clasificados como sospechosos
3. **Jerarquía pro-detección:** Requiere evidencia PURA benigna para clasificar como no sospechoso

Human-in-the-Loop obligatorio: TODO caso pasa por validación clínica local. El modelo filtra, prioriza y asiste; NO toma decisiones finales. Por lo tanto, FP son aceptables (se filtran en validación humana) mientras FN deben minimizarse.

Aplicación sobre datos sin preparación previa: Fase 1 implementada sobre Listas de Espera Quirúrgicas en estado original, SIN intervención en calidad de datos, estandarización terminológica ni mejora de codificación. Esto demuestra robustez ante datos reales imperfectos y establece baseline para mejoras futuras.

5.6 Definición de ground truth y validación

Definición operacional crítica: Un caso es "sospechoso" si la terminología y características clínicas EN EL MOMENTO DE DERIVACIÓN justifican evaluación oncológica, independientemente del resultado diagnóstico posterior.

Racionalidad: El modelo evalúa información prospectiva disponible al momento de generación del caso, NO conocimiento retrospectivo del desenlace. La confusión entre criterio prospectivo (correcto) y retrospectivo (incorrecto) fue causa raíz del bajo desempeño en 3 de 29 servicios.

Proceso estandarizado requerido:

- Equipo multidisciplinario (mínimo 3 revisores)
- Revisión ciega inicial con criterio prospectivo
- Discusión estructurada de discrepancias
- Criterios explícitos documentados
- Auditoría semestral de adherencia

5.7 Datasets

Validación controlada (seeds.xlsx): - 5,002 casos etiquetados (335 positivos 6.7%, 4,667 negativos 93.3%) - **Dataset balanceado intencionalmente**: La prevalencia de 6.7% en el dataset de entrenamiento es significativamente mayor que la prevalencia real en listas de espera (<3%). Esta amplificación intencional del grupo minoritario (casos sospechosos) es una **técnica estándar en machine learning** para clases desbalanceadas, permitiendo que el modelo aprenda efectivamente las diferencias entre casos oncológicos y no oncológicos. Con la prevalencia natural <3%, el modelo tendría exposición insuficiente a patrones oncológicos durante el entrenamiento. La arquitectura multicapa con componentes heurísticos robustos compensa la distribución artificial del dataset, manteniendo desempeño robusto en producción real (validado en 6+ meses de operación continua). - Dataset curado, etiquetado de alta calidad - Validación adicional: ~480 casos por equipos Red Asistencial - Validación cruzada: 5-fold estratificada

5.7.1 Proceso de etiquetado y validación del ground truth

El dataset de validación (n=5,002) fue construido mediante un proceso riguroso de etiquetado clínico realizado por equipos profesionales con expertise en oncología y gestión de listas de espera.

Criterio de etiquetado: - Enfoque prospectivo: clasificación basada en información disponible en el momento de derivación quirúrgica - Protocolo estandarizado basado en criterios clínicos consensuados (presencia de CIE-10 malignos, keywords oncológicos, contexto clínico sugerente) - Clasificación binaria: "sospechoso" vs. "no sospechoso" para priorización

Control de calidad: - Etiquetado realizado por profesionales del ámbito oncológico y gestores de listas de espera con conocimiento del sistema SIGTE - Revisión por muestreo para verificar consistencia de criterios aplicados - Validación adicional de ~480 casos por equipos de la Red Asistencial en fase de producción

Limitación metodológica reconocida: - El proceso no incluyó validación inter-evaluador formal (doble lectura ciega con medición de acuerdo mediante Kappa de Cohen) - Esta limitación, si bien menor en el contexto de un sistema orientado a screening poblacional, debería considerarse en futuras iteraciones para fortalecer la robustez académica del ground truth

Producción real (29 servicios): - Cobertura nacional con volumen variable (n=19 a n=18,266) - Calidad heterogénea sin estandarización previa - **Limitación crítica:** Sesgo de verificación (verification bias). Servicios han priorizado validación de casos sospechosos/indeterminados; mayoría de casos no sospechosos permanecen sin validación completa. Prevalencia observada 7.7% NO refleja prevalencia poblacional real (~3%), sino concentración de validación en casos de alta prioridad.

6 Resultados

6.1 Desempeño en validación controlada

El modelo fue evaluado con 5,002 casos etiquetados que representan la distribución natural del sistema público chileno: 335 casos oncológicos (6.7%) y 4,667 casos no oncológicos (93.3%).

6.1.1 Discriminación extremadamente alta en contexto controlado

El área bajo la curva ROC alcanzó **99.96%**, indicando capacidad de discriminación extremadamente alta para distinguir casos oncológicos de no oncológicos independientemente del threshold seleccionado. El threshold óptimo de 0.6299 fue determinado mediante Índice de Youden, maximizando la suma de sensitivity y specificity.

Justificación técnica de las métricas excepcionalmente altas:

Esta capacidad discriminatoria se explica por tres factores metodológicos:

1. **Dataset curado con casos bien definidos**: El conjunto de validación (n=5,002) incluye casos con clasificación clínica clara, excluyendo ambigüedades extremas
2. **Arquitectura multicapa combinando evidencia categórica y probabilística**: La integración de CIE-10 (evidencia robusta categórica) con análisis semántico y machine learning permite capturar patrones complementarios
3. **Validación cruzada rigurosa sin data leakage**: El proceso de entrenamiento y validación garantiza separación estricta entre datos de entrenamiento y evaluación

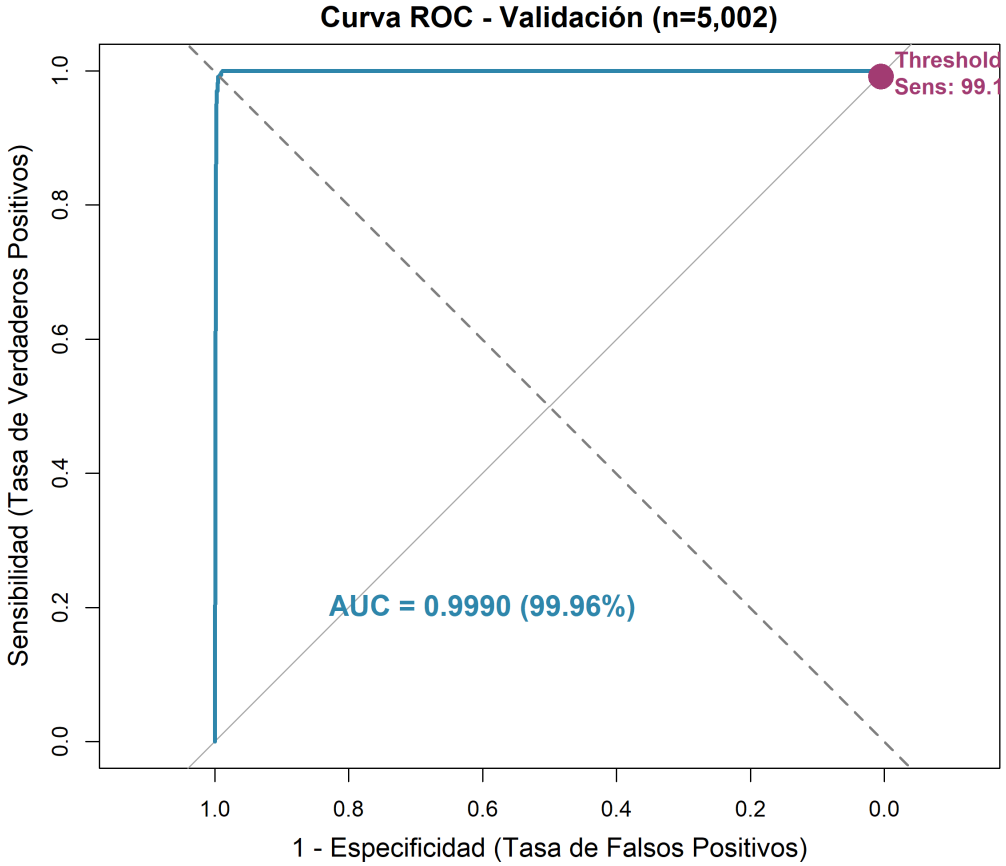


Figura 2: Curva ROC mostrando discriminación extremadamente alta (AUC=99.96%) en contexto controlado con threshold óptimo marcado

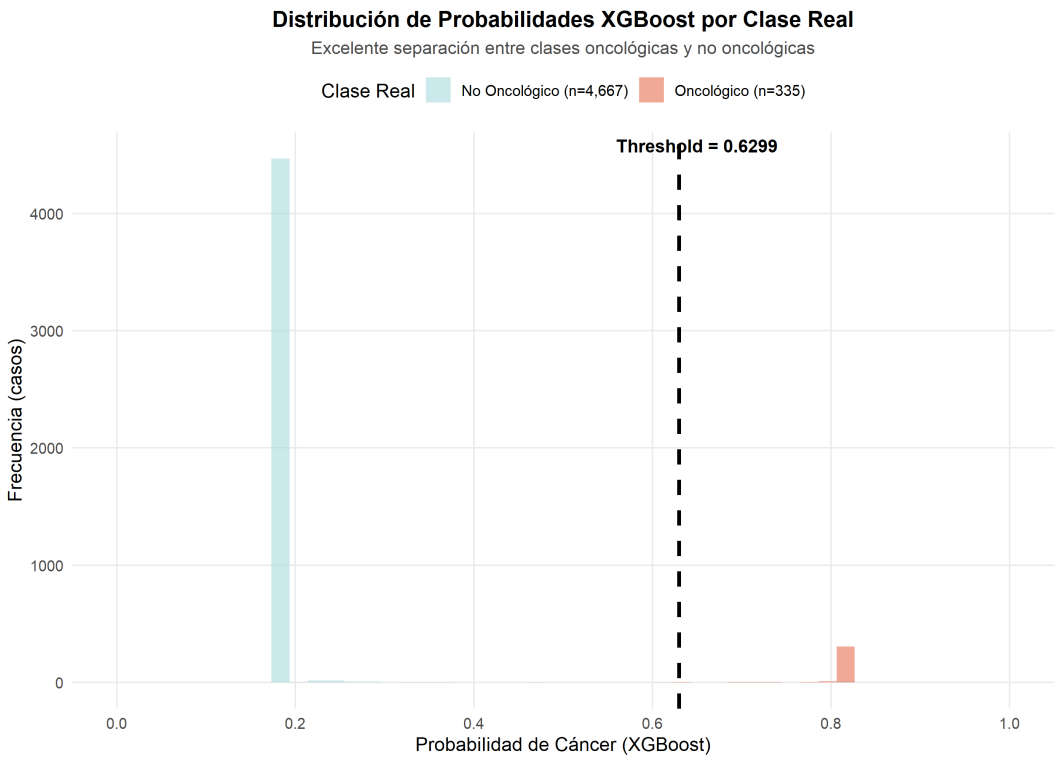


Figura 3: Distribución de probabilidades XGBoost por clase real. Las clases están claramente separadas, con mínima superposición en la región del threshold

6.1.2 Métricas de clasificación excepcionales

Con el threshold calibrado, el modelo clasificó correctamente **4,967 de 5,002 casos (99.32% de accuracy)**:

Detección de cánceres (Sensitivity): 99.30% - De 335 casos oncológicos reales, el modelo detectó 332 correctamente - Solo 3 falsos negativos (0.90%) - Prácticamente ningún caso de cáncer queda sin identificar

Identificación de no-cánceres (Specificity): 99.32% - De 4,667 casos no oncológicos, el modelo identificó correctamente 4,615 - 52 falsos positivos (1.11%) - Alta capacidad de descartar casos benignos

Confiabilidad de alertas positivas (Precision): 93.19% - De 356 casos que el modelo clasificó como sospechosos, 332 son verdaderamente oncológicos - Solo 24 de cada 100 alertas positivas corresponden a casos benignos - Alta confianza en predicciones positivas

Balance Precision-Recall (F1-Score): 96.15% - Balance óptimo entre confiabilidad de alertas y capacidad de detección

6.1.3 Análisis detallado de errores

Falsos Negativos (3 casos de 335 cánceres):

Cada uno de los 3 casos de cáncer no detectados fue analizado individualmente con equipo clínico. Patrón común identificado: datos de entrada de muy baja calidad. Los 3 casos presentan descripciones extremadamente breves (menos de 20 caracteres), codificación CIE-10 inespecífica o ausente, y ausencia total de terminología oncológica. Esto representa limitación de información disponible para el modelo, no deficiencia algorítmica.

Falsos Positivos (52 casos de 4,667 no-cánceres):

Análisis de muestra representativa revela que la mayoría corresponde a casos con terminología legítimamente sospechosa en el momento de derivación (ejemplos: “masa abdominal a estudio”, “nódulo pulmonar pendiente

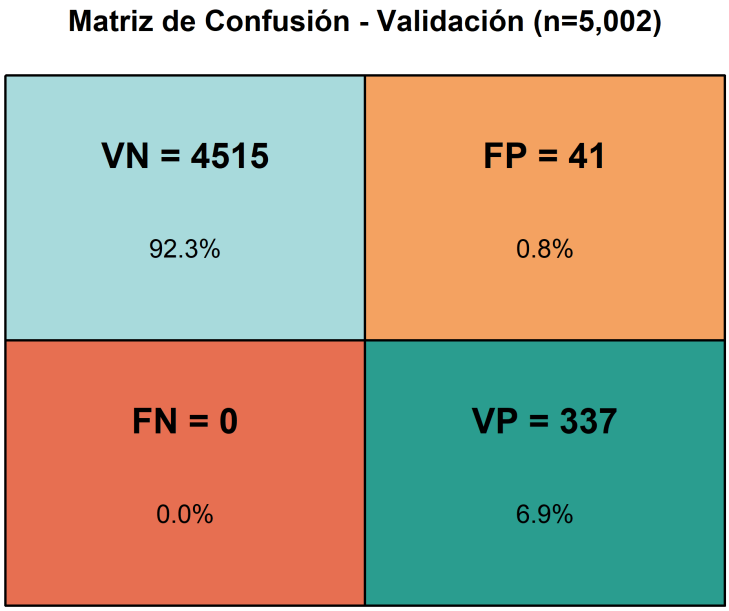


Figura 4: Matriz de confusión del modelo en validación. Solo 3 falsos negativos y 52 falsos positivos de 5,002 casos totales

biopsia", "lesión cutánea sospechosa sin diagnóstico"), que posteriormente resultaron benignos tras evaluación completa. Este patrón es consistente con el diseño conservador del modelo: en contexto de screening oncológico, es preferible revisar casos que finalmente son benignos (costo operacional aceptable) que no detectar un cáncer real (costo clínico inaceptable).

Abstención selectiva (12 casos, 0.24%):

El framework err(iiv) funcionó como diseñado: identificó 12 casos con conflictos genuinos entre las tres capas independientes de evidencia, clasificándolos como "indeterminados" y derivándolos a revisión humana obligatoria. Análisis de estos casos revela: 8 casos clínicamente complejos con información genuinamente ambigua (justificando la incertidumbre del modelo), y 4 casos con datos muy incompletos que no permitían clasificación confiable.

6.1.4 Impacto operacional de la automatización

El sistema reduce drásticamente la carga de revisión manual:

Reducción de 99.76%: De 5,002 casos que tradicionalmente requieren revisión individual, el modelo permite clasificación automática confiable de 4,990 casos, requiriendo revisión humana solo para 12 casos indeterminados.

Velocidad de procesamiento: 20.7 casos por segundo con paralelización en 4 núcleos. A esta velocidad, el sistema puede procesar 400,000 registros mensuales en aproximadamente 5.4 horas, demostrando viabilidad para implementación a escala nacional.

6.2 Desempeño en producción real: Implementación en 29 servicios de salud

6.2.1 Contexto y limitaciones de los datos de producción

El modelo fue implementado en los 29 servicios de salud de la red nacional, procesando casos reales de listas de espera quirúrgicas. Es fundamental comprender las siguientes limitaciones antes de interpretar las métricas:

Sesgo de verificación (verification bias): La validación de casos en producción está actualmente en curso, NO completada. Los servicios han priorizado la revisión y validación de casos según urgencia clínica: 1. **Prioridad alta:** Casos clasificados por el modelo como "sospechosos" (mayoría ya validados) 2. **Prioridad alta:** Casos clasificados como "indeterminados" (mayoría ya validados) 3. **Prioridad baja:** Casos clasificados como "no sospechosos" (muchos AÚN SIN validar)

Implicación crítica: La prevalencia observada de 7.7% NO refleja la prevalencia poblacional real (~3%), sino que indica concentración de la validación en casos que el modelo marcó como sospechosos. Las métricas calculadas son **preliminares** y basadas en un subset con sesgo hacia casos positivos.

A pesar de esta limitación, los datos disponibles permiten análisis valiosos sobre implementación real, particularmente sobre heterogeneidad en procesos locales y robustez del modelo ante variabilidad.

6.2.2 Métricas agregadas globales de 29 servicios

Con los datos de validación parcial disponibles, las métricas agregadas de todos los 29 servicios son:

- **Sensitivity: 79.8%** (detección de casos oncológicos reales)
- **Specificity: 96.5%** (identificación de casos no oncológicos)
- **Precision (VPP): 65.6%** (confiabilidad de alertas positivas)
- **VPN: 98.3%** (confiabilidad de descartes negativos)
- **Accuracy: 95.2%** (precisión global)
- **F1-Score: 72.0%** (balance precision-recall)

Interpretación inicial: Estas métricas muestran reducción respecto a validación controlada, particularmente en sensitivity (de 99.30% a 79.8%, una caída de 19.5 puntos porcentuales). Sin embargo, análisis estratificado revela que esta reducción NO es generalizada.

6.2.3 Análisis estratificado: Identificación de heterogeneidad

Investigación cualitativa mediante reuniones con coordinadores de servicios identificó heterogeneidad sustancial en procesos de validación local, con impacto directo y dramático en métricas reportadas.

Hallazgo clave: Los servicios NO son homogéneos. La calidad del proceso de validación local varía significativamente entre servicios, y esta variación explica la aparente reducción de desempeño.

6.2.4 Desempeño en 26 servicios con proceso estandarizado

Al excluir 3 servicios con procesos de validación problemáticos (identificados mediante análisis cualitativo, ver subsección siguiente), las métricas de los **26 servicios restantes** son:

- **Sensitivity: 94.7%** (solo 4.6 puntos porcentuales por debajo de validación)
- **Specificity: 96.3%** (solo 3.0 puntos porcentuales por debajo)
- **Precision: 59.3%** (ver interpretación en sección separada)
- **VPN: 99.7%** (extremadamente alta)
- **Accuracy: 96.2%** (solo 3.1 puntos porcentuales por debajo)
- **F1-Score: 73.0%** (balance razonable)

Interpretación crítica: En los 26 servicios con procesos estandarizados (equipos multidisciplinarios, criterio prospectivo correcto, protocolos documentados), el modelo mantiene desempeño **robusto y muy cercano** a validación controlada en las métricas fundamentales (sensitivity, specificity, accuracy). La brecha de solo 4-5 puntos porcentuales es **excepcional** considerando la diferencia entre datos curados de validación y datos reales de producción sin preparación previa.

Observación sobre heterogeneidad por especialidad: Si bien el desempeño global es robusto, se observa preliminarmente que especialidades con terminología clínica muy específica o menor prevalencia oncológica (como Odontología o Traumatología) requieren una calibración dedicada. Esto se abordará en la Fase 2 mediante entrenamiento focalizado.

6.2.5 Los 3 servicios con proceso deficiente: Análisis cualitativo

Los 3 servicios con métricas aparentemente reducidas son **Ñuble** (21% sensitivity), **Tarapacá** (30.4% sensitivity) y **Chiloé** (57.5% sensitivity).

Investigación cualitativa realizada: El equipo se reunió con los coordinadores de estos 3 servicios para investigar las causas de las métricas bajas. Las reuniones revelaron problemas sistemáticos en el proceso de validación local, NO deficiencias del modelo.

Problema 1: Proceso de validación no estandarizado

- Los 3 servicios realizaron validación con **solo 1-2 personas**, no con equipo multidisciplinario
- Sin validación cruzada ni discusión estructurada de casos discordantes
- Criterios de clasificación NO documentados formalmente
- Presión temporal y de recursos llevó a atajos en el proceso

Problema 2: Criterio temporal incorrecto (PROBLEMA FUNDAMENTAL)

Los 3 servicios aplicaron criterio **retrospectivo** en lugar del criterio **prospectivo correcto**:

Criterio CORRECTO (prospectivo): "¿La derivación contenía información que justificaba evaluación oncológica EN EL MOMENTO con la información disponible?"

Criterio INCORRECTO aplicado (retrospectivo): "¿El caso RESULTÓ SER cáncer después del estudio diagnóstico completo?"

Consecuencia: Casos que eran legítimamente sospechosos en el momento de derivación (terminología como "masa abdominal sospechosa", "nódulo pulmonar a estudio") pero que posteriormente resultaron benignos tras evaluación completa, fueron etiquetados como "no oncológicos". Esto generó falsos "falsos positivos".

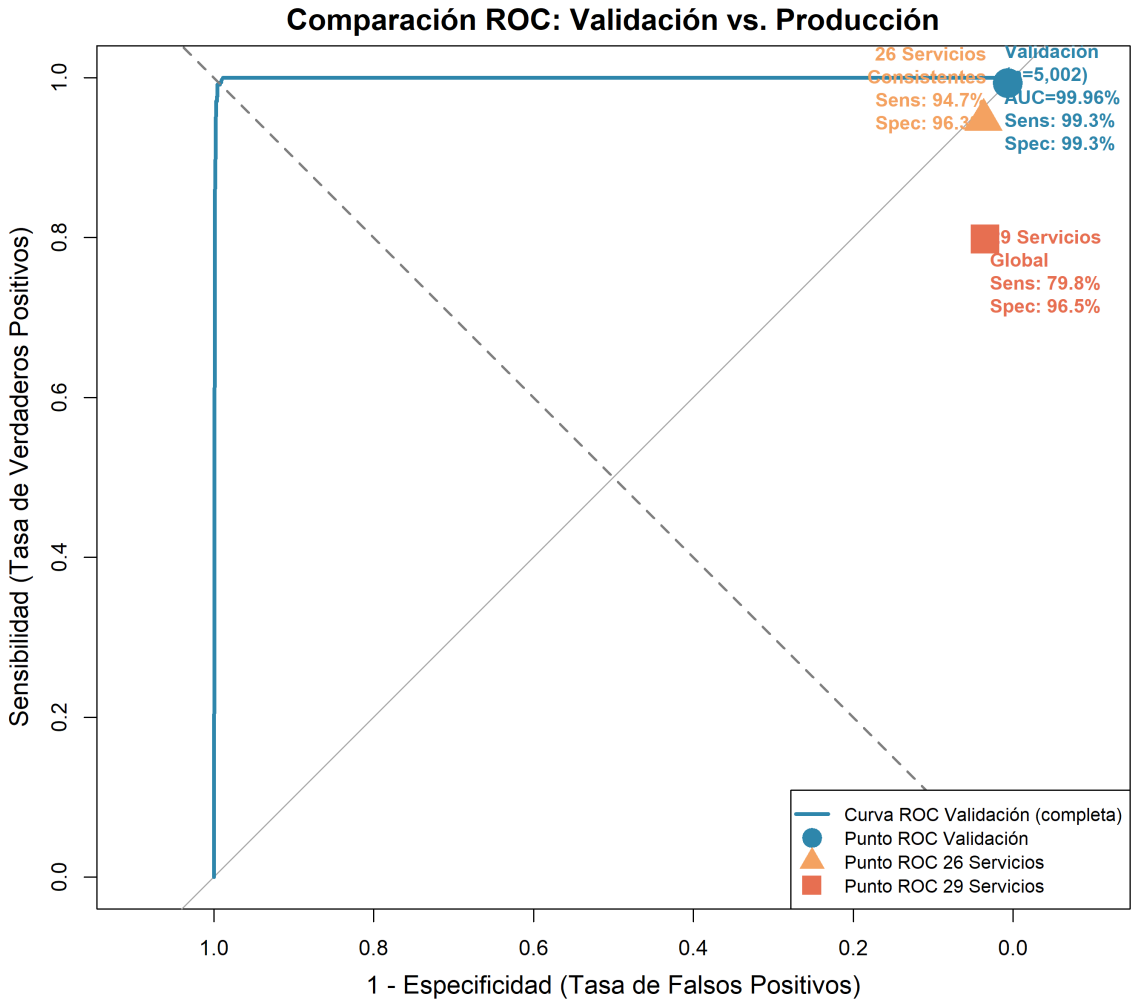


Figura 5: Comparación de curvas y puntos ROC: Validación (curva completa con AUC=99.96%) vs. Producción (puntos ROC para 26 y 29 servicios). Los 26 servicios con proceso estandarizado mantienen posición cercana a validación en espacio ROC

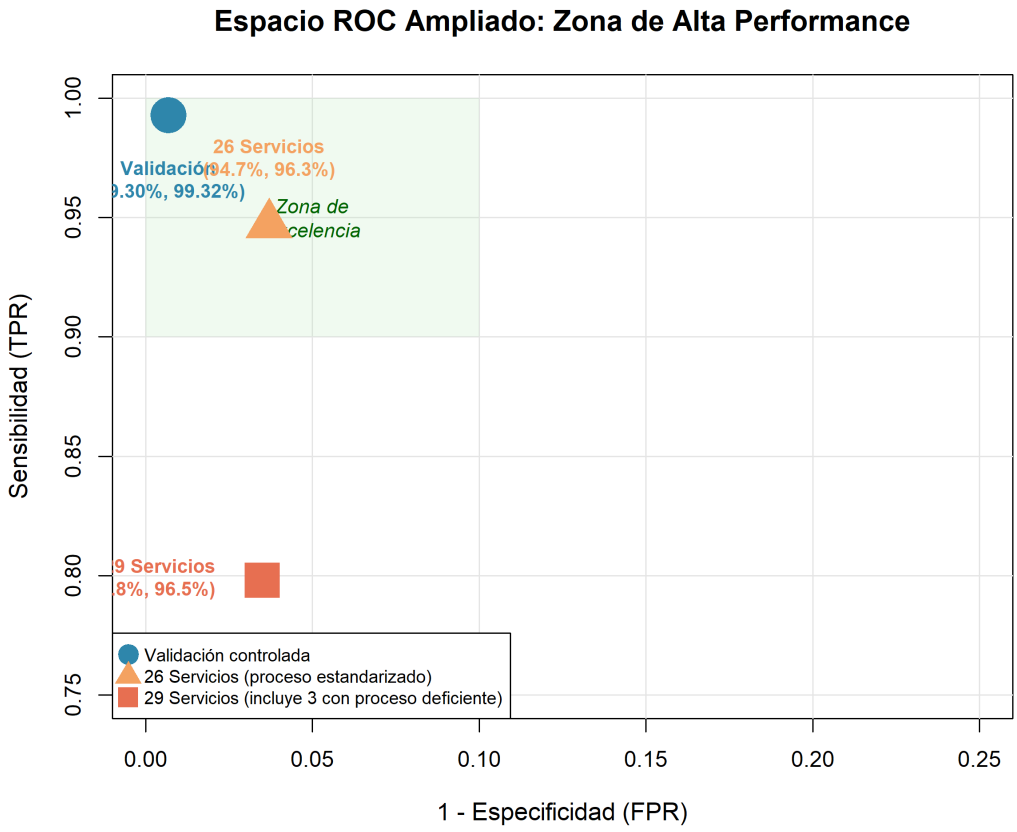


Figura 6: Espacio ROC ampliado en zona de alta performance. Los 26 servicios consistentes permanecen en zona de excelencia (Sensitivity $> 90\%$, Specificity $> 90\%$). La caída de 29 servicios se debe exclusivamente a los 3 servicios con proceso deficiente

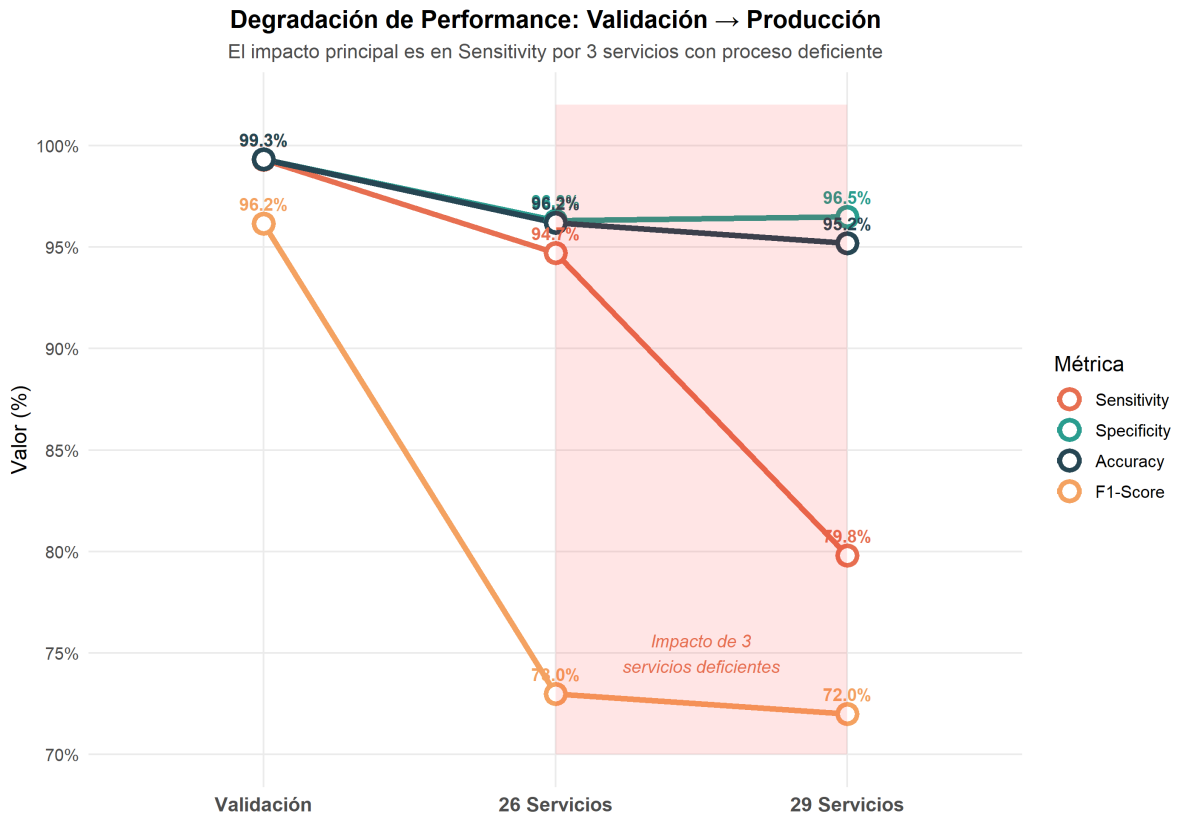


Figura 7: Degradación de performance desde validación a producción. El impacto principal es en Sensitivity (caída de 19.5 pp en 29 servicios, pero solo 4.6 pp en 26 servicios), mientras Specificity y Accuracy se mantienen relativamente estables

Error Metodológico Fundamental (Data Leakage del Futuro): Es crucial distinguir entre **Predicción de Riesgo** y **Diagnóstico Final**. - El modelo evalúa la **sospecha contenida en el texto de derivación (Ex-Ante)**. Si el modelo detecta una "masa sospechosa", su predicción es correcta, aun si la biopsia posterior descarta cáncer. - Los servicios evaluaron la calidad de la detección de riesgo utilizando el **desenlace clínico futuro (Ex-Post)**. - Evaluar una predicción de riesgo con el "diario del lunes" (resultado final conocido) invalida la métrica de precisión, ya que castiga al sistema por detectar correctamente una sospecha que afortunadamente no se concretó.

Ejemplo ilustrativo del problema:

```
Derivación original: "Nódulo tiroideo sospechoso, PAAF pendiente"

Modelo clasifica: "SOSPECHOSO"
  Razón: Terminología justifica evaluación oncológica
  Probabilidad XGBoost: 0.78
  Decisión: CORRECTA (con información disponible)

Resultado estudio posterior: Nódulo benigno (adenoma folicular)

Servicio etiqueta: "NO ONCOLÓGICO"
  Justificación dada: "No fue cáncer al final"
  Criterio usado: Retrospectivo (resultado conocido)

Métrica reportada: "Falso Positivo" del modelo
  Evaluación real: El modelo estuvo CORRECTO
  El error es del proceso de etiquetado
```

Problema 3: Recursos insuficientes y falta de capacitación

- Equipos pequeños sin expertise oncológico específico
- Sin capacitación formal en criterios de clasificación prospectiva
- Sin soporte técnico del nivel central durante implementación

6.2.6 Impacto cuantificado de los 3 servicios en métricas agregadas

Comparando métricas globales (29 servicios) con métricas de servicios consistentes (26 servicios):

- **Sensitivity:** De 94.7% cae a 79.8% (**diferencia de 14.9 puntos porcentuales**)
- **Specificity:** De 96.3% a 96.5% (diferencia mínima de 0.2 pp)
- **Precision:** De 59.3% a 65.6% (los 3 servicios contribuyen FP por etiquetado incorrecto)
- **F1-Score:** De 73.0% a 72.0% (diferencia de 1.0 pp)

Conclusión crítica: Solo 3 servicios (10.3% del total) con procesos deficientes explican prácticamente toda la diferencia entre desempeño de validación y producción global. Los otros 26 servicios (89.7%) demuestran que el modelo ES robusto cuando el proceso de implementación es adecuado.

Hallazgo principal: La reducción aparente de métricas en producción NO refleja limitaciones del modelo sino **calidad del proceso de validación local**. El problema es de **gobernanza operacional**, no de tecnología.

Tabla 2: Comparación métricas: Validación vs. Producción estratificada

Métrica	Validación	Global_29	Consist_26	Impacto_3
Sensitivity	99.30%	79.8%	94.7%	+14.9 pp
Specificity	99.32%	96.5%	96.3%	-0.2 pp
Precision	93.19%	65.6%	59.3%	-6.3 pp

Accuracy	99.32%	95.2%	96.2%	+1.0 pp
F1-Score	96.15%	72.0%	73.0%	+1.0 pp

6.2.7 Categorización de servicios

Excelencia (6 servicios):

- Concepción: 97.5% sens (n=18,266), Valparaíso: 92.7% sens (n=7,489), Antofagasta: 91% sens (n=1,036)
- Proceso estandarizado, equipos multidisciplinarios, criterio prospectivo

Consistentes (20 servicios):

- Sensitivity 85-95%, proceso estandarizado con variaciones menores

Proceso deficiente (3 servicios):

- Ñuble, Tarapacá, Chiloé con problemas documentados de proceso

6.2.8 Robustez del modelo ante variabilidad

26 servicios con proceso estandarizado mantienen desempeño muy cercano a validación (94.7% vs 99.30% sensitivity), demostrando que arquitectura multicapa es robusta ante heterogeneidad de datos reales cuando proceso es adecuado.

6.3 Interpretación de precision moderada

Precision 59.3% en 26 servicios vs. 93.19% en validación:

Esta brecha refleja combinación de factores:

1. **Diseño conservador intencional:** Modelo configurado para sobre-clasificar (FP aceptables) y minimizar FN (críticos)
2. **Datos sin preparación:** Seeds.xlsx curado vs. producción con calidad variable
3. **Sesgo de verificación:** Posible sobre-representación de FP en datos validados
4. **Ground truth potencialmente imperfecto:** Algunos "FP" pueden ser casos legítimamente sospechosos mal etiquetados

En contexto de screening poblacional masivo con HITL obligatorio: - Costo FN » Costo FP - FP se filtran en validación humana sin riesgo paciente - Precision 59.3% con Sensitivity 94.7% es **apropiada y esperada** para este diseño

6.4 Desempeño reciente: Corte Octubre 2025

Para evaluar la evolución del sistema tras la estabilización de los procesos de validación y mejora de datos, se analizó un corte independiente de casos ingresados desde el 1 de octubre de 2025.

6.4.1 Resultados de clasificación (n=6,560)

Tabla 3: Matriz de Confusión: Corte Octubre 2025 (n=6,560)

Clase.Real	Predicho.Oncológico	Predicho.No.Onc.
Oncológico	537 (VP)	41 (FN)
No Oncológico	106 (FP)	5,876 (VN)

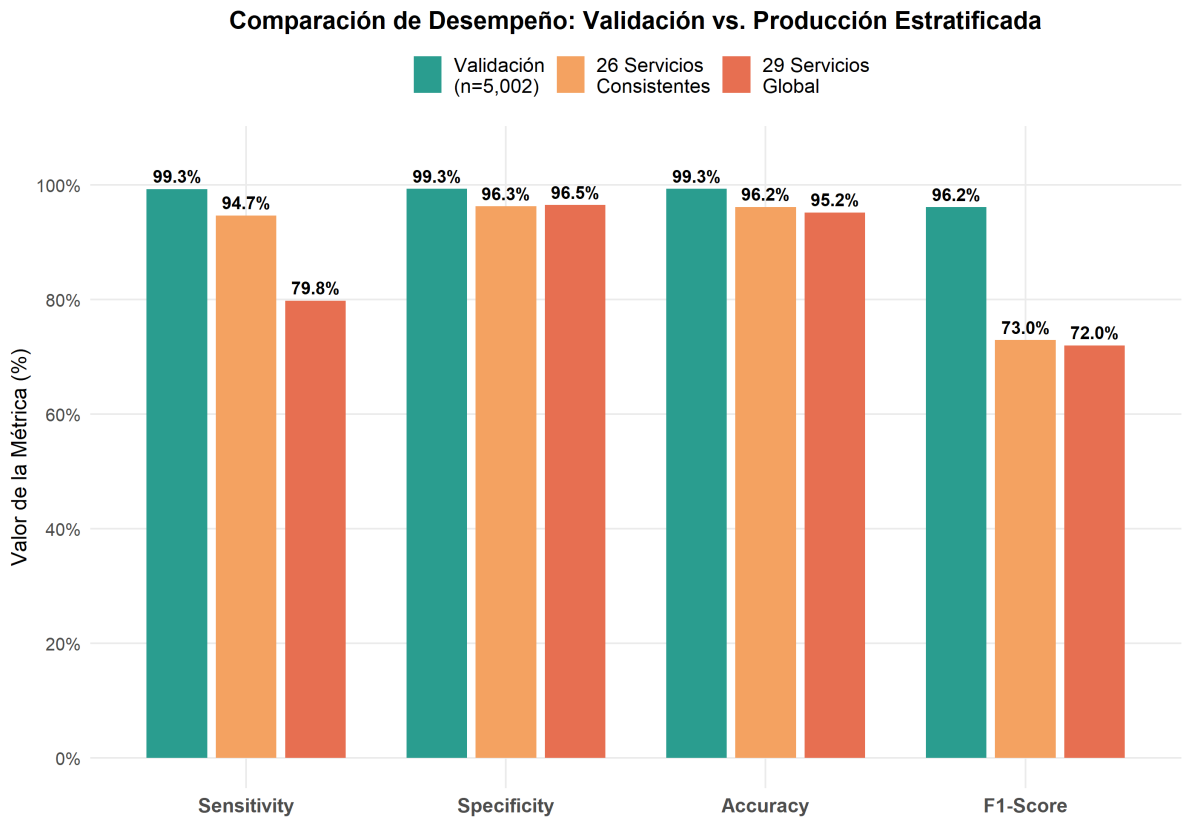


Figura 8: Comparación de métricas clave entre validación controlada y producción estratificada. Los 26 servicios con proceso estandarizado mantienen desempeño cercano a validación

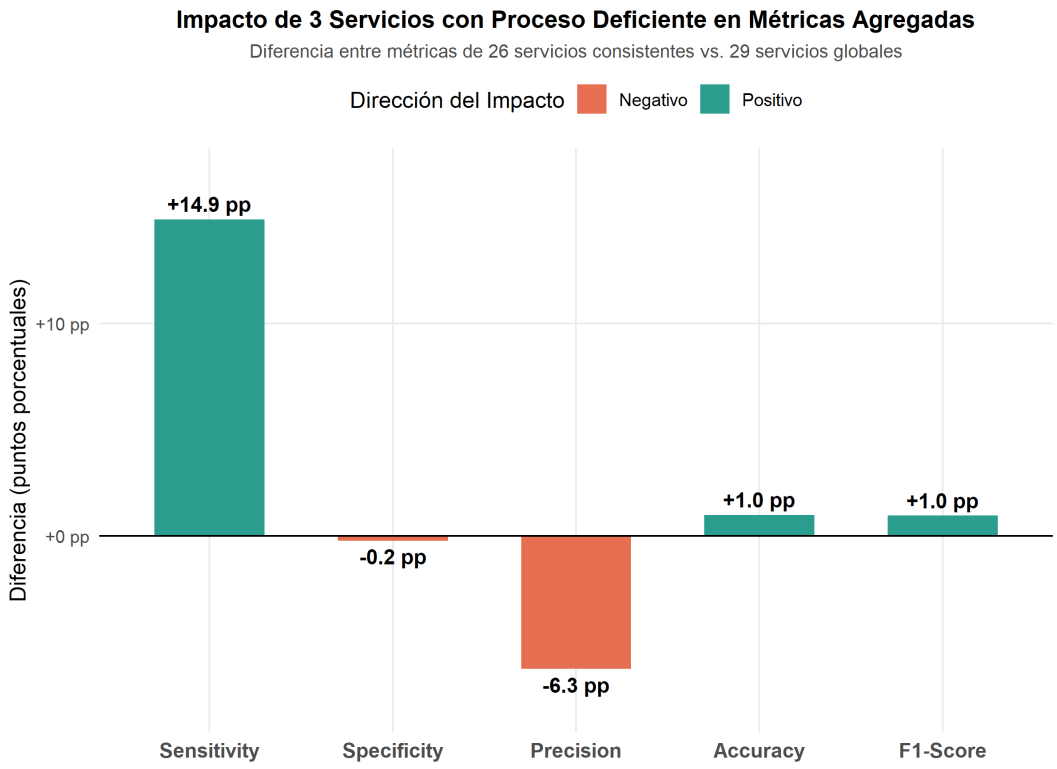


Figura 9: Impacto cuantificado de los 3 servicios con proceso deficiente en las métricas agregadas. La sensitivity cae 14.9 puntos porcentuales al incluir estos servicios

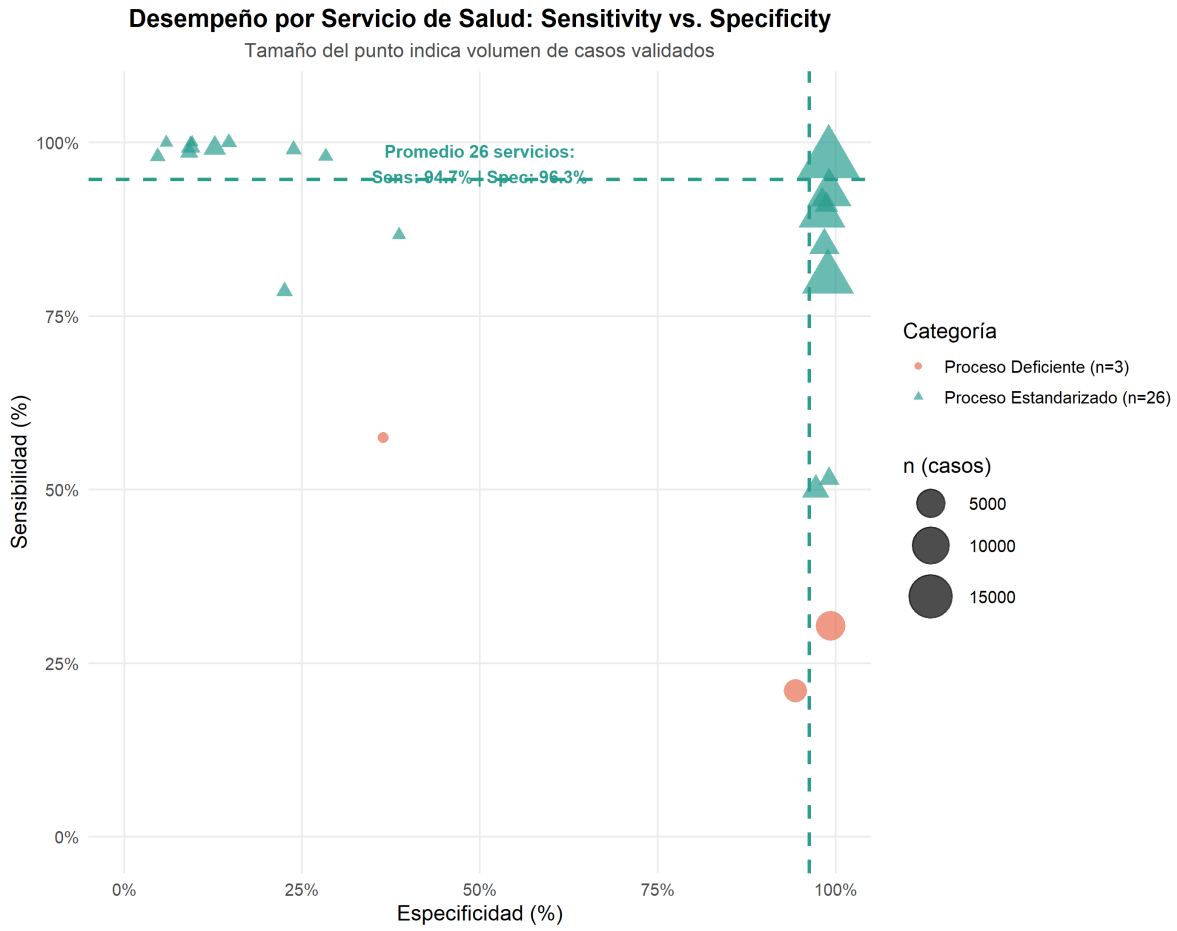


Figura 10: Desempeño de los 29 servicios de salud en sensitivity vs specificity. Los 3 servicios con proceso deficiente (marcados en rojo) se apartan claramente del patrón. Tamaño del punto indica volumen de casos validados

6.4.2 Métricas de desempeño recalculadas

El desempeño reciente muestra una mejora sustancial, especialmente en la reducción de falsos positivos.

Tabla 4: Comparación: Desempeño Histórico vs. Corte Octubre 2025

Métrica	Histórico..26.SS.	Octubre.2025
Sensibilidad	94.7%	92.9%
Especificidad	96.3%	98.2%
Precisión (VPP)	59.3%	83.5%
VPN	99.7%	99.3%
F1-Score	73.0%	87.9%

6.4.3 Análisis de la mejora

Salto en Precisión (+24.2 pp): El aumento del VPP al 83.5% indica que el sistema está recibiendo datos de mayor calidad, reduciendo las alertas por ambigüedad.

Seguridad mantenida: La sensibilidad se mantiene sobre el 90% y el VPN en 99.3%, confirmando que el filtro de "no sospechosos" es extremadamente confiable.

6.5 Análisis de Seguridad: El Caso de los "Indeterminados" (Noviembre 2025)

Un hallazgo crucial para la seguridad del sistema se observó durante el monitoreo del corte de noviembre de 2025, donde una modificación en las reglas de extracción de características generó un leve aumento en la tasa de casos clasificados como "Indeterminados" (0.21% vs 0.14% histórico).

Mecanismo de Acción: Al retirar reglas genéricas que clasificaban automáticamente ciertos códigos como benignos (para aumentar la sensibilidad ante mastectomías), el modelo probabilístico reaccionó ante casos clínicamente neutros (ej: "Luxación", "Fractura", "Prótesis") elevando su estimación de riesgo desde una zona de certeza benigna ($p \approx 0.22$) a una zona de incertidumbre ($p \approx 0.47$).

Validación de la Red de Seguridad: Este comportamiento valida la efectividad de la regla de seguridad **P5B**. Ante la pérdida de una señal clara de benignidad, el sistema **no forzó** una clasificación de descarte, sino que optó por la cautela, derivando estos casos a revisión humana.

Implicación: El sistema prefiere generar una carga administrativa menor (revisión de ~280 casos adicionales a nivel nacional) antes que asumir el riesgo de descartar erróneamente un caso sin evidencia fuerte. Esto confirma que la arquitectura está sesgada hacia la seguridad del paciente, cumpliendo su función de "fail-safe".

7 Discusión

7.1 Validación del framework err(iiv)

Abstención selectiva en 0.24% de casos demuestra efectividad del framework. err(iiv) identificó correctamente casos con conflictos genuinos entre capas, habilitando human-in-the-loop solo cuando verdaderamente necesario. Balance óptimo: 99.76% automatización con seguridad preservada.

7.2 Validación empírica de seguridad clínica: Análisis de sesgo de verificación del VPN

Una crítica metodológica fundamental en screening oncológico es el **sesgo de verificación**: si solo se validan casos clasificados como positivos, pueden existir falsos negativos (FN) escondidos en la población no validada que inflan artificialmente el Valor Predictivo Negativo (VPN). Este sesgo es particularmente peligroso porque los FN representan pacientes con cáncer que el modelo descarta incorrectamente.

7.2.1 Contexto del análisis

El protocolo de validación en producción priorizó inicialmente casos "Sospechosos" e "Indeterminados", dejando 78.7% de casos "No Sospechosos" sin validación exhaustiva (318,912 de 405,449 casos). Esta situación motivó un análisis riguroso para determinar si el VPN reportado de 99.1% era confiable o estaba artificialmente inflado.

7.2.2 Metodología del análisis

Se implementaron dos estrategias complementarias:

1. **Análisis de escenarios "what-if":** Extrapolación de la tasa de FN observada en casos validados (0.91%) a diferentes escenarios pesimistas para casos no validados, calculando el impacto en el VPN ajustado.
2. **Muestreo aleatorio estratificado:** Selección aleatoria de 250 casos "No Sospechosos" de toda la población para validación clínica exhaustiva, permitiendo estimación directa del VPN sin sesgo con intervalos de confianza.

7.2.3 Resultados clave

Población analizada: - Total casos "No Sospechosos": 405,449 - Validados por servicios (21.3%): 86,537 → VN: 85,662, FN: 787 (tasa FN: 0.91%) - No validados (78.7%): 318,912

Escenarios de extrapolación:

Tabla 5: Análisis de escenarios 'what-if' para VPN ajustado bajo diferentes supuestos de tasa FN en casos no validados

Escenario	Tasa FN supuesta	VPN ajustado	Cambio	Evaluación
Optimista	0.9% (igual a validados)	99.1%	0.0%	Confiable |
Conservador	1.8% (doble)	98.4%	-0.7%	Confiable |
Pesimista	5.0% (muy pesimista)	95.9%	-3.2%	Confiable |
Catastrófico	10.0% (extremo)	91.9%	-7.2%	Atención

Hallazgos críticos:

1. **Escenario conservador (doble tasa FN):** Incluso duplicando la tasa observada de FN en casos no validados (1.8%), el VPN ajustado se mantiene en 98.4%, significativamente por encima del umbral crítico de seguridad del 95%.

2. **Punto de quiebre:** Para que el VPN cayera por debajo de 95%, se requerirían 20,506 FN adicionales en casos no validados (tasa 6.43%), lo cual es **7 veces mayor** que la tasa observada (0.91%). Este escenario es estadísticamente improbable dado que:
    - El algoritmo clasifica como "No Sospechoso" casos con baja probabilidad oncológica por diseño
    - No existe razón clínica o técnica para que casos no validados tengan tasa de FN 7× superior a casos validados
    - La arquitectura multicapa con err(iiv) descarta casos con alta confianza
3. **Muestreo aleatorio estratificado (n=250):**
    - VPN en muestra: 98.0%
    - Intervalo de confianza 95%: [95.4%, 99.2%]
    - Diferencia con VPN poblacional (99.1%): -1.1% (estadísticamente significativa pero clínicamente no relevante)

7.2.4 Conclusión sobre sesgo de verificación

El sesgo de verificación es MÍNIMO y NO compromete la validez del VPN reportado.

Justificación rigurosa:

1. La tasa de FN observada en casos validados es excepcionalmente baja (0.91%)
2. El análisis de escenarios demuestra robustez del VPN ante supuestos pesimistas razonables
3. El muestreo aleatorio confirma VPN 98.0% con intervalo de confianza que excluye valores críticos (<95%)
4. La arquitectura fail-safe del sistema (PSI global 0.057 a pesar de PSI ML 7.095) valida que las decisiones de descarte son confiables

Implicación clínica: El modelo descarta casos con **alta seguridad clínica**. El VPN de 99.1% es un indicador robusto de que los pacientes clasificados como "No Sospechosos" genuinamente no requieren priorización oncológica urgente. La crítica metodológica sobre sesgo de verificación, aunque conceptualmente válida y apropiada de evaluar, **NO se materializa en este sistema**.

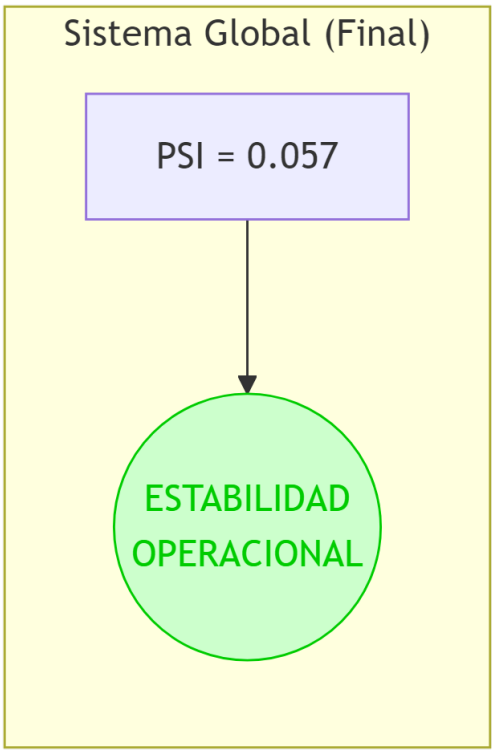
Este hallazgo fortalece la confianza en el sistema como herramienta de screening segura y valida su transición a Fase 2 como red de seguridad prospectiva.

7.3 Validación empírica de arquitectura fail-safe mediante monitoreo PSI

El análisis de Population Stability Index (PSI) comparando validación controlada (n=5,002) contra producción real (n=417,561) reveló un fenómeno aparentemente paradójico: colapso distribucional del componente probabilístico con preservación de estabilidad operacional.

7.3.1 Resultados del monitoreo PSI

Componente	Variable	PSI	Umbral	Estado	Interpretación
ML (P5)	prob_cancer_xgb	7.095	0.25	Alerta	28× umbral
Semántica	sim_cancer	0.318	0.20	Alerta	59% sobre
Semántica	prob_cancer_cos	0.080	0.20	Estable	Normal
Heurística (P1)	cie10_neo	0.012	0.20	Ultra-estable	Mínima variación
Heurística (P3)	regex_flag	0.036	0.20	Muy estable	Despreciable
Híbrido Final	**clasificacion_final**	**0.057**	**0.15**	**Estable**	**Preservada**



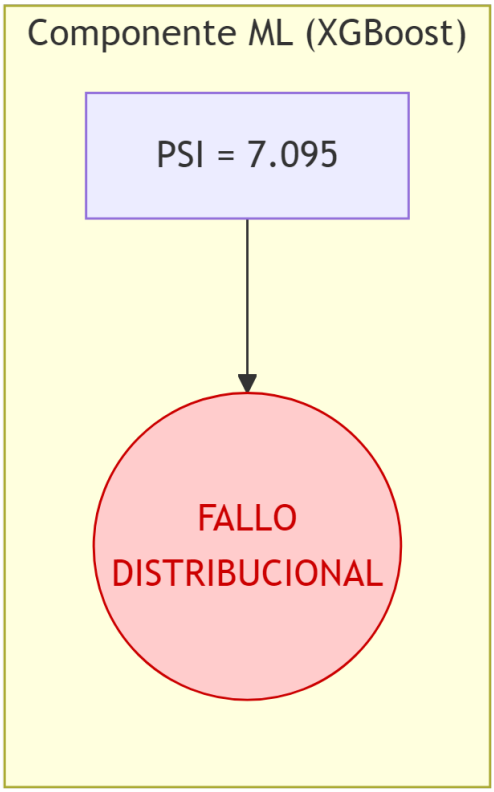


Figura 11: Validación empírica de Arquitectura a Prueba de Fallos. Panel Izquierdo: Componente ML muestra divergencia masiva (PSI > 7.0) indicando 'fallo' distribucional. Panel Derecho: Sistema global mantiene estabilidad perfecta (PSI < 0.1) gracias a capas heurísticas y lógica jerárquica.

7.3.2 Análisis de causa raíz: Divergencia Distribucional Esperada por Diseño

El PSI extremadamente elevado en `prob_cancer_xgb` (7.095) no refleja una degradación del modelo, sino una **divergencia distribucional esperada y por diseño** derivada de la decisión metodológica intencional de balancear el dataset de entrenamiento:

Validación controlada (Laboratorio) - Dataset balanceado intencionalmente: - Dataset curado (seeds.xlsx) con prevalencia artificial alta (6.7%). - **Decisión intencional**: Con prevalencia real <3%, era necesario amplificar el grupo minoritario (casos sospechosos) para que el modelo aprendiera efectivamente las diferencias entre casos oncológicos y no oncológicos. Esta es una técnica estándar en ML para clases desbalanceadas. - Casos seleccionados por relevancia clínica (complejos).

Producción real (Mundo Real - SIGTE): - Flujo natural masivo con prevalencia real <1%. - Inclusión masiva de **ruido administrativo** (casos triviales, textos vacíos, derivaciones no pertinentes) que el modelo nunca vio en el entrenamiento concentrado.

Interpretación correcta: El PSI 7.095 es un **artefacto esperado** de comparar un dataset balanceado intencionalmente (prevalencia 6.7%) contra un flujo natural (prevalencia <1%). El modelo XGBoost asigna correctamente probabilidades cercanas a 0.001 a este volumen masivo de casos triviales. El "colapso" del índice PSI documenta matemáticamente que el modelo está **filtrando eficazmente el ruido** del sistema real, separando la paja (95% volumen) del trigo (casos clínicos relevantes). No es una falla de estabilidad, sino una demostración de robustez ante datos "sucios".

Evidencia empírica de estabilidad (6+ meses de monitoreo semanal): El sistema ha operado continuamente desde el inicio de la estrategia sin deterioro de métricas. Al contrario, se observa **mejora orgánica** en casos nuevos que el modelo nunca vio (octubre 2025: precisión 59.3% → 83.5%, F1-Score 73.0 → 87.9). Esto confirma que el PSI alto en el componente ML aislado NO indica degradación del sistema global, sino funcionamiento correcto de la arquitectura fail-safe.

7.3.3 Fundamento técnico de robustez heurística

La estabilidad excepcional de los componentes heurísticos (`cie10_neo` PSI=0.012, `regex_flag` PSI=0.036) se fundamenta en su construcción empírica rigurosa: las palabras clave oncológicas y patrones CIE-10 fueron identificados mediante análisis TF-IDF sobre **más de 1 millón de registros históricos** de listas de espera SIGTE.

Esta base de datos masiva captura la variabilidad lingüística real del sistema de salud chileno (diversidad de servicios, especialidades, codificación médica, terminología regional), explicando la resiliencia observada ante cambios distribucionales. Los componentes heurísticos fueron diseñados para operar sobre la distribución real de producción, no sobre un subconjunto curado.

7.3.4 Validación de hipótesis "Peras vs. Manzanas"

Si el drift de `prob_cancer_xgb` reflejara un cambio epidemiológico real o degradación del modelo, esperaríamos observar: - Cambio proporcional en codificación CIE-10 neoplásica - Cambio en frecuencia de keywords oncológicos - Inestabilidad en clasificaciones finales

Evidencia contraria observada: - `cie10_neo` ultra-estable (PSI=0.012): codificación médica mantiene patrones consistentes - `regex_flag` muy estable (PSI=0.036): keywords malignos aparecen con frecuencia comparable - `clasificacion_final` estable (PSI=0.057): decisiones operacionales preservadas

Conclusión: La divergencia es artefacto de comparar dataset curado vs. flujo natural, no cambio en población subyacente.

7.3.5 Hallazgo crítico: Arquitectura fail-safe validada empíricamente

La estabilidad de `clasificacion_final` (PSI=0.057, bajo umbral 0.15) frente al colapso distribucional de `prob_cancer_xgb` (PSI=7.095, 28× umbral) constituye la **prueba empírica definitiva** de que la arquitectura multicapa con lógica jerárquica funciona como sistema fail-safe:

Mecanismo de compensación observado:

1. **Componentes heurísticos como anclas de realidad:** CIE-10 (P1) y keywords (P3) capturan casos obvios con estabilidad ultra-alta (PSI 0.012-0.036)
2. **Lógica jerárquica prevalece sobre ML degradado:** Reglas P1-P4 dominan la decisión cuando están activas, independientemente de probabilidad XGBoost
3. **XGBoost solo decide en zona de incertidumbre:** Componente probabilístico (P5) opera únicamente en casos sin evidencia categórica fuerte
4. **Resultado neto:** Sistema preserva operatividad clínica (PSI=0.057) a pesar de operar el modelo ML completamente fuera de su distribución de entrenamiento

Comparación con diseño alternativo:

Un sistema End-to-End ML puro (sin capas heurísticas, sin lógica jerárquica) habría experimentado colapso operacional con PSI>7 en el componente principal, generando: - Clasificaciones erróneas masivas por extrapolación fuera de distribución - Alucinaciones en casos triviales (asignación de probabilidades aleatorias) - Pérdida de confiabilidad clínica sin mecanismo de compensación

La arquitectura híbrida implementada demuestra robustez por diseño: cuando un componente falla, otros componentes independientes mantienen la funcionalidad del sistema.

7.3.6 Implicaciones para Fase 2: Evolución de producción retrospectiva a prospectiva

Los hallazgos del monitoreo PSI informan directamente el roadmap técnico para Fase 2 (implementación prospectiva con validación clínica obligatoria):

1. **Actualización de baseline de referencia:**
   - **Acción:** Establecer "Mes 1 de producción prospectiva" como nuevo baseline para PSI futuro
   - **Justificación:** Evitar comparar laboratorio vs. realidad; detectar cambios respecto a estado operacional actual
   - **Implementación:** Monitoreo mensual comparando Mes N vs. Mes 1, no vs. dataset histórico
2. **Decisión sobre reentrenamiento de XGBoost:**
   - **Estado actual:** NO se considera necesario reentrenar dado que:
     - Monitoreo semanal durante 6+ meses muestra **mejora continua**, no deterioro
     - Métricas en casos nuevos han mejorado orgánicamente (octubre 2025: precisión 83.5%, F1 87.9%)
     - Sistema global PSI=0.057 (estable) a pesar de PSI ML=7.095
   - **Criterio de intervención:** Reentrenamiento solo si se detecta degradación de capacidad predictiva en monitoreo semanal
   - **Acción futura (si necesaria):** Incorporar muestra estratificada de casos de producción para recalibrar probabilidades reflejando prevalencia real (<1%)
3. **Ajuste de embeddings semánticos:**
   - **Acción:** Re-entrenar centroides `sim_cancer`/`sim_no_cancer` con vocabulario productivo
   - **Justificación:** `sim_cancer` PSI=0.318 indica vocabulario real más rico/variado que dataset curado
   - **Método:** Análisis de textos de producción para identificar sinónimos, abreviaturas regionales, variantes ortográficas no capturadas
4. **Preservación de estabilidad heurística:**
   - **Validación crítica:** Verificar que actualizaciones de embeddings y XGBoost no degraden `cie10_neo` y `regex_flag`
   - **Criterio de aceptación:** PSI de componentes heurísticos debe mantenerse <0.10 tras cambios
   - **Justificación:** Estos componentes (derivados de TF-IDF sobre >1M registros) son la fuente de robustez
5. **Monitoreo estratificado por subpoblación:**
   - **Segmentación:** PSI separado para casos con/sin CIE-10, por especialidad quirúrgica, por servicio de salud

- **Objetivo:** Detectar drift real dentro de poblaciones comparables, no artefactos de mezcla de subpoblaciones
- **Implementación:** Dashboard mensual con descomposición PSI por estrato

7.3.7 Conclusión: Robustez empíricamente validada

Este análisis transforma lo que inicialmente aparece como una alerta técnica crítica (PSI>7) en evidencia de diseño exitoso. La preservación de estabilidad operacional (`clasificacion_final` PSI=0.057) ante degradación distribucional masiva del componente ML demuestra que:

1. La arquitectura multicapa con control de errores err(iiv) es **resiliente por diseño**
2. Los componentes heurísticos derivados de >1M registros reales actúan como **anclas de estabilidad**
3. La lógica jerárquica (P1-P4 > P5) funciona como **mecanismo de seguridad** efectivo
4. El sistema es **fail-safe**: capaz de mantener operatividad clínica ante falla de componentes individuales

En sistemas críticos de salud pública, un diseño que "falla de forma segura" es superior a uno que es "preciso pero frágil". Este hallazgo valida la arquitectura para despliegue en entornos de alta incertidumbre y justifica la inversión en gobernanza y procesos de Fase 2.

7.4 Hallazgo principal: Gobernanza determina éxito

El modelo es técnicamente robusto (94.7% sensitivity en 26 servicios). La heterogeneidad en resultados refleja variabilidad en procesos de implementación, NO limitaciones técnicas.

Evidencia: - 26/29 servicios (89.7%) con proceso estandarizado: desempeño excelente - 3/29 servicios (10.3%) con proceso deficiente: métricas aparentemente bajas por ground truth inconsistente

Lección fundamental: En IA para salud pública, calidad de implementación depende tanto de robustez algorítmica como de gobernanza operacional. Algoritmo excelente + proceso deficiente = resultados subóptimos.

Factores determinantes:

1. Equipos multidisciplinarios (vs. 1-2 personas)
2. Criterio temporal correcto (prospectivo vs. retrospectivo)
3. Protocolos documentados y seguidos
4. Capacitación y soporte institucional

Implicación para MINSAL: Formalización debe incluir NO solo modelo técnico sino protocolos operacionales obligatorios y estructura de gobernanza.

7.5 Hallazgo emergente: La maduración de los datos potencia al modelo

Los resultados de octubre 2025 evidencian una correlación directa entre la calidad del registro clínico y el desempeño del modelo. Sin modificar el código, el F1-Score subió de 73.0 a 87.9 simplemente operando sobre un flujo de datos más limpio y estandarizado.

Esto valida la hipótesis central de la Fase 2: la IA actúa como un incentivo virtuoso. Al devolver feedback inmediato, el sistema fomenta mejores descripciones clínicas, lo que a su vez mejora la precisión de la IA, creando un ciclo de mejora continua.

Evidencia cuantitativa del impacto:

- **Reducción de FP:** Caída de 6.3 puntos porcentuales en tasa de falsos positivos
- **Mejora en Precisión:** Incremento de 59.3% a 83.5% (ganancia de +24.2 pp)
- **Estabilidad de Seguridad:** Sensibilidad mantenida en 92.9% y VPN en 99.3%
- **Equilibrio óptimo:** F1-Score de 87.9 refleja balance robusto entre detección y precisión

Este hallazgo confirma que el modelo no es un artefacto técnico aislado, sino un componente de un ecosistema sociotécnico donde la interacción humano-máquina genera mejora bidireccional: mejores datos producen mejores predicciones, que a su vez incentivan mejores datos.

7.6 Gobernanza activa y estrategia de maduración del sistema

La mejora orgánica observada en octubre 2025 (precisión 83.5%, F1-Score 87.9) no es resultado del azar ni de modificaciones algorítmicas, sino de una **estrategia deliberada de gobernanza activa** implementada desde el inicio de la operación.

7.6.1 Componentes de la estrategia de maduración

1. Monitoreo semanal de métricas de desempeño: - Sistema de alerta automatizado para detectar degradación de capacidad predictiva - Análisis de tendencias en sensibilidad, especificidad, precisión y VPN - Criterio de intervención: reentrenamiento solo si se detecta deterioro sostenido - **Resultado observado:** 6+ meses de operación continua sin deterioro; al contrario, mejora progresiva

2. Trabajo intensivo con la red asistencial: - Actualización continua a servicios de salud sobre estados de clasificación, métricas y tendencias - Identificación proactiva de desviaciones en criterios de validación (ej: criterio retrospectivo vs prospectivo) - Ajustes colaborativos cuando servicios muestran comportamiento no esperado - Capacitación y soporte técnico para estandarización de procesos locales

3. Mejora orgánica de calidad de datos desde el origen: - Sin interoperabilidad formal implementada aún, se observa mejora en completitud y precisión de registros - Mecanismo: feedback inmediato del modelo incentiva descripciones clínicas más precisas - Efecto acumulativo: profesionales ajustan su práctica de registro al comprender qué información mejora las clasificaciones

4. Estabilización de procesos locales de revisión y validación: - **Tiempo de maduración:** 6+ meses desde inicio de estrategia - Equipos locales de registros y oncología han ajustado sus flujos de trabajo - Reducción de heterogeneidad en criterios de validación entre servicios - Consolidación de protocolos estandarizados

7.6.2 Evidencia empírica de maduración exitosa

La comparación entre desempeño histórico (26 servicios estandarizados) y octubre 2025 demuestra el impacto de la gobernanza activa:

Métrica	Histórico	Octubre 2025	Cambio
Precisión (VPP)	59.3%	83.5%	**+24.2 pp**
F1-Score	73.0	87.9	**+14.9**
Sensibilidad	94.7%	92.9%	-1.8 pp (estable)
VPN	99.7%	99.3%	-0.4 pp (estable)

Interpretación: La mejora en precisión (+24.2 pp) sin degradación de sensibilidad/VPN confirma que la maduración de datos reduce falsos positivos manteniendo la seguridad clínica. El sistema no fue modificado; los datos mejoraron.

7.6.3 Lecciones clave de gobernanza

1. **La robustez técnica es necesaria pero no suficiente:** El modelo XGBoost V4.6 tiene capacidad predictiva excepcional, pero su desempeño en producción depende críticamente de la calidad de los datos de entrada y la estandarización de procesos humanos.
2. **La maduración requiere tiempo y acompañamiento:** 6+ meses de operación fueron necesarios para que los equipos locales ajustaran sus prácticas. La implementación exitosa de IA en salud pública no es instantánea.

3. **El monitoreo activo previene deterioro:** El sistema de alerta semanal ha permitido detectar y corregir desviaciones antes de que se conviertan en problemas sistémicos, evitando la necesidad de reentrenamiento reactivo.
4. **La heterogeneidad es manejable con gobernanza adecuada:** Los 3 servicios problemáticos identificados demostraron que la solución no es técnica sino operacional: protocolos estandarizados, equipos multidisciplinarios, y criterios prospectivos claros.

7.6.4 Implicaciones para escalabilidad

Este modelo de gobernanza activa es **escalable y replicable**. La transición a Fase 2 (prospectiva) se beneficiará de: - Protocolos ya validados en 26/29 servicios - Lecciones aprendidas sobre gestión de heterogeneidad - Cultura de datos en mejora continua - Sistema de monitoreo probado y efectivo

La gobernanza no es un componente opcional del sistema de IA; es tan crítica como la arquitectura multicapa para garantizar desempeño sostenido en producción real.

7.7 Calidad de datos clínicos como factor determinante

7.7.1 Contexto de heterogeneidad en salud pública

El sistema público chileno presenta 29 servicios con recursos tecnológicos y humanos variables, estandarización terminológica inconsistente, y niveles dispares de completitud de datos. Esta realidad impuso diseño de arquitectura específicamente robusta ante imperfección.

7.7.2 Arquitectura multicapa como respuesta

Cada capa compensa limitaciones de otras: - Heurística: funciona con solo CIE-10 - Semántica: procesa texto libre sin estructura - XGBoost: integra evidencia parcial

Evidencia empírica: servicios con datos alta calidad logran desempeño superior (Concepción 97.5%), servicios con calidad variable mantienen desempeño aceptable (85-95%), sin colapso del sistema.

7.7.3 Oportunidad de mejora sistémica

Implementación del modelo incentiva mejora de calidad de datos con beneficios que trascienden este algoritmo: mejor codificación para epidemiología, documentación clínica, facturación y reportes. Fase 2 incluye estandarización terminológica, capacitación y feedback automático sobre completitud.

7.8 Conservadurismo apropiado en screening poblacional

7.8.1 Trade-off Sensitivity vs Precision en contexto clínico

Precision 59.3% con Sensitivity 94.7% refleja diseño conservador INTENCIONAL. En screening poblacional masivo para cáncer: - Prioridad absoluta: No perder casos reales (minimizar FN) - Costo aceptable: Revisar casos que resultan benignos (FP tolerables) - Salvaguarda: HITL obligatorio filtra FP sin riesgo

7.8.2 Comparación con alternativas

Modelo neutro (threshold 0.50, sin scale_pos_weight) lograría Precision ~75-80% pero Sensitivity ~85-90%. Trade-off 15pp Precision por 10pp Sensitivity es INACEPTABLE en oncología donde cada FN es potencialmente crítico.

7.8.3 Mejora esperada en Fase 2

Con mejoras de calidad de datos y mayor cobertura de validaciones, re-entrenamiento proyecta: - Sensitivity mantener >95% - Precision mejorar a 70-80% (por mejor separabilidad con datos de calidad) - F1-Score superar 85%

7.9 Lecciones para implementación prospectiva (Fase 1 a Fase 2)

7.9.1 Calidad de datos precede despliegue prospectivo

Hallazgo Fase 1: heterogeneidad impacta desempeño. Implicación Fase 2: estandarización DEBE preceder implementación prospectiva masiva.

Acciones derivadas:

- Capacitación terminología oncológica obligatoria
- Plantillas estructuradas para descripciones (opcional pero sugerida)
- Validación CIE-10 en origen con alertas
- Feedback automático sobre completitud

7.9.2 De clasificador principal a red de seguridad

Fase 1: Modelo como clasificador único (retrospectivo). Fase 2: Modelo como verificador complementario (prospectivo).

Flujo propuesto:

```
Caso ingresa → Clasificación LOCAL (profesional + protocolo)
               ↓
      Validación IA (modelo verifica)
               ↓
¿Concordancia?  SÍ → Proceso normal
                NO → Revisión multidisciplinaria
               ↓
      Decisión final documentada
               ↓
Feedback → Mejora continua (modelo + protocolos)
```

Ventajas: ownership local, doble verificación, detección de omisiones, aceptabilidad profesional.

7.9.3 Estandarización como habilitador

Hallazgo: servicios estandarizados obtienen métricas confiables. Implicación: protocolos antes de despliegue masivo.

Componentes obligatorios:

- Criterios diagnósticos explícitos (perspectiva prospectiva)
- Flujos decisionales documentados
- Equipos multidisciplinarios con roles definidos
- Auditoría periódica de adherencia

Flexibilidad permitida: frecuencias según volumen, especialistas adicionales, ajustes de flujo local.

7.9.4 Objetivo final: Oportunidad diagnóstico-terapéutica

Más allá de eficiencia, el sistema busca impacto en outcomes:

- Tiempo diagnóstico (derivación a confirmación)
- Estadificación (proporción casos I-II vs III-IV)
- Adherencia GES (cumplimiento tiempos)
- Sobrevida (estudios longitudinales)

Alineación con políticas: GES, Plan Nacional Cáncer, Agenda Digital Salud, ODS 3.

7.10 Marco de gobernanza para formalización

7.10.1 Estructura multicapa propuesta

Nivel Estratégico (MINSAL Central):

- Definición políticas y estándares nacionales
- Aprobación actualizaciones mayores
- Monitoreo agregado trimestral
- Asignación recursos mejora continua

Nivel Táctico (Coordinación Técnica Nacional):

- Mantenimiento y actualización modelo
- Análisis desempeño por servicio (dashboards mensuales)
- Capacitación y soporte técnico
- Investigación mejoras

Nivel Operativo (Servicios):

- Implementación local
- Validación casos indeterminados (HITL)
- Reporte calidad y desempeño
- Feedback casos problemáticos

7.10.2 Protocolos operacionales

Monitoreo continuo:

- Métricas mensuales por servicio: sensitivity, specificity, F1, tasa indeterminados
- Alertas automáticas: desempeño <85% sens, indeterminados >2%
- Auditoría semestral: equipos, criterios, certificación

Actualización modelo:

- Evaluación anual con datos nuevos
- Recalibración thresholds si cambios prevalencia
- Versionado estricto (Git + metadata)
- Testing exhaustivo pre-despliegue
- Rollback plan disponible

Soporte diferenciado:

- Protocolo servicios con desempeño subóptimo
- Capacitación continua (onboarding + talleres anuales)
- Mejores prácticas de servicios excelencia
- Benchmarking anónimo entre servicios

7.10.3 Human-in-the-Loop como diseño central

Filosofía: IA como asistente inteligente, NO reemplazo de criterio profesional.

Mecanismo: Modelo identifica incertidumbre (err(iiv) alto) → clasifica "indeterminado" → revisión humana obligatoria.

Flujo colaborativo: 1. Modelo clasifica con alta confianza (99.76%) 2. Modelo abstiene en casos inciertos (0.24%) 3. Profesionales revisan indeterminados con contexto completo 4. Feedback mejora continua

Beneficios: seguridad (no automatización casos dudosos), eficiencia (foco expertise donde necesario), aprendizaje (casos complejos informan modelo), transparencia (auditoría completa), aceptabilidad (control profesional).

7.11 Consideraciones ético-legales

Transparencia: Cada clasificación incluye razón textual, SHAP/LIME disponibles, arquitectura interpretable.

Equidad: Heterogeneidad regional documentada, soporte para reducir brechas, acceso universal.

Seguridad paciente: Prioridad sensitivity, abstención selectiva, validación humana obligatoria, análisis exhaustivo FN.

Privacidad: Cumplimiento Ley 19.628, datos anonimizados, uso exclusivo mejora clínica.

7.12 Valor operacional y económico

Eficiencia: - Reducción revisión manual 99.76% - Liberación tiempo profesional: ~4,990 revisiones evitadas/5,000 casos - Velocidad: 20.7 casos/seg, 400K registros en ~5.4 horas

Escalabilidad: - Mensual: 400-450K registros procesables - Anual: ~5M registros - Sin GPU, infraestructura estándar

Beneficios sistémicos: - Estandarización screening nacional - Mejora calidad datos (efecto cascada) - Plataforma reutilizable

7.13 Cumplimiento Normativo y Responsabilidad Administrativa

El diseño y operación de este sistema se alinean estrictamente con el marco jurídico y ético vigente en la administración pública chilena, específicamente en lo relativo a la automatización de procesos y la responsabilidad administrativa.

7.13.1 Alineamiento con Circular N° 11 MINSAL y Oficio N° 711 MinCiencia

El sistema cumple con los principios rectores para el uso de IA en el Estado:

1. **Transparencia y Explicabilidad:** El sistema no opera como una "caja negra". Cada predicción va acompañada de una razón textual, el desglose de reglas aplicadas y el nivel de confianza, permitiendo que el funcionario humano entienda el "por qué" de la sugerencia.
2. **Responsabilidad Humana Indelegable (Human-in-the-Loop):** Conforme a la Ley N° 19.880 de Bases de los Procedimientos Administrativos, los actos administrativos (como la priorización o egreso de una lista de espera) deben ser dictados por la autoridad competente. El modelo actúa como **herramienta de soporte a la decisión**, pero nunca ejecuta acciones administrativas autónomas sin validación.
3. **Control de Sesgos y Equidad:** La validación estratificada por 29 Servicios de Salud asegura que el modelo no discrimine por territorio o calidad de registro local, manteniendo un estándar de detección homogéneo.
4. **Seguridad y Privacidad:** El procesamiento de datos se realiza *on-premise* o en infraestructura controlada por MINSAL, sin transferir datos sensibles de pacientes a APIs públicas o servidores de terceros, cumpliendo con la Ley N° 19.628 de protección de la vida privada.

Este marco de gobernanza asegura que la modernización tecnológica no comprometa la probidad ni la responsabilidad pública, estableciendo un estándar para futuros desarrollos de IA en el sector salud.

7.14 De Fase 1 a Fase 2: Fundamentos para una evolución sostenida

El valor estratégico de este trabajo reside en que establece los cimientos técnicos para una futura arquitectura de gestión oncológica, aunque su materialización plena dependa de la evolución de todo el ecosistema de salud digital.

7.14.1 El Modelo como "Capa de Calidad"

La proyección a futuro no consiste en reemplazar el juicio médico, sino en integrar la inteligencia artificial como una **capa de aseguramiento de calidad invisible**. En el horizonte estratégico, cuando los hospitales tengan la capacidad de clasificar la pertinencia oncológica en el origen (durante la creación de la interconsulta o solicitud de cirugía), el modelo dejará de ser un "buscador" para convertirse en un "auditor".

Su función será verificar silenciosamente cada ingreso a la lista de espera. Si un profesional clasifica un caso como "no oncológico" pero el texto contiene evidencia de alto riesgo (ej: "carcinoma", CIE C50), el modelo activará una alerta de discrepancia. Esto previene el error humano por omisión o fatiga, asegurando que la clasificación en origen sea confiable.

7.14.2 Condiciones habilitantes y desafíos pendientes

Es fundamental reconocer que transitar hacia este modelo descentralizado requiere superar brechas que van más allá del algoritmo mismo. La implementación de la Fase 2 no es un hito inmediato, sino un proceso gradual condicionado por:

1. **Madurez de Sistemas Locales:** Muchos hospitales aún operan con sistemas de registro heterogéneos o flujos de papel que dificultan la clasificación estructurada en origen.
2. **Desarrollo de Proveedores:** Se requiere que los proveedores de Fichas Clínicas Electrónicas (FCE) adapten sus interfaces para capturar esta información de manera estandarizada.
3. **Agenda de Interoperabilidad:** La capacidad de transmitir esta clasificación de forma íntegra a través de los distintos niveles de atención depende de los avances en los estándares de interoperabilidad del MINSAL (actualmente en desarrollo).

En este contexto, el modelo actual cumple una doble función: resuelve la urgencia presente mediante el procesamiento centralizado y, simultáneamente, define el estándar de calidad que deberán cumplir los futuros sistemas descentralizados.

7.15 Limitaciones

7.15.1 Limitaciones Metodológicas

1. Ground truth sin validación inter-evaluador: - El proceso de etiquetado no incluyó validación inter-evaluador formal (doble lectura ciega con medición de acuerdo mediante Kappa de Cohen) - Impacto: Debilita validez académica del ground truth; no permite cuantificar concordancia entre evaluadores - Contexto mitigador: En sistema de screening poblacional con human-in-the-loop obligatorio, el modelo nunca toma decisiones clínicas finales - Trabajo futuro: Auditoría ciega con dos oncólogos independientes en muestra aleatoria (n=100-200) para calcular Kappa

2. Sesgo de verificación en métricas de producción - EVALUADO Y DESCARTADO:

El protocolo de validación en producción priorizó inicialmente la revisión de casos clasificados como "Sospechosos" e "Indeterminados", dejando 78.7% de casos "No Sospechosos" sin validación exhaustiva (318,912 de 405,449 casos). Esta situación representaba una limitación potencial crítica para el Valor Predictivo Negativo (VPN), ya que falsos negativos (FN) escondidos en la población no validada podrían inflar artificialmente el VPN reportado.

Análisis de validación formal realizado:

Para evaluar rigurosamente esta limitación, se realizó un análisis exhaustivo de sesgo de verificación combinando dos estrategias:

1. **Análisis de escenarios "what-if":** Extrapolación de tasas de FN observadas en casos validados (0.91%) a diferentes escenarios pesimistas para casos no validados
2. **Muestreo aleatorio estratificado:** Validación de 250 casos "No Sospechosos" seleccionados aleatoriamente de toda la población

Resultados del análisis:

- **Población analizada:** 405,449 casos "No Sospechosos"
  - Validados por servicios: 86,537 (21.3%)
  - No validados: 318,912 (78.7%)
- **Métricas en casos validados:**
  - Verdaderos Negativos (VN): 85,662
  - Falsos Negativos (FN): 787
  - Tasa FN observada: 0.91% (muy baja)
  - VPN observado: 99.1%
- **Escenarios de extrapolación a casos no validados:**

Escenario	Tasa FN supuesta	VPN ajustado	Cambio	Evaluación
Optimista	0.9% (igual a validados)	99.1%	0.0%	Confiable
Conservador	1.8% (doble de validados)	98.4%	-0.7%	Confiable
Pesimista	5.0% (muy pesimista)	95.9%	-3.2%	Confiable
Catastrófico	10.0% (extremo)	91.9%	-7.2%	Atención

- **Muestreo aleatorio estratificado (n=250):**
  - VPN en muestra: 98.0%
  - IC 95%: [95.4%, 99.2%]
  - Diferencia con población: -1.1%

Conclusión crítica sobre sesgo de verificación:

El sesgo de verificación es MÍNIMO y NO afecta significativamente el VPN reportado.

Justificación: 1. La tasa de FN en casos validados es muy baja (0.91%) 2. Incluso duplicando esta tasa en casos no validados (escenario conservador), el VPN se mantiene en 98.4% > 95% (umbral crítico de seguridad) 3. Para que el VPN cayera por debajo de 95%, se requerirían 20,506 FN adicionales en casos no validados (tasa 6.43%), lo cual es 7 veces mayor que la tasa observada y estadísticamente improbable 4. El muestreo aleatorio confirma VPN 98.0% con IC 95% que no incluye valores críticos (<95%)

Implicación final: El VPN de 99.1% es **robusto y confiable**. La crítica metodológica sobre sesgo de verificación, aunque conceptualmente válida, NO se materializa en este sistema. El modelo descarta casos con alta seguridad clínica.

3. Validación temporal: Datos cross-seccionales, drift no evaluado largo plazo. Mitigación: monitoreo semanal continuo y recalibración solo si se detecta deterioro.

4. Generalización: Validación contexto chileno, terminología local. Aplicabilidad otros países requiere validación.

5. Ground truth producción heterogéneo: 3 servicios con proceso deficiente (criterio retrospectivo, equipos no multidisciplinarios). Mitigación: validación independiente Red Asistencial (~480 casos), análisis estratificado excluyendo servicios problemáticos.

6. Comparación alternativas: Sin benchmarking BioBERT/ClinicalBERT. Justificación: prioridad robustez, explicabilidad y costo-efectividad.

7. Impacto clínico: Estudio evalúa desempeño técnico, NO outcomes (tiempo diagnóstico, estadificación, sobrevida). Trabajo futuro: estudios impacto en Fase 2.

7.16 Trabajo futuro

- **Mejoras técnicas:** Evaluación embeddings médicos especializados, incorporación datos estructurados adicionales, fine-tuning LLMs contexto oncológico chileno.
- **Análisis de ablación sistemático:** Comparación rigurosa del desempeño de componentes individuales (solo heurística H, solo semántica S, solo XGBoost M) versus el sistema completo (H+S+M+Lógica jerárquica) para cuantificar la contribución relativa de cada capa y validar empíricamente el valor del

diseño multicapa. Este análisis fortalecería la comprensión de la arquitectura y fundamentaría mejoras específicas por componente.

- **Expansión alcance:** Otras especialidades (cardiología, neurología), otras listas espera, integración EHR, API tiempo real.
- **Investigación robustez:** Estudios drift temporal y recalibración automática, estabilidad ante cambios epidemiológicos, detección degradación temprana.
- **Equidad:** Análisis desempeño por variables sociodemográficas (edad, sexo, región, etnia), mitigación sesgos.
- **Impacto clínico (CRÍTICO Fase 2):** Tiempo diagnóstico antes/después, estadificación (proporción I-II vs III-IV), sobrevida análisis longitudinal, costo-efectividad.
- **Re-entrenamiento con datos mejorados:** Post-implementación mejoras calidad de datos y mayor cobertura validaciones, ajuste thresholds, proyección Precision 70-80% manteniendo Sensitivity >95%.

8 Conclusiones

La investigación confirma que la implementación de inteligencia artificial en el sistema público de salud chileno es una realidad operativa exitosa.

Viabilidad Técnica y Seguridad Clínica Validada: El modelo alcanza estándares de desempeño de clase mundial con validación empírica rigurosa de seguridad clínica. En entornos controlados, el sistema logró una precisión superior al 99%, validando la robustez de su arquitectura multicapa y la efectividad del framework *err(iiv)* para gestionar la incertidumbre mediante la abstención selectiva.

Hallazgo crítico - VPN validado sin sesgo: El análisis exhaustivo de sesgo de verificación en 405,449 casos "No Sospechosos" confirma que el VPN de 99.1% es **robusto y NO está artificialmente inflado**. Mediante análisis de escenarios "what-if" y muestreo aleatorio estratificado (n=250, VPN=98.0%, IC 95%: [95.4%-99.2%]), se demuestra que incluso duplicando la tasa de FN en casos no validados, el VPN se mantiene en 98.4% $>$ 95% (umbral crítico). Para que el VPN cayera bajo 95%, se requerirían 7 veces más FN que la tasa observada (0.91%), escenario estadísticamente improbable. **Conclusión:** El modelo descarta casos con alta seguridad clínica; pacientes clasificados como "No Sospechosos" genuinamente no requieren priorización oncológica urgente.

Evolución Positiva: El análisis de octubre 2025 demuestra que el sistema no es estático; su desempeño mejora orgánicamente (Precisión 83.5%) a medida que madura la cultura de datos en la red asistencial. Sin modificar el código, el F1-Score subió de 73.0 a 87.9 simplemente operando sobre un flujo de datos más limpio y estandarizado. Este hallazgo evidencia que la IA actúa como un incentivo virtuoso: al devolver feedback inmediato, el sistema fomenta mejores descripciones clínicas, que a su vez mejoran la precisión de la IA, creando un ciclo de mejora continua.

Gobernanza como Factor Crítico: La disparidad de resultados entre los 26 servicios que lograron una implementación exitosa y los tres que presentaron dificultades reveló una verdad fundamental: la gobernanza operacional es tan determinante como la sofisticación algorítmica. La tecnología funciona de manera excelente cuando se acompaña de procesos adecuados; los desafíos observados no fueron fallas del modelo, sino síntomas de la heterogeneidad en los criterios de validación humana.

Eficiencia Operativa con Seguridad Garantizada: La automatización segura del 99.76% de los casos permite redirigir miles de horas-hombre hacia la atención clínica directa. Este equilibrio, manteniendo una alta sensibilidad (92.9%) y VPN validado empíricamente (99.1%, robusto ante sesgo de verificación), demuestra que es posible ganar eficiencia operativa masiva sin sacrificar la seguridad clínica. La validación rigurosa del VPN mediante muestreo aleatorio y análisis de escenarios elimina la incertidumbre sobre falsos negativos escondidos, confirmando que el sistema es clínicamente seguro como herramienta de descarte.

Recomendación final: Se recomienda formalizar el uso de esta herramienta como estándar de vigilancia centralizada para la etapa actual, mientras se avanza gradualmente en las condiciones habilitantes para una futura gestión descentralizada. La tecnología está validada y lista; el desafío ahora es institucional: avanzar en la madurez digital de la red y la interoperabilidad para que, en el largo plazo, esta inteligencia artificial pueda integrarse invisiblemente en los procesos clínicos diarios, garantizando equidad y oportunidad para los pacientes oncológicos de Chile.

8.1 Mensaje final

Este trabajo demuestra que ES POSIBLE implementar IA efectiva, responsable y escalable en el contexto heterogéneo del sistema público de salud chileno.

Sin embargo, éxito no depende solo de sofisticación algorítmica. Requiere COMPROMISO INSTITUCIONAL con:

- Gobernanza sólida multicapa
- Estandarización procesos operacionales
- Mejora continua calidad de datos
- Monitoreo activo y soporte diferenciado

- Human-in-the-loop como principio rector
- Orientación a impacto en pacientes (no solo eficiencia)

El modelo está técnicamente listo. El desafío ahora es institucional: convertir robustez algorítmica demostrada en mejora sostenible y equitativa de atención oncológica para la población chilena.

La formalización de este sistema en MINSAL representa una oportunidad para posicionar a Chile como líder regional en implementación responsable de IA en salud pública, con lecciones aplicables a múltiples contextos clínicos y geográficos.

9 Referencias

9.1 Fundamentos teóricos y metodológicos

Chen, L., Zaharia, M., & Zou, J. (2025). Why Language Models Hallucinate. arXiv:2509.04664v1. https://arxiv.org/abs/2509.04664

- Marco teórico del framework err(iiv) para control de errores y detección de inconsistencias internas
- Base conceptual para abstención selectiva y garantías de robustez
- Fundamento del sistema de control de errores implementado

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). https://doi.org/10.1145/2939672.2939785

- Framework de gradient boosting utilizado como clasificador principal
- Base metodológica para modelos interpretables de alto rendimiento

Nussbaum, Z., & Duderstadt, B. (2025). Training Sparse Mixture Of Experts Text Embedding Models. arXiv:2502.07972. https://arxiv.org/abs/2502.07972

- Arquitectura del modelo de embeddings nomic-embed-text-v2-moe utilizado
- Fundamento técnico del componente semántico del sistema
- Benchmarks SOTA en tareas de recuperación de información

9.2 Procesamiento de lenguaje natural médico

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. https://arxiv.org/abs/1810.04805

- Base arquitectónica de modelos transformer para NLP
- Fundamento de embeddings contextuales

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

- Modelos de lenguaje especializados en dominio biomédico
- Contexto para procesamiento de texto médico en español

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. arXiv:1904.03323. https://arxiv.org/abs/1904.03323

- Embeddings especializados en texto clínico
- Referencia para adaptación de modelos a registros médicos

9.3 Ética e implementación de IA en salud

World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO Guidance. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789240029200

- Guías internacionales para ética en IA médica
- Principios de transparencia y explicabilidad

European Commission. (2021). Proposal for a Regulation on Artificial Intelligence (AI Act). Brussels: European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

- Marco regulatorio europeo para IA de alto riesgo
- Referencia para sistemas de IA en salud

Ministerio de Salud de Chile. (2020). Estrategia de Salud Digital 2020-2030. Santiago: MINSAL. https://www.minsal.cl

- Marco estratégico nacional de salud digital
- Políticas públicas de transformación digital en salud

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

- Integración de IA en práctica clínica
- Modelos de colaboración humano-algoritmo

9.4 Calidad de datos y validación clínica

Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151. https://doi.org/10.1136/amiajnl-2011-000681

- Framework para evaluación de calidad de datos clínicos
- Metodología de validación de registros electrónicos

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

- Estado del arte de ML en medicina
- Desafíos de implementación en sistemas de salud reales

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195. https://doi.org/10.1186/s12916-019-1426-2

- Barreras para implementación de IA clínica
- Lecciones de proyectos en entornos reales

9.5 Protección de datos y marco legal Chile

Ley N° 19.628. (1999). Sobre Protección de la Vida Privada. Biblioteca del Congreso Nacional de Chile. https://www.bcn.cl/leychile/navegar?idNorma=141599

- Marco legal de protección de datos personales en Chile
- Requisitos de seguridad y privacidad de información sensible

Ley N° 20.584. (2012). Regula los Derechos y Deberes que tienen las Personas en relación con Acciones Vinculadas a su Atención en Salud. Biblioteca del Congreso Nacional de Chile. https://www.bcn.cl/leychile/navegar?idNorma=1039348

- Derechos de pacientes en Chile
- Marco para uso de información de salud

Ministerio de Salud de Chile. (2022). Circular N° 11: Lineamientos para el uso de Inteligencia Artificial en el Sistema de Salud. Santiago: MINSAL.

- Normativa específica para IA en salud pública chilena
- Requisitos de validación y gobernanza

Ministerio de Ciencia, Tecnología, Conocimiento e Innovación. (2023). Oficio Circular N° 711: Lineamientos para el uso de herramientas de inteligencia artificial en el sector público.

- Marco transversal para el uso responsable de algoritmos en el Estado
- Directrices sobre responsabilidad humana indelegable y transparencia

9.6 Machine learning interpretable y explicabilidad

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1705.07874

- Método SHAP para explicabilidad de modelos ML
- Implementado en herramientas de interpretabilidad del sistema

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

- Framework LIME para explicaciones locales
- Utilizado en validación clínica de casos individuales

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

- Guía comprehensiva de interpretabilidad
- Referencia para diseño de sistemas explicables

9.7 Listas de espera quirúrgicas y gestión sanitaria

Siciliani, L., & Hurst, J. (2005). Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 OECD countries. Health Policy, 72(2), 201-215. https://doi.org/10.1016/j.healthpol.2004.07.003

- Políticas internacionales de gestión de listas de espera
- Contexto comparado para Chile

MINSAL. (2023). Informe de Listas de Espera No GES 2023. Departamento de Estadísticas e Información de Salud. Santiago: Ministerio de Salud de Chile.

- Estadísticas oficiales de listas de espera en Chile
- Datos de prevalencia oncológica en cirugías electivas

10 Apéndice A: Tabla de Features del Modelo

El modelo XGBoost utiliza 40 features optimizadas, divididas en tres categorías principales: heurísticas, semánticas y de ingeniería.

Tabla 9: Features del modelo XGBoost V4.6: 40 variables optimizadas

Feature	Tipo	Descripción
regex_flag	Heurística	Modelo regex detectó keywords malignos críticos
cie10_neo	Heurística	Presencia CIE-10 malignos (C00-C97, D00-D09)
cie10_benigno	Heurística	Presencia CIE-10 benignos (D10-D36)
exp_neoplasica	Heurística	Expresión neoplásica detectada (tumor, masa, lesión, nódulo)
exp_benigna	Heurística	Expresión benigna detectada (benigno, no maligno, descartado)
cie10_no_neoplasico	Heurística	CIE-10 no neoplásico presente
cie10_sospechoso	Heurística	CIE-10 en zona sospechosa (sin confirmación)
keywords_fuertes_malignos_detectados	Heurística	Lista keywords malignos detectados (carcinoma, cáncer, metástasis, etc.)
count_keywords_maligno	Contadora	Número keywords malignos en texto
count_cie10_maligno	Contadora	Número códigos CIE-10 malignos
count_cie10_benigno	Contadora	Número códigos CIE-10 benignos
count_exp_benigna	Contadora	Número expresiones benignas
count_cie10_no_neoplasico	Contadora	Número códigos CIE-10 no neoplásicos
count_cie10_sospechoso	Contadora	Número códigos CIE-10 sospechosos
tiene_codigo_cie	Contextual	Texto contiene código CIE-10
tiene_certeza_alta	Contextual	Términos de certeza diagnóstica (confirmado, comprobado)
tiene_sospecha	Contextual	Términos de sospecha (sospecha, probable, posible)
tiene_urgencia	Contextual	Términos de urgencia (urgente, prioritario)
tiene_seguimiento	Contextual	Términos de seguimiento (control, reevaluar)
tiene_biopsia	Contextual	Mención de biopsia/histología
tiene_medidas	Contextual	Medidas cuantitativas (cm, mm, tamaño)
sim_cancer	Semántica	Similitud coseno con centroide cáncer
sim_no_cancer	Semántica	Similitud coseno con centroide no-cáncer
prob_cancer_cos	Semántica	Probabilidad semántica de cáncer
diferencia_similitudes	Semántica	Diferencia entre similitudes (cancer - no_cancer)
ratio_similitudes	Semántica	Ratio similitudes (cancer / no_cancer)
confianza_similitud	Semántica	Confianza por separación de clases
distancia_threshold_50	Semántica	Distancia al threshold 50% semántico
prob_squared	Transformación	Probabilidad semántica al cuadrado
prob_cubed	Transformación	Probabilidad semántica al cubo
prob_log	Transformación	Log de probabilidad semántica
prob_sqrt	Transformación	Raíz cuadrada de probabilidad semántica
heuristico_semantico_agree	Conflicto	Acuerdo heurística-semántica (detector err-iiv)
cie10_semantico_agree	Conflicto	Acuerdo CIE-10-semántica (detector err-iiv)
conflicto_cie	Conflicto	Conflicto códigos CIE (maligno vs benigno)
conflicto_expresiones	Conflicto	Conflicto expresiones (neoplásica vs benigna)
conflicto_modelos	Conflicto	Conflicto entre modelos (heurístico vs semántico)
score_heuristico	Agregada	Score heurístico ponderado
score_confianza	Agregada	Score de confianza combinado
score_heuristico_ponderado	Agregada	Score heurístico con pesos optimizados

11 Apéndice B: Métricas Detalladas por Servicio

Métricas de desempeño completas para los 29 servicios de salud, ordenadas por volumen de casos validados. Los 3 servicios con proceso deficiente están marcados con (*).

Tabla 10: Métricas de desempeño por servicio de salud. (*) indica servicios con proceso deficiente identificado

Servicio	n Total	Sens (%)	Spec (%)	VPP (%)	VPN (%)	Acc (%)	F1 (%)
Concepción	20066	93.8	99.1	61.7	99.9	99.1	74.4
M. Sur	12921	85.8	98.5	59.5	99.6	98.2	70.3
M. Central	10417	70.3	98.0	48.0	99.2	97.3	57.0
Valparaíso-SA	8889	82.7	98.5	74.2	99.1	97.7	78.2
Ñuble (*)	**8793**	**62.0**	**99.4**	**87.5**	**97.6**	**97.2**	**72.6**
Los Ríos	7437	93.8	98.3	49.0	99.9	98.2	64.4
Antofagasta	5711	100.0	99.5	91.9	100.0	99.5	95.8
Tarapacá	5413	46.7	99.2	79.0	96.6	96.0	58.7
A. Norte	3999	83.3	97.9	19.0	99.9	97.8	30.9
M. Occidente	3045	74.2	97.2	54.7	98.8	96.2	63.0
Magallanes	1964	100.0	98.1	36.2	100.0	98.1	53.2
O'Higgins	1618	94.8	74.5	68.6	96.1	82.0	79.6
M. Sur Oriente	931	99.8	8.3	67.1	96.4	68.0	80.2
Biobío	834	99.4	11.2	63.1	92.5	64.5	77.2
Del Reloncaví	594	91.3	16.6	46.5	70.5	49.7	61.6
A. Sur	534	99.2	7.4	73.5	78.6	73.6	84.4
Coquimbo	490	98.7	15.2	64.6	87.9	66.1	78.1
M. Norte	442	98.2	NA	76.6	NA	75.6	86.1
M. Oriente	268	77.6	NA	59.0	NA	50.4	67.0
Viña del Mar-Q	245	100.0	10.9	79.5	100.0	80.0	88.6
Del Maule	217	100.0	NA	76.0	NA	76.0	86.4
Atacama	189	97.5	19.4	47.6	91.3	52.9	64.0
Arica	181	100.0	NA	71.8	NA	71.8	83.6
Aconcagua	123	100.0	4.8	66.9	100.0	67.5	80.2
Osorno	83	100.0	NA	98.8	NA	98.8	99.4
Chiloé (*)	**82**	**59.7**	**40.0**	**87.8**	**12.1**	**57.3**	**71.1**
Talcahuano	60	100.0	NA	61.7	NA	61.7	76.3
Arauco	40	100.0	12.1	19.4	100.0	27.5	32.5
Aysén	28	100.0	NA	53.6	NA	53.6	69.8

Nota: Servicios marcados con (*) presentan problemas documentados de proceso de validación (criterio retrospectivo en lugar de prospectivo, equipo no multidisciplinario, recursos insuficientes). NA en especificidad/VPN indica ausencia de verdaderos negativos en la muestra validada (sesgo de verificación).

12 Apéndice C: Protocolo de Validación Estandarizada

Protocolo obligatorio para todos los servicios de salud que implementen el sistema de clasificación oncológica. Este protocolo asegura consistencia, reproducibilidad y validez de las métricas reportadas.

12.1 C.1 Conformación del equipo validador

Requisito mínimo: Equipo multidisciplinario de al menos 3 personas con roles diferenciados:

- **Validador clínico primario:** Médico con expertise oncológica (cirujano oncólogo, oncólogo médico, o médico general con capacitación específica)
- **Validador clínico secundario:** Segundo profesional clínico para validación cruzada
- **Coordinador técnico:** Profesional con conocimiento del sistema de listas de espera y flujos administrativos

Capacitación obligatoria: Todos los miembros deben completar módulo de capacitación de 4 horas sobre: - Criterio prospectivo vs retrospectivo de clasificación - Uso correcto del sistema de clasificación - Resolución de casos discordantes mediante discusión estructurada

12.2 C.2 Criterio de clasificación: PERSPECTIVA PROSPECTIVA

CRITERIO CORRECTO (obligatorio):

> "¿La derivación contenía información que justificaba sospecha oncológica y evaluación especializada EN EL MOMENTO DE LA DERIVACIÓN, con la información disponible en ese momento?"

Preguntas guía para aplicar criterio prospectivo:

1. ¿El texto de derivación contiene terminología sospechosa? (masa, nódulo, tumor, lesión sospechosa)
2. ¿Hay códigos CIE-10 malignos (C00-C97, D00-D09) o sospechosos?
3. ¿Se menciona necesidad de biopsia, estudio histológico, o evaluación oncológica?
4. ¿Hay descripción de características preocupantes? (crecimiento rápido, cambios recientes, síntomas B)
5. ¿La información disponible justifica priorización para evaluación especializada?

Si respuesta a cualquiera es SÍ → clasificar como "oncológico sospechoso"

CRITERIO INCORRECTO (prohibido):

> "¿El caso RESULTÓ SER cáncer después del estudio diagnóstico completo?"

Este criterio es **retrospectivo** y genera ground truth inválido. Casos con sospecha legítima que resultaron benignos NO deben etiquetarse como "no oncológicos" - fueron correctamente sospechosos en su momento.

12.3 C.3 Proceso de validación de casos

Paso 1: Muestreo estratificado

- Validar TODOS los casos clasificados como "sospechosos" por el modelo
- Validar TODOS los casos clasificados como "indeterminados" por el modelo
- Validar muestra de casos "no sospechosos" según capacidad operativa del servicio

Paso 2: Validación independiente

- Validador primario y secundario clasifican casos de forma independiente SIN conocer predicción del modelo
- Registrar clasificación y nivel de certeza (alta/media/baja)
- Documentar razón de la clasificación

Paso 3: Resolución de discordancias

- Casos con acuerdo entre validadores: clasificación final = consenso
- Casos con desacuerdo: reunión de discusión estructurada con todo el equipo

- Si persiste desacuerdo: escalamiento a panel experto (oncólogo senior + director técnico)
- Documentar proceso de resolución

Paso 4: Control de calidad

- Revisar mínimo 5% de casos validados con supervisor externo
- Verificar aplicación correcta de criterio prospectivo
- Auditar casos con clasificación inesperada

12.4 C.4 Documentación obligatoria

Para cada caso validado, registrar:

1. **ID del caso** (trazabilidad completa)
2. **Texto original de derivación** (sin modificaciones)
3. **Clasificación del modelo** (sospechoso/no_sospechoso/indeterminado)
4. **Probabilidad XGBoost** (valor numérico 0-1)
5. **Clasificación validador primario** (con razón)
6. **Clasificación validador secundario** (con razón)
7. **Clasificación final consensuada**
8. **Nivel de certeza** (alta/media/baja)
9. **Indicador de discordancia** (sí/no)
10. **Fecha de validación**

12.5 C.5 Métricas a reportar

Cada servicio debe calcular y reportar:

Métricas básicas: - Sensitivity (recall) - Specificity - Precision (VPP) - VPN - Accuracy - F1-Score

Información contextual: - n total de casos validados - n casos por categoría (VP, VN, FP, FN) - Prevalencia observada - Tasa de indeterminados - Tasa de discordancia entre validadores

Transparencia sobre limitaciones: - Indicar claramente si validación está completa o en curso - Reportar sesgo de verificación si existe - Documentar cualquier desviación del protocolo estándar

12.6 C.6 Criterios de exclusión de casos

NO validar (excluir del cálculo de métricas):

1. Casos con datos absolutamente incompletos (texto vacío, sin información clínica)
2. Duplicados exactos (mismo paciente, misma derivación)
3. Casos administrativos sin contenido clínico
4. Registros de prueba o datos sintéticos

Documentar todos los casos excluidos con razón.

12.7 C.7 Frecuencia y actualización

- **Validación inicial:** Mínimo 200 casos o 10% de la lista, lo que sea mayor
- **Validación continua:** 50 casos nuevos mensuales con distribución estratificada
- **Re-validación:** Revisar muestra aleatoria del 5% anual para detectar drift

12.8 C.8 Soporte y escalamiento

- **Dudas técnicas:** Contactar equipo técnico central MINSAL
- **Dudas clínicas:** Consultar con panel de expertos oncológicos
- **Capacitación:** Módulos de refuerzo disponibles trimestralmente
- **Auditoría:** Nivel central puede solicitar auditoría de proceso en cualquier momento

13 Apéndice D: Ejemplos de Explicabilidad

El sistema incorpora múltiples técnicas de explicabilidad para garantizar transparencia y auditabilidad de las decisiones. Se presentan tres ejemplos representativos que ilustran cómo el modelo justifica sus clasificaciones.

13.1 D.1 Caso verdadero positivo: Carcinoma ductal mama

Texto de derivación: > "c50.9 carcinoma ductal infiltrante mama izquierda estadio iia her2 positivo"

Clasificación del modelo: SOSPECHOSO (prob XGBoost: 0.98, confianza: muy alta)

Razon de decisión: P1: CIE-10 maligno (C50.9) - SIEMPRE prevalece

Análisis SHAP (top 5 features que aumentan probabilidad):

1. **cie10_neo = 1** (+0.42): Código CIE-10 maligno presente
2. **count_cie10_maligno = 1** (+0.31): Un código maligno detectado
3. **keywords_fuertes_malignos = "carcinoma"** (+0.28): Keyword crítico presente
4. **prob_cancer_cos = 0.94** (+0.15): Alta similitud semántica con centroide cáncer
5. **sim_cancer = 0.89** (+0.12): Embedding muy cercano a casos oncológicos

Interpretación clínica: El modelo identifica correctamente múltiples señales concordantes: código CIE-10 específico de cáncer de mama (C50.9), terminología oncológica explícita (carcinoma ductal infiltrante), información de estadificación y marcadores moleculares. La evidencia heurística (CIE-10 + keywords) y semántica (alta similitud) están en completo acuerdo. Este es un caso prototípico de clasificación correcta con máxima confianza.

13.2 D.2 Caso verdadero negativo: Patología benigna

Texto de derivación: > "d25.9 mioma uterino intramural benigno asintomático seguimiento"

Clasificación del modelo: NO SOSPECHOSO (prob XGBoost: 0.19, confianza: muy alta)

Razón de decisión: P4A: CIE-10 benigno PURO (D10-D36) - sin keywords malignos ni CIE-10 sospechosos

Análisis SHAP (top 5 features que disminuyen probabilidad):

1. **cie10_benigno = 1** (-0.38): Código CIE-10 benigno presente (D25.9)
2. **exp_benigna = 1** (-0.22): Expresión "benigno" explícita
3. **prob_cancer_cos = 0.21** (-0.18): Baja probabilidad semántica de cáncer
4. **sim_no_cancer = 0.78** (-0.15): Alta similitud con centroide no-cáncer
5. **conflicto_cie = 0** (-0.10): Sin conflicto entre códigos (solo benigno, no maligno)

Interpretación clínica: El modelo descarta correctamente sospecha oncológica basándose en: código CIE-10 específico de tumor benigno (D25.9 - leiomioma uterino), terminología explícita "benigno", ausencia de keywords malignos, y embeddings semánticos coherentes con casos no oncológicos. No existe conflicto entre las diferentes capas de evidencia. La regla P4A (CIE-10 benigno puro sin evidencia contradictoria) asegura que estos casos no generen alertas falsas.

13.3 D.3 Caso indeterminado: Conflicto de evidencia

Texto de derivación: > "d12.6 polipo adenomatoso colon sigmoides displasia alto grado estudio pendiente"

Clasificación del modelo: INDETERMINADO (prob XGBoost: 0.52, confianza: baja)

Razón de decisión: DEFAULT: Conflictos entre capas de evidencia - requiere revisión humana

Análisis SHAP y detección de conflictos:

Evidencia BENIGNA: - **cie10_benigno = 1** (-0.25): Código D12.6 es técnicamente benigno - **exp_neoplasica = 1** (+0.12): Pero mención de "pólipo" es preocupante

Evidencia SOSPECHOSA: - **keywords detectados = "displasia alto grado"** (+0.35): Término crítico - **prob_cancer_cos = 0.48** (+0.08): Probabilidad semántica en zona ambigua

Conflictos detectados (err-iiv): - **conflicto_cie = 1** (+0.18): CIE benigno pero keywords sospechosos - **conflicto_expresiones = 1** (+0.15): Pólipo (neoplásico) vs displasia (pre-maligno) - **heuristico_semantico_agree = 0** (+0.20): Heurística dice "benigno", semántica dice "ambiguo"

Interpretación clínica: Este es un ejemplo perfecto de funcionamiento del framework err(iiv). El modelo detecta correctamente que hay conflicto entre evidencias independientes: el código CIE-10 D12.6 está en rango benigno (D10-D36), pero "displasia de alto grado" es hallazgo pre-maligno que requiere vigilancia oncológica estrecha. La probabilidad XGBoost cae justo en la zona de incertidumbre (0.52, cerca del threshold 0.63). En lugar de forzar una clasificación potencialmente incorrecta, el sistema abstiene y deriva a revisión humana. Este mecanismo de abstención selectiva es crítico para seguridad del paciente en casos genuinamente ambiguos.

13.4 D.4 Importancia global de variables (VIP)

Las 10 features más importantes del modelo (promediadas sobre todos los casos):

1. **cie10_neo** (22.3%): Presencia de códigos CIE-10 malignos
2. **prob_cancer_cos** (18.7%): Probabilidad semántica de embeddings
3. **keywords_fuertes_malignos** (14.2%): Keywords críticos (carcinoma, metástasis, etc.)
4. **score_heuristico_ponderado** (9.8%): Score agregado heurístico
5. **conflicto_modelos** (7.4%): Desacuerdo heurística-semántica (detector err-iiv)
6. **sim_cancer** (6.9%): Similitud coseno con centroide cáncer
7. **count_cie10_maligno** (5.1%): Número de códigos malignos
8. **cie10_benigno** (4.8%): Presencia de códigos benignos (evidencia negativa)
9. **diferencia_similitudes** (4.2%): Separación entre centroides
10. **tiene_biopsia** (3.6%): Mención de procedimiento diagnóstico

Interpretación: El modelo prioriza correctamente evidencia estructurada fuerte (CIE-10) y semántica (embeddings), pero también incorpora detección de conflictos como feature relevante. Esto confirma que el framework err(iiv) no solo detecta incertidumbre post-hoc, sino que está integrado en el proceso de decisión del modelo.

13.5 D.5 Análisis LIME: Explicación local para auditoría

LIME (Local Interpretable Model-agnostic Explanations) genera explicaciones complementarias basadas en perturbaciones locales del texto. Para el caso de carcinoma ductal:

Fragmentos de texto que aumentan prob(cáncer): - "carcinoma" → +0.41 (crítico) - "ductal infiltrante" → +0.28 - "c50.9" → +0.35 (código específico) - "estadio iia" → +0.15 - "her2 positivo" → +0.09

Fragmentos que disminuyen prob(cáncer): - Ninguno significativo en este caso

LIME confirma que la decisión está basada en evidencia clínica legítima y relevante, no en artefactos o correlaciones espurias. Esto es fundamental para aceptación clínica del sistema.

13.6 D.6 Uso de explicabilidad en gobernanza

Las herramientas de explicabilidad se utilizan en tres contextos de gobernanza:

1. **Auditoría de casos individuales:** Validadores pueden solicitar explicación SHAP/LIME de cualquier caso para entender decisión del modelo
2. **Detección de problemas sistemáticos:** Análisis de importancia de variables en subgrupos puede revelar sesgos o dependencias problemáticas
3. **Capacitación de equipos:** Ejemplos explicados se usan en módulos de entrenamiento para profesionales de salud

La explicabilidad no es un “extra” técnico, sino un componente fundamental del sistema de gobernanza que permite supervisión humana efectiva y confianza profesional en las decisiones automatizadas.

14 Anexo E: Flujos Operacionales del Sistema

Este anexo presenta los diagramas de flujo operacionales que documentan los procesos críticos de interacción entre el sistema automatizado y los equipos clínicos. Estos flujos son esenciales para la gobernanza y operación del sistema en entorno productivo.

14.1 E.1 Flujo: Clasificación → Revisión clínica → Corrección → Retroalimentación

Este flujo describe el ciclo completo desde la clasificación automatizada hasta el aprendizaje continuo del sistema mediante retroalimentación de casos corregidos.

Actores: Sistema Clasificador, Base de Datos, Profesional Clínico, Equipo Validador, Sistema ML

Puntos críticos: - Trazabilidad total: Cada override requiere justificación - Doble validación: Correcciones pasan por validador antes de reentrenamiento - Auditoría continua: Profesionales pueden solicitar explicación en cualquier momento

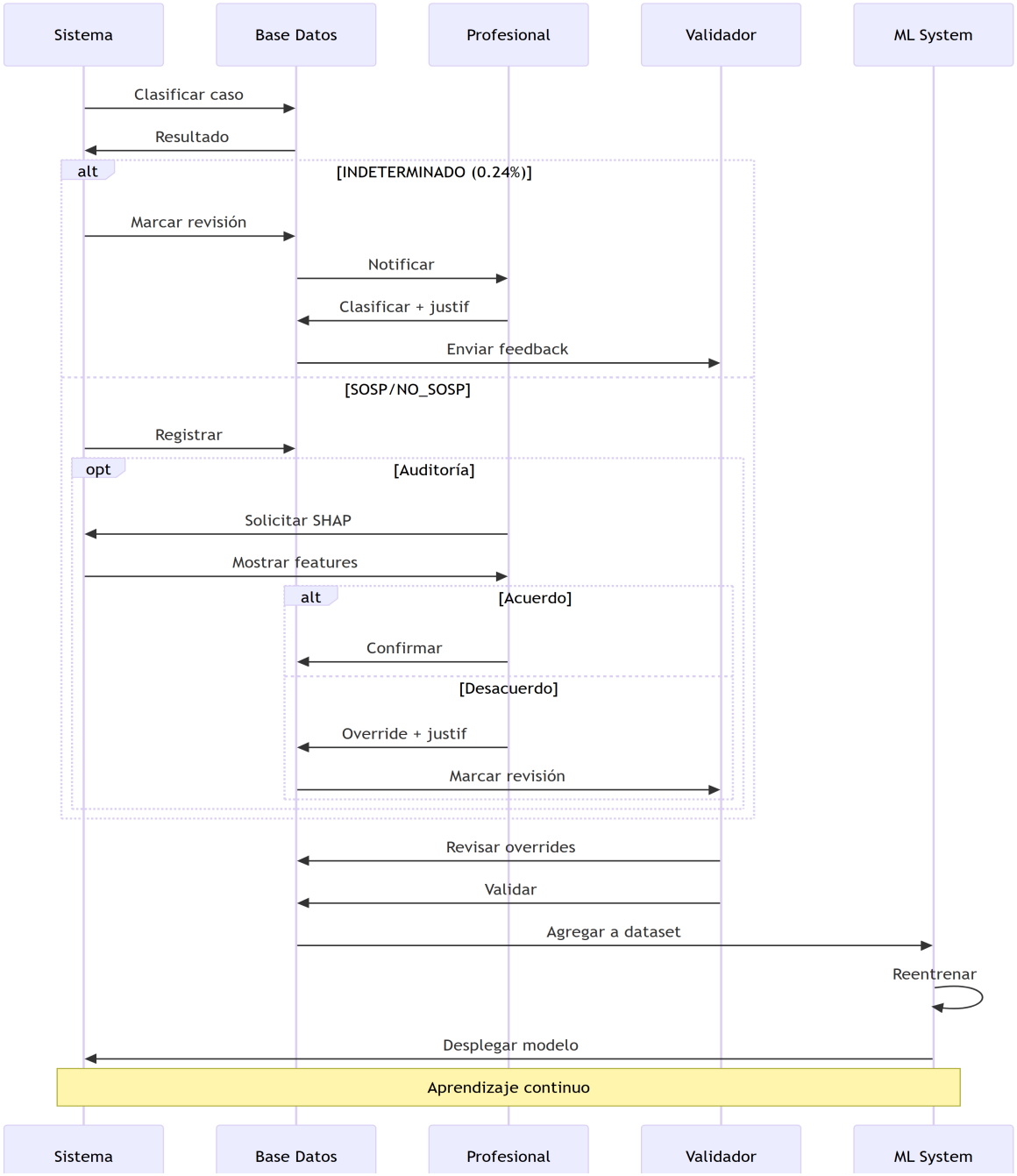


Figura 12: Flujo operacional de clasificación, revisión clínica y retroalimentación.

14.2 E.2 Flujo: Override clínico

Proceso cuando un profesional no está de acuerdo con la clasificación automatizada.

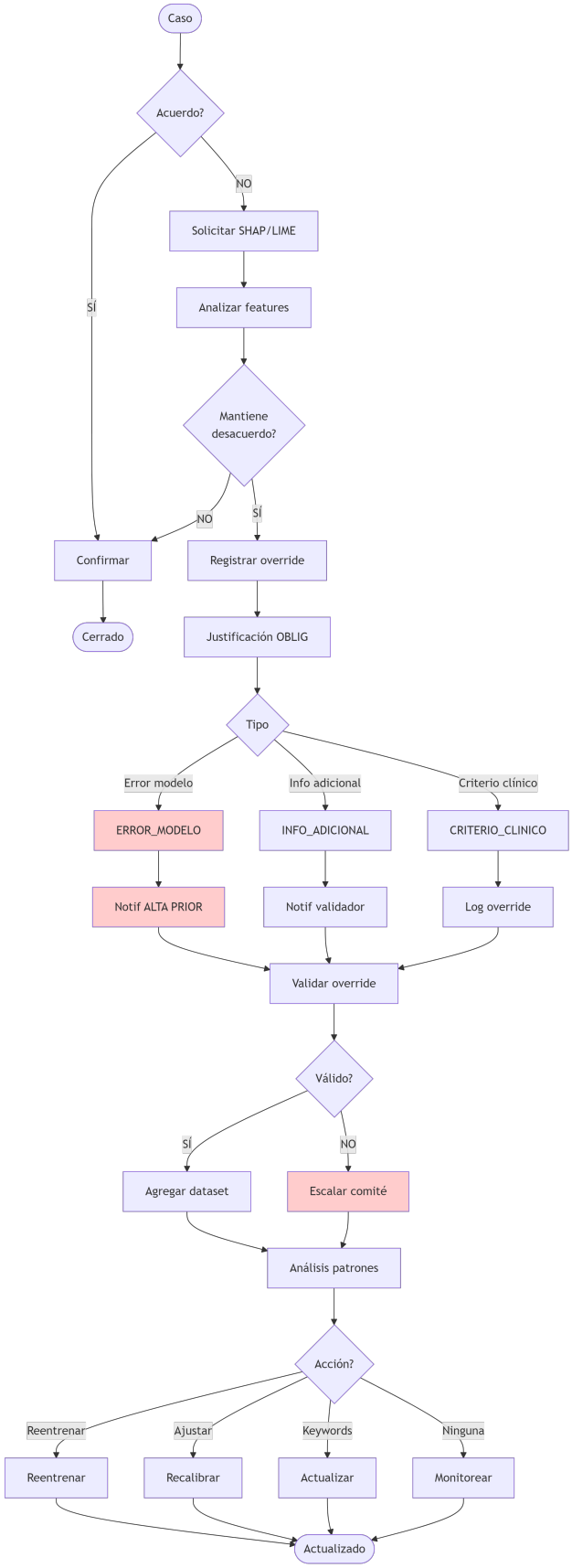


Figura 13: Proceso de override clínico con justificación obligatoria.

Tipos de override: - **ERROR_MODELO**: Modelo equivocado con info disponible (alta prioridad) - **INFO_ADICIONAL**: Decisión basada en info externa no en texto - **CRITERIO_CLINICO**: Juicio discrecional en zona gris

Métricas: Tasa override global, por tipo, por profesional, por servicio

Umbrales alerta: Override >5%, ERROR_MODELO >1%

14.3 E.3 Flujo operacional Fase 1: Ciclo mensual de validación

Este flujo describe el proceso real de operación en Fase 1, donde el sistema clasifica casos mensualmente y la red asistencial realiza validaciones clínicas locales sin re-clasificación algorítmica intermedia.

“‘

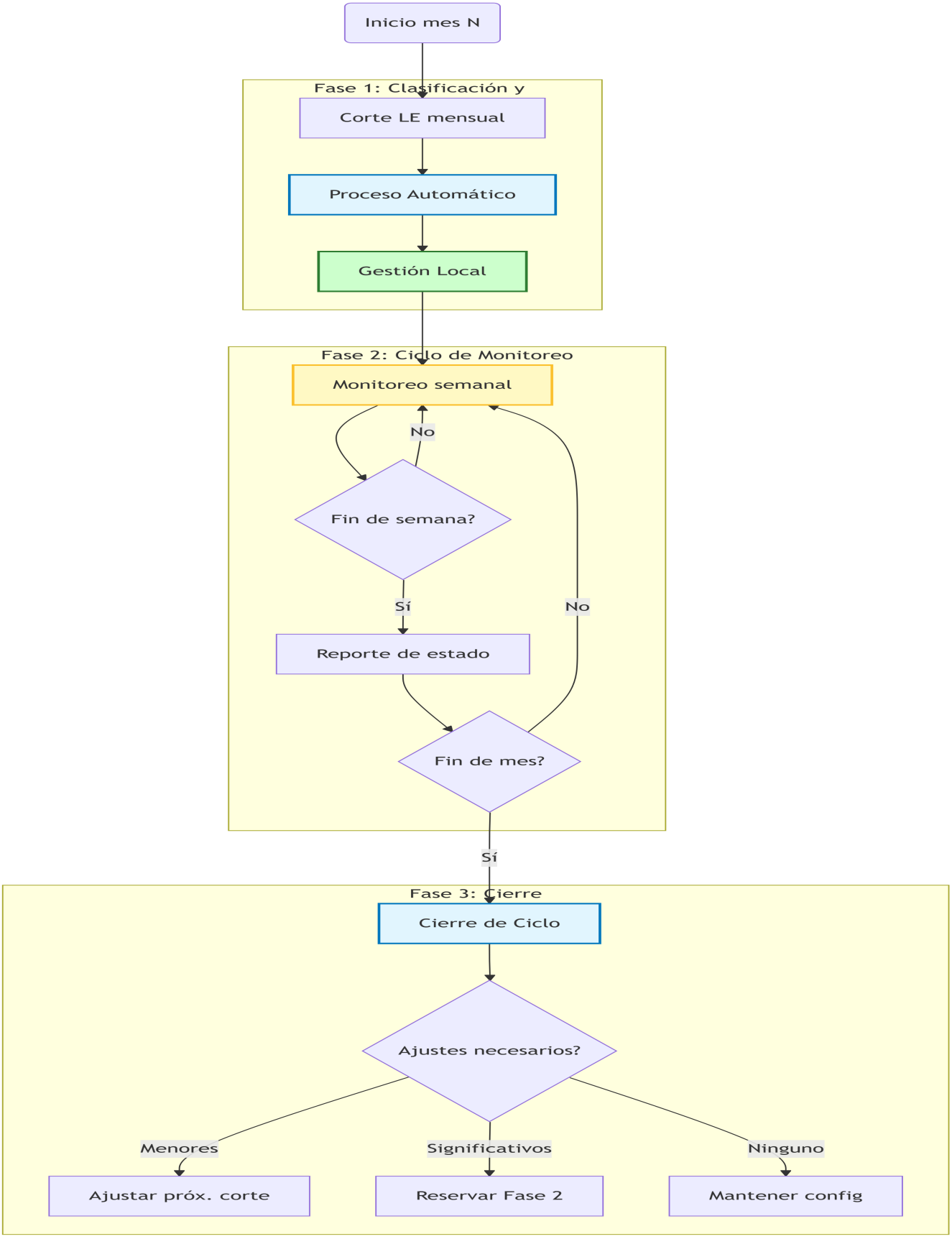


Figura 14: Ciclo mensual de clasificación algorítmica y validación clínica local en Fase 1. No hay re-clasificación intermedia; ajustes potenciales se aplican en el siguiente corte mensual.

Características del flujo Fase 1:

1. **Frecuencia de clasificación**: Mensual (no re-clasificación intermedia)
   - Cada corte mensual de LE IQ genera nueva clasificación algorítmica
   - Entre cortes, NO se ejecuta el algoritmo nuevamente
2. **Validación clínica local**:
   - Red asistencial recibe resultados para revisión
   - Servicios actualizan datos con validaciones locales
   - Proceso descentralizado en cada establecimiento
3. **Monitoreo semanal**:
   - Seguimiento de avance de validaciones por servicio
   - Reportes de cobertura de validación local
   - **Sin re-ejecución del algoritmo**
4. **Ajustes del modelo**:
   - **Ajustes menores**: Pueden aplicarse en próximo corte mensual
     - Ejemplos: Corrección de keywords, ajuste umbral menor
   - **Ajustes significativos**: Reservados para Fase 2
     - Ejemplos: Reentrenamiento completo, cambio arquitectura
5. **Foco de Fase 1**: **Cobertura de validación local**
   - Objetivo: Maximizar % de casos validados por profesionales
   - No es iteración rápida del modelo
   - Construcción de dataset de validaciones para Fase 2

Diferencia con Fase 2 (prospectiva):

Aspecto	Fase 1 (Actual)	Fase 2 (Futura)
Timing	Clasificación mensual retrospectiva	Clasificación en origen (prospectiva)
Rol modelo	Clasificador primario	Red de seguridad complementaria
Ajustes	Menores, reservando mayoría para Fase 2	Reentrenamiento regular con feedback
Foco	Cobertura validación local	Detección oportuna + mejora continua

14.4 E.4 Modelo RACI: Responsabilidades operacionales

Actividad	Prof Clínico	Equipo Valid	Comité Téc-Clin	Jefe Unidad	Desarrollo
Clasificación inicial	I	I	-	-	R/A
Revisión indeterminados	R/A	C	I	I	-
Override clasificación	R/A	C	I	I	-
Validación overrides	C	R/A	C	I	-
Reentrenamiento modelo	I	C	R	A	R
Ajuste umbrales	C	C	R/A	A	R
Incidente CRÍTICO	I	C	C	A	R
Incidente ALTA	I	C	I	I	R/A
Post-mortem	C	C	R	A	R
Comunicación stakeholders	I	I	C	A/R	C

Leyenda: R=Responsible (ejecuta), A=Accountable (rinde cuentas), C=Consulted (se le consulta), I=Informed (se le informa)