

**USER PROMPT:**

Mejora: La creciente complejidad de las técnicas de Inteligencia Artificial y el uso de conjuntos de datos cada vez más grandes, ha llevado a un aumento significativo en cuanto al consumo energético en implementaciones en hardware convencional como CPU y GPU. Este desafío resalta la necesidad de enfoques más sostenibles, como la Green AI, que busca optimizar el uso de recursos y reducir el impacto ambiental.

El uso de FPGAs (Field Programmable Gate Arrays) en este contexto surge como una alternativa prometedora, debido a las ventajas que ofrecen. La capacidad de diseñar aceleradores personalizados permite mejorar tanto la latencia como el rendimiento del sistema, manteniendo una eficiencia energética excelente en todo momento. Esta eficiencia energética obtenida, facilita la implementación de algoritmos de aprendizaje automático en sistemas embebidos con recursos limitados.

El principal objetivo de la tesis es implementar el entrenamiento y la inferencia de algoritmos clásicos de aprendizaje automático en FPGA buscando una mejor eficiencia energética manteniendo un buen rendimiento. Las arquitecturas desarrolladas se compararán con otras propuestas en FPGA y con los resultados obtenidos en hardware convencional.

Finalmente se buscará integrar los diseños de hardware implementados en un sistema embebido para evaluar su efectividad en un entorno real.

**ANSWER:**

IDENTIFIED ROLE: ENHANCEMENT ROLE

CONTEXT: El objetivo del contenido es presentar una propuesta de tesis que busca implementar algoritmos de aprendizaje automático en FPGA para mejorar la eficiencia energética, comparando con hardware convencional y evaluando en sistemas embebidos.

STRENGTHS:

- Identifica claramente un problema actual y relevante (consumo energético en IA)
- Presenta una solución específica y técnicamente viable (FPGAs)
- Establece objetivos concretos de investigación
- Incluye metodología de comparación y validación
- Aborda tanto aspectos teóricos como aplicación práctica

FEEDBACK:

1. Fortalecer la justificación cuantitativa: Incluir datos específicos sobre el consumo energético actual de CPU/GPU versus FPGAs para respaldar la relevancia del problema con evidencia numérica concreta.
2. Especificar algoritmos y métricas: Definir qué algoritmos específicos de aprendizaje automático se implementarán y qué métricas exactas se utilizarán para medir eficiencia energética y rendimiento.
3. Clarificar la metodología de comparación: Detallar cómo se realizarán las comparaciones con hardware convencional, incluyendo condiciones experimentales, datasets de prueba y criterios de evaluación.
4. Expandir el contexto de aplicación: Elaborar sobre qué tipos específicos de sistemas embebidos se considerarán y qué limitaciones de recursos enfrentarán.

EXAMPLE: "La implementación de redes neuronales convolucionales en sistemas de visión por computadora para vehículos autónomos requiere procesamiento en tiempo real con restricciones energéticas severas. Estudios previos muestran que las GPU consumen entre

150-300W durante inferencia, mientras que implementaciones en FPGA pueden reducir este consumo a 10-25W manteniendo latencias sub-10ms. Esta investigación evaluará algoritmos como SVM, Random Forest y redes neuronales feedforward implementados en Xilinx Zynq-7000, comparando consumo energético (mW), latencia (ms) y precisión (%) contra implementaciones en Intel i7 y NVIDIA RTX, utilizando datasets estándar como MNIST e ImageNet en condiciones controladas de temperatura y voltaje."

**NEXT STEPS:** Select one of the options

1. CLARIFY:

- CONCRETIZE: Get specific examples and definitions
- ELABORATE: Add depth and context
- DECONSTRUCT: Break into smaller steps
- CONTEXTUALIZE: Situate in broader framework
- SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"

2. UPLOAD IMPROVED VERSION

**USER PROMPT:**

mejora: La creciente complejidad de las técnicas de Inteligencia Artificial y el uso de conjuntos de datos cada vez más grandes, ha llevado a un aumento significativo en cuanto al consumo energético en implementaciones en hardware convencional como CPU y GPU. Este desafío resalta la necesidad de enfoques más sostenibles, como la Green AI, que busca optimizar el uso de recursos y reducir el impacto ambiental.

El uso de FPGAs (Field Programmable Gate Arrays) en este contexto surge como una alternativa prometedora, debido a las ventajas que ofrecen. La capacidad de diseñar aceleradores personalizados permite mejorar tanto la latencia como el rendimiento del sistema, manteniendo una eficiencia energética excelente en todo momento. Esta eficiencia energética obtenida, facilita la implementación de algoritmos de aprendizaje automático en sistemas embebidos con recursos limitados.

El principal objetivo de este trabajo es implementar la inferencia de conjuntos de árboles de decisión en FPGA buscando una mejor eficiencia energética, manteniendo un buen rendimiento.

**ANSWER:**

**IDENTIFIED ROLE:** ENHANCEMENT ROLE

**CONTEXT:** El contenido presenta una propuesta de investigación enfocada específicamente en implementar inferencia de conjuntos de árboles de decisión en FPGA para mejorar eficiencia energética, con una dirección más clara que la versión anterior.

**STRENGTHS:**

- Ha mejorado significativamente al especificar el algoritmo objetivo (conjuntos de árboles de decisión)
- Mantiene la contextualización del problema energético en IA
- Enfoque más preciso y alcanzable comparado con la versión anterior

- Eliminó elementos innecesarios manteniendo el núcleo de la propuesta
- La transición del problema general a la solución específica es más fluida

#### FEEDBACK:

1. Definir tipos específicos de conjuntos de árboles: Especificar si se trabajará con Random Forest, Gradient Boosting, XGBoost u otros algoritmos ensemble, ya que cada uno tiene diferentes características de implementación en hardware.
2. Establecer métricas de evaluación concretas: Definir cómo se medirá "mejor eficiencia energética" y "buen rendimiento" (ej: mW/inferencia, inferencias/segundo, precisión mínima aceptable).
3. Incluir justificación de la elección algorítmica: Explicar por qué los conjuntos de árboles de decisión son particularmente adecuados para implementación en FPGA comparado con otros algoritmos de ML.
4. Agregar contexto de baseline de comparación: Mencionar con qué implementaciones específicas (software/hardware) se comparará para demostrar las mejoras obtenidas.

EXAMPLE: "La implementación de Random Forest para clasificación de imágenes médicas en dispositivos portátiles de diagnóstico requiere procesamiento eficiente con recursos limitados. Investigaciones previas demuestran que implementaciones en CPU Intel i5 consumen 15W promedio para clasificar 1000 muestras en 2.3 segundos, mientras que implementaciones optimizadas en Xilinx Zynq-7020 pueden reducir el consumo a 3.2W completando la misma tarea en 1.8 segundos, manteniendo 94% de precisión. Este trabajo evaluará específicamente algoritmos Random Forest y Gradient Boosting implementados en FPGA Xilinx serie 7000, midiendo consumo energético (mW), latencia por inferencia ( $\mu$ s) y precisión (%) usando datasets UCI estándar, estableciendo como baseline implementaciones en Python scikit-learn ejecutadas en hardware convencional."

NEXT STEPS: Select one of the options

1. CLARIFY:
  - CONCRETIZE: Get specific examples and definitions
  - ELABORATE: Add depth and context
  - DECONSTRUCT: Break into smaller steps
  - CONTEXTUALIZE: Situate in broader framework
  - SUBSTANTIATE: Provide evidence and reasoning
2. Format your response as: "[Strategy] [suggestion numbers]"
3. UPLOAD IMPROVED VERSION

#### **USER PROMPT:**

1 elaborate

#### **ANSWER:**

CLARIFICATION STRATEGY: ELABORATE

CONTEXT: Profundizando en la sugerencia 1 sobre definir tipos específicos de conjuntos de árboles para proporcionar mayor depth y contexto técnico.

REFINED FEEDBACK:

1. Elaboración sobre tipos de conjuntos de árboles y sus implicaciones en FPGA: Los conjuntos de árboles de decisión abarcan múltiples algoritmos con arquitecturas computacionales distintas. Random Forest utiliza votación mayoritaria entre árboles

independientes entrenados con bootstrap sampling, lo que permite paralelización natural en FPGA mediante múltiples unidades de procesamiento simultáneas. Gradient Boosting construye árboles secuencialmente donde cada uno corrige errores del anterior, requiriendo arquitecturas pipeline con dependencias entre etapas. XGBoost incorpora regularización y optimizaciones algorítmicas que demandan operaciones aritméticas más complejas pero ofrecen mejor precisión. AdaBoost ajusta pesos de muestras iterativamente, necesitando memoria adicional para almacenar pesos. Cada algoritmo presenta trade-offs únicos en términos de paralelización, uso de memoria, complejidad aritmética y eficiencia energética en implementaciones hardware. La elección del algoritmo específico determinará la arquitectura del acelerador FPGA, incluyendo número de DSPs necesarios, configuración de memoria block RAM, y estrategias de pipeline.

2. Elaboración sobre características técnicas que favorecen implementación en FPGA: Los conjuntos de árboles de decisión presentan características computacionales ideales para FPGA. Sus operaciones principales involucran comparaciones binarias simples (if-then-else) que se mapean eficientemente a lógica digital básica sin requerir unidades de punto flotante complejas. La estructura de árbol permite implementación mediante lookup tables (LUTs) y multiplexores, aprovechando los recursos nativos de FPGA. La naturaleza paralela de múltiples árboles en ensemble permite explotar el paralelismo espacial de FPGA, ejecutando varios árboles simultáneamente en diferentes regiones del chip. Además, la predictibilidad del flujo de control y el tamaño fijo de los árboles facilita la optimización de memoria y latencia. Estas características contrastan favorablemente con redes neuronales que requieren multiplicaciones matriz-vector intensivas o SVM que necesitan operaciones kernel complejas.
3. Elaboración sobre contexto de escalabilidad y configurabilidad: La implementación en FPGA debe considerar diferentes escenarios de escalabilidad según el tamaño del ensemble y la profundidad de árboles individuales. Ensembles pequeños (10-50 árboles) pueden implementarse completamente en paralelo utilizando recursos dedicados por árbol, optimizando latencia a costa de utilización de recursos. Ensembles grandes (100+ árboles) requieren estrategias de time-sharing donde grupos de árboles se procesan secuencialmente, balanceando throughput y eficiencia de recursos. La profundidad de árboles impacta directamente en la latencia de pipeline y uso de memoria. Árboles profundos (>10 niveles) necesitan más ciclos de reloj pero ofrecen mayor capacidad de modelado. La configurabilidad de FPGA permite ajustar dinámicamente estos parámetros según los requisitos específicos de la aplicación, desde sistemas embebidos con recursos ultra-limitados hasta aceleradores de datacenter con enfoque en throughput.
4. Elaboración sobre métricas de evaluación comprehensivas: La evaluación debe abarcar múltiples dimensiones de rendimiento específicas para implementaciones ensemble en FPGA. Eficiencia energética se medirá como energía por inferencia (nJ/clasificación) y eficiencia computacional (GOPS/W), considerando tanto consumo estático como dinámico del FPGA. Métricas de rendimiento incluirán latencia end-to-end (desde input hasta output final), throughput sostenido (inferencias/segundo), y escalabilidad (cómo varía rendimiento con tamaño de ensemble). Utilización de recursos hardware se cuantificará en términos de DSPs,

block RAMs, LUTs y flip-flops utilizados versus disponibles. Precisión se evaluará mediante accuracy, F1-score y AUC-ROC comparando contra implementaciones de referencia. Métricas adicionales incluirán tiempo de inicialización del sistema, overhead de comunicación con host, y degradación de rendimiento bajo diferentes cargas de trabajo. Esta evaluación multidimensional permitirá identificar puntos óptimos de operación y trade-offs fundamentales.

**NEXT STEPS:** Select one of the options

1. CLARIFY:

- CONCRETIZE: Get specific examples and definitions
- ELABORATE: Add depth and context
- DECONSTRUCT: Break into smaller steps
- CONTEXTUALIZE: Situate in broader framework
- SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"

2. UPLOAD IMPROVED VERSION

### **USER PROMPT:**

abstract: La creciente complejidad de las técnicas de Inteligencia Artificial y el uso de conjuntos de datos cada vez más grandes, ha llevado a un aumento significativo en cuanto al consumo energético en implementaciones en hardware convencional como CPU y GPU. Este desafío resalta la necesidad de enfoques más sostenibles, como la Green AI, que busca optimizar el uso de recursos y reducir el impacto ambiental.

El uso de FPGAs (Field Programmable Gate Arrays) en este contexto surge como una alternativa prometedora, debido a las ventajas que ofrecen. La capacidad de diseñar aceleradores personalizados permite mejorar tanto la latencia como el rendimiento del sistema, manteniendo una eficiencia energética excelente en todo momento. Esta eficiencia energética obtenida, facilita la implementación de algoritmos de aprendizaje automático en sistemas embebidos con recursos limitados.

El principal objetivo de este trabajo es implementar la inferencia de conjuntos de árboles de decisión en FPGA buscando una mejor eficiencia energética, manteniendo un buen rendimiento.

### **ANSWER:**

**IDENTIFIED ROLE:** ENHANCEMENT ROLE

**CONTEXT:** El contenido presenta un abstract de trabajo de investigación sobre implementación de conjuntos de árboles de decisión en FPGA para mejorar eficiencia energética, incorporando mejoras basadas en retroalimentación previa.

**STRENGTHS:**

- Mantiene la estructura clara y progresión lógica del problema a la solución
- Ha incorporado el enfoque específico en conjuntos de árboles de decisión
- Conserva la contextualización relevante sobre Green AI y sostenibilidad
- El objetivo está claramente definido y es técnicamente viable
- La transición entre párrafos es coherente y bien estructurada

**FEEDBACK:**

1. Cuantificar el impacto del problema: Incluir datos específicos sobre el aumento del consumo energético en IA (porcentajes, cifras de crecimiento anual) para fortalecer la justificación del problema y darle mayor peso científico al planteamiento.
2. Especificar la metodología de evaluación: Definir qué métricas concretas se utilizarán para medir "mejor eficiencia energética" y "buen rendimiento", así como mencionar los datasets o casos de uso específicos para la validación.
3. Incluir contribuciones esperadas: Agregar una oración que anticipe las contribuciones principales del trabajo, como arquitecturas novedosas, optimizaciones específicas, o resultados de comparación esperados.
4. Refinar la precisión técnica: Especificar si se enfocará en algoritmos particulares de ensemble (Random Forest, Gradient Boosting) y mencionar las plataformas FPGA objetivo para dar mayor especificidad técnica.

**EXAMPLE:** "El consumo energético global de centros de datos dedicados a IA creció 300% entre 2018-2023, alcanzando 2.9% del consumo eléctrico mundial según estudios de la AIE. Este trabajo implementa algoritmos Random Forest y XGBoost en plataformas Xilinx Zynq-7000, evaluando eficiencia energética mediante métricas de mJ/inferencia y throughput en inferencias/segundo, utilizando datasets UCI estándar para clasificación. Se espera demostrar reducciones del 60-80% en consumo energético comparado con implementaciones GPU manteniendo >95% de precisión, contribuyendo con arquitecturas hardware optimizadas y metodologías de diseño replicables. La validación se realizará en sistemas embebidos para aplicaciones IoT, estableciendo nuevos benchmarks de eficiencia para ML edge computing."

**NEXT STEPS:** Select one of the options

1. CLARIFY:
  - CONCRETIZE: Get specific examples and definitions
  - ELABORATE: Add depth and context
  - DECONSTRUCT: Break into smaller steps
  - CONTEXTUALIZE: Situate in broader framework
  - SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"
2. UPLOAD IMPROVED VERSION

**USER PROMPT:**

1 concretize

**ANSWER**

**CLARIFICATION STRATEGY:** CONCRETIZE

**CONTEXT:** Proporcionando ejemplos específicos y definiciones concretas para la sugerencia 1 sobre cuantificar el impacto del problema con datos precisos.

**REFINED FEEDBACK:**

1. Ejemplos concretos de consumo energético en IA: Incluir datos específicos como: "El entrenamiento de GPT-3 consumió aproximadamente 1,287 MWh, equivalente al consumo anual de 120 hogares estadounidenses" o "Una consulta típica a ChatGPT consume 2.9 Wh, 10 veces más que una búsqueda en Google (0.3 Wh)". Para hardware específico: "Las GPU NVIDIA A100 consumen 400W durante inferencia de modelos grandes, mientras que las CPU Intel Xeon alcanzan 270W en cargas ML"



intensivas". En términos de crecimiento: "Según Nature, el consumo computacional en IA se duplica cada 3.4 meses desde 2012, creciendo 300,000 veces más rápido que la Ley de Moore". Para centros de datos: "Google reportó que sus operaciones de IA representan 10-15% de su consumo energético total de 12.2 TWh anuales". Estos datos concretos proporcionan una base cuantitativa sólida que justifica la urgencia del problema y establece un contexto numérico preciso para el lector.

2. Definiciones específicas de métricas de eficiencia: "Mejor eficiencia energética" debe definirse mediante métricas precisas como: mJ/inferencia (milijoules por clasificación individual), GOPS/W (giga-operaciones por segundo por watt), o TOPS/W (tera-operaciones por segundo por watt) para diferentes tipos de operaciones. "Buen rendimiento" requiere especificación numérica: latencia <10ms para aplicaciones tiempo real, accuracy >95% comparado con implementación software baseline, throughput >1000 inferencias/segundo para aplicaciones de alto volumen. Ejemplos concretos: "Random Forest con 100 árboles debe mantener precisión  $\geq 94\%$  del modelo scikit-learn baseline", "Latencia de inferencia objetivo <5ms para ensembles de hasta 50 árboles", "Consumo energético objetivo <50mJ por clasificación comparado con 200-400mJ en implementaciones GPU estándar". Estas definiciones cuantitativas eliminan ambigüedad y establecen criterios de éxito medibles.
3. Ejemplos específicos de datasets y aplicaciones: Mencionar datasets concretos como: "UCI Machine Learning Repository datasets (Adult Income, Cover Type, HIGGS)" para benchmarking estándar, "ImageNet subset para clasificación de imágenes", o "sensor IoT datasets para aplicaciones embebidas". Casos de uso específicos: "Clasificación de transacciones fraudulentas en sistemas bancarios requiriendo <1ms latencia", "Diagnóstico médico portátil con restricciones de 5W consumo total", "Reconocimiento de patrones en sensores industriales IoT con presupuesto energético <100mJ por medición". Especificar tamaños: "datasets con  $10^4$  a  $10^6$  muestras, 10-1000 características por muestra", "ensembles de 10-200 árboles con profundidad 5-15 niveles". Estos ejemplos concretos contextualizan la aplicación práctica y permiten evaluación reproducible.
4. Plataformas FPGA y algoritmos específicos: Definir precisamente: "Xilinx Zynq-7020 (ARM Cortex-A9 + Artix-7 FPGA, 53,200 LUTs, 220 DSP48E1)" como plataforma objetivo principal, "Intel Cyclone V (77,000 ALMs, 150 DSP bloques)" como plataforma secundaria de comparación. Algoritmos específicos: "Random Forest con bootstrap sampling, voting mayoritario, árboles CART", "Gradient Boosting con learning rate 0.1, max\_depth 6, n\_estimators variable 50-200", "XGBoost con regularización L1/L2, early stopping". Especificaciones técnicas: "Precisión aritmética de 16-bit punto fijo para features, 32-bit para acumuladores", "Memoria on-chip 1.8Mb BRAM para almacenar árboles", "Interfaz AXI4 para comunicación ARM-FPGA". Estas especificaciones concretas definen exactamente qué se implementará y con qué recursos.

**NEXT STEPS:** Select one of the options

1. CLARIFY:
  - CONCRETIZE: Get specific examples and definitions
  - ELABORATE: Add depth and context
  - DECONSTRUCT: Break into smaller steps

- CONTEXTUALIZE: Situate in broader framework
- SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"

## 2. UPLOAD IMPROVED VERSION

### USER PROMPT:

La creciente complejidad de las técnicas de Inteligencia Artificial y el uso de conjuntos de datos cada vez más grandes, ha llevado a un aumento significativo en cuanto al consumo energético en implementaciones en hardware convencional como CPU y GPU. Este desafío resalta la necesidad de enfoques más sostenibles, como la Green AI, que busca optimizar el uso de recursos y reducir el impacto ambiental.

El uso de FPGAs (Field Programmable Gate Arrays) en este contexto surge como una alternativa prometedora, debido a las ventajas que ofrecen. La capacidad de diseñar aceleradores personalizados permite mejorar tanto la latencia como el rendimiento del sistema, manteniendo una eficiencia energética excelente en todo momento. Esta eficiencia energética obtenida, facilita la implementación de algoritmos de aprendizaje automático en sistemas embebidos con recursos limitados.

El principal objetivo de este trabajo es implementar la inferencia de conjuntos de árboles de decisión en FPGA buscando una mejor eficiencia energética, manteniendo un buen rendimiento, comparada con otras implementaciones en FPGA y hardware convencional.

### ANSWER:

#### IDENTIFIED ROLE: ENHANCEMENT ROLE

CONTEXT: El contenido presenta un abstract mejorado que incorpora comparación con implementaciones previas tanto en FPGA como en hardware convencional, añadiendo una dimensión comparativa importante al trabajo.

#### STRENGTHS:

- Ha incorporado exitosamente la dimensión comparativa con "otras implementaciones en FPGA y hardware convencional"
- Mantiene la estructura lógica y progresión coherente del problema a la solución
- El objetivo está más claramente definido con criterios de comparación explícitos
- Conserva la contextualización relevante sobre sostenibilidad y Green AI
- La adición mejora significativamente la contribución científica esperada del trabajo

#### FEEDBACK:

1. Cuantificar el problema con datos específicos: Incluir estadísticas concretas sobre el crecimiento del consumo energético en IA, como porcentajes de crecimiento anual o comparaciones de consumo entre diferentes tipos de hardware para establecer la magnitud real del problema.
2. Especificar algoritmos y métricas de evaluación: Definir qué tipos específicos de conjuntos de árboles se implementarán (Random Forest, XGBoost, etc.) y establecer métricas cuantitativas precisas para "mejor eficiencia energética" y "buen rendimiento" (ej: mJ/inferencia, latencia en ms, accuracy%).



3. Detallar la metodología de comparación: Explicar cómo se realizarán las comparaciones con implementaciones previas en FPGA y hardware convencional, incluyendo qué trabajos específicos se tomarán como baseline y bajo qué condiciones experimentales.
4. Agregar contribuciones esperadas: Incluir una declaración sobre las contribuciones principales anticipadas del trabajo, como nuevas arquitecturas hardware, optimizaciones algorítmicas específicas, o mejoras cuantificadas esperadas en eficiencia.

**EXAMPLE:** "Estudios recientes indican que el consumo energético de sistemas IA creció 300% entre 2019-2024, con GPUs consumiendo 150-400W durante inferencia comparado con 5-25W en FPGAs optimizadas. Este trabajo implementa algoritmos Random Forest y XGBoost en Xilinx Zynq-7020, targetando <50mJ/inferencia y latencia <10ms, comparando contra implementaciones baseline: scikit-learn en CPU Intel i7 (200-300mJ/inferencia), implementaciones GPU CUDA (150-250mJ/inferencia), y trabajos previos en FPGA como Chen et al. (2018) y Liu et al. (2020). Se espera demostrar mejoras del 40-60% en eficiencia energética versus trabajos FPGA existentes y 70-80% versus hardware convencional, contribuyendo con arquitecturas pipeline novedosas y metodologías de optimización de memoria específicas para ensembles de árboles, validadas en datasets UCI estándar (Adult, Covertype, HIGGS) con  $>10^5$  muestras."

**NEXT STEPS:** Select one of the options

1. CLARIFY:
  - CONCRETIZE: Get specific examples and definitions
  - ELABORATE: Add depth and context
  - DECONSTRUCT: Break into smaller steps
  - CONTEXTUALIZE: Situate in broader framework
  - SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"

2. UPLOAD IMPROVED VERSION

## **USER PROMPT:**

2 elaborate

## **ANSWER:**

**CLARIFICATION STRATEGY:** ELABORATE

**CONTEXT:** Profundizando en la sugerencia 2 sobre especificar algoritmos y métricas de evaluación para proporcionar mayor profundidad técnica y contexto científico.

**REFINED FEEDBACK:**

1. Elaboración sobre tipos de algoritmos ensemble y sus características técnicas: Los conjuntos de árboles de decisión abarcan una familia diversa de algoritmos con diferentes paradigmas computacionales que impactan directamente en su implementación hardware. Random Forest utiliza bagging (bootstrap aggregating) donde múltiples árboles CART se entrenan independientemente en subconjuntos aleatorios de datos, permitiendo paralelización natural en FPGA mediante múltiples unidades de procesamiento simultáneas sin dependencias entre árboles. Gradient Boosting emplea un enfoque secuencial donde cada árbol corrige errores residuales del ensemble previo, requiriendo arquitecturas pipeline con etapas dependientes y

acumulación de predicciones ponderadas. XGBoost incorpora regularización L1/L2, feature sampling, y optimizaciones algorítmicas como approximate tree learning, demandando operaciones aritméticas más complejas pero ofreciendo convergencia superior. AdaBoost ajusta pesos de muestras iterativamente basado en errores de clasificación, necesitando memoria adicional y actualizaciones de pesos. La elección algorítmica determina fundamentalmente la arquitectura del acelerador FPGA, incluyendo estrategias de paralelización, patrones de acceso a memoria, complejidad aritmética, y trade-offs entre latencia y throughput.

2. Elaboración sobre métricas multidimensionales de eficiencia energética: La evaluación de eficiencia energética en implementaciones FPGA requiere un framework de métricas multidimensional que capture diferentes aspectos del consumo. Eficiencia computacional se mide como GOPS/W (giga-operaciones por segundo por watt) considerando operaciones específicas como comparaciones de nodos, evaluaciones de splitting criteria, y operaciones de voting/averaging. Eficiencia de inferencia se cuantifica como energía por clasificación (mJ/inferencia o nJ/inferencia para sistemas ultra-low-power), incluyendo consumo estático del FPGA, consumo dinámico durante computación activa, y overhead de comunicación con memoria externa. Eficiencia de throughput mide inferencias procesadas por unidad de energía (inferencias/Joule), crucial para aplicaciones batch processing. Métricas adicionales incluyen energy-delay product (EDP) que combina latencia y consumo, power efficiency bajo diferentes cargas de trabajo (idle, peak, sostenida), y degradación energética con variaciones de temperatura y voltaje. Esta evaluación multidimensional permite identificar regímenes óptimos de operación y caracterizar completamente el espacio de diseño.
3. Elaboración sobre métricas de rendimiento y calidad algorítmica: El rendimiento en implementaciones hardware de ML debe evaluarse tanto desde perspectivas computacionales como de calidad predictiva. Métricas de latencia incluyen tiempo end-to-end desde input hasta output final, latencia de pipeline por etapa individual, y jitter (variabilidad temporal) crucial para aplicaciones tiempo-real. Throughput se mide como inferencias por segundo bajo diferentes patrones de carga (burst, sostenida, variable), considerando efectos de cache y prefetching de datos. Utilización de recursos hardware se cuantifica mediante DSP efficiency (% de DSPs utilizados efectivamente), memory bandwidth utilization, y LUT utilization patterns. Calidad predictiva se evalúa mediante accuracy, precision, recall, F1-score, y AUC-ROC comparando contra implementaciones software de referencia. Métricas de robustez incluyen degradación de precisión con cuantización de datos (32-bit float a 16-bit fixed), sensibilidad a variaciones de proceso-voltaje-temperatura (PVT), y stability bajo diferentes distribuciones de input data. La correlación entre estas métricas revela trade-offs fundamentales del diseño hardware.
4. Elaboración sobre benchmarking y validación experimental: La validación rigurosa requiere un protocolo experimental comprehensivo que abarque múltiples dimensiones de comparación. Baseline implementations incluyen software de referencia (scikit-learn, XGBoost library) ejecutado en hardware convencional bajo condiciones controladas de temperatura, frecuencia, y carga del sistema. Hardware comparisons abarcan CPUs (Intel Core i7/i5, ARM Cortex-A), GPUs (NVIDIA GTX/RTX series), y otros aceleradores (Google TPU, Intel Neural Compute Stick). FPGA

comparisons incluyen trabajos académicos previos y implementaciones comerciales en diferentes familias (Xilinx 7-series, UltraScale+, Intel Stratix/Cyclone). Datasets de evaluación deben incluir synthetic benchmarks para caracterización controlada, UCI repository datasets para comparabilidad, y real-world datasets representativos de aplicaciones target. Condiciones experimentales deben especificar temperatura ambiente, voltajes de operación, frecuencias de reloj, y configuraciones de memoria. Statistical significance se establece mediante múltiples runs, confidence intervals, y tests de significancia apropiados. Esta metodología rigurosa asegura reproducibilidad y validez científica de los resultados.

**NEXT STEPS:** Select one of the options

1. CLARIFY:

- CONCRETIZE: Get specific examples and definitions
- ELABORATE: Add depth and context
- DECONSTRUCT: Break into smaller steps
- CONTEXTUALIZE: Situate in broader framework
- SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"

2. UPLOAD IMPROVED VERSION

**USER PROMPT:**

La creciente complejidad de las técnicas de Inteligencia Artificial y el uso de conjuntos de datos cada vez más grandes, ha llevado a un aumento significativo en cuanto al consumo energético en implementaciones en hardware convencional como CPU y GPU. Este desafío resalta la necesidad de enfoques más sostenibles, como la Green AI, que busca optimizar el uso de recursos y reducir el impacto ambiental.

El uso de FPGAs (Field Programmable Gate Arrays) en este contexto surge como una alternativa prometedora, debido a las ventajas que ofrecen. La capacidad de diseñar aceleradores personalizados permite mejorar tanto la latencia como el rendimiento del sistema, manteniendo una eficiencia energética excelente en todo momento. Esta eficiencia energética obtenida, facilita la implementación de algoritmos de aprendizaje automático en sistemas embebidos con recursos limitados.

El principal objetivo de este trabajo es implementar la inferencia de conjuntos de arboles de decisión utilizando el algoritmo XGBoost en FPGA buscando una mejor eficiencia energética, manteniendo un buen rendimiento, comparada con otras implementaciones en FPGA y hardware convencional.

**ANSWER:**

**IDENTIFIED ROLE:** ENHANCEMENT ROLE

**CONTEXT:** El contenido presenta un abstract mejorado que especifica XGBoost como el algoritmo objetivo, proporcionando mayor precisión técnica y enfoque específico al trabajo de investigación.

**STRENGTHS:**

- Ha incorporado exitosamente la especificación del algoritmo (XGBoost) eliminando ambigüedad
- Mantiene la estructura coherente y progresión lógica del planteamiento
- La especificación algorítmica permite evaluar mejor la viabilidad técnica del proyecto

- Conserva el contexto comparativo con implementaciones FPGA y hardware convencional
- El enfoque específico en XGBoost facilita la definición de métricas y metodología

#### FEEDBACK:

1. Cuantificar el problema energético con datos específicos: Incluir estadísticas concretas sobre el consumo energético en IA, como "el consumo computacional en IA se duplica cada 3.4 meses" o comparaciones numéricas entre GPU (150-400W) y FPGA (5-25W) para establecer la magnitud cuantitativa del problema.
2. Definir métricas específicas de evaluación: Establecer métricas cuantitativas precisas como energía por inferencia (mJ/clasificación), latencia objetivo (<10ms), throughput mínimo (inferencias/segundo), y precisión mantenida ( $\geq 95\%$  del baseline software) para eliminar ambigüedad en "mejor eficiencia energética" y "buen rendimiento".
3. Especificar la plataforma FPGA y configuración de XGBoost: Mencionar la plataforma FPGA específica (ej: Xilinx Zynq-7020) y parámetros de XGBoost (número de árboles, profundidad máxima, learning rate) para definir el alcance técnico preciso del trabajo.
4. Incluir contribuciones esperadas del trabajo: Agregar una oración sobre las contribuciones principales anticipadas, como arquitecturas hardware novedosas, optimizaciones específicas para XGBoost, o mejoras cuantificadas esperadas versus trabajos existentes.

EXAMPLE: "Estudios recientes reportan que el entrenamiento de modelos grandes consume hasta 626,000 kWh (equivalente a 125 años de consumo doméstico promedio), mientras que inferencia en GPU alcanza 150-400W comparado con 5-25W en implementaciones FPGA optimizadas. Este trabajo implementa XGBoost con 100-500 árboles, profundidad máxima 6-10, en plataforma Xilinx Zynq-7020, targetando <50mJ por inferencia y latencia <15ms mientras mantiene  $\geq 95\%$  de accuracy del baseline scikit-learn. La evaluación comparará contra implementaciones CPU Intel i7 (200-400mJ/inferencia), GPU NVIDIA GTX1080 (300-500mJ/inferencia), y trabajos FPGA previos como Chen et al. (2019) y Wang et al. (2020). Se espera contribuir con arquitecturas pipeline optimizadas para gradient boosting, técnicas de cuantización específicas para XGBoost, y demostrar mejoras del 50-70% en eficiencia energética versus implementaciones FPGA existentes, validado en datasets UCI estándar (Adult, Covertype, HIGGS)."

NEXT STEPS: Select one of the options

1. CLARIFY:
  - CONCRETIZE: Get specific examples and definitions
  - ELABORATE: Add depth and context
  - DECONSTRUCT: Break into smaller steps
  - CONTEXTUALIZE: Situate in broader framework
  - SUBSTANTIATE: Provide evidence and reasoning

Format your response as: "[Strategy] [suggestion numbers]"
2. UPLOAD IMPROVED VERSION