UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

RECONSTRUCCIÓN CRANEAL Y SEGMENTACIÓN ROBUSTA DE IMÁGENES MÉDICAS MEDIANTE APRENDIZAJE PROFUNDO

Victor Franco Matzkin

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERÍA
Mención en Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2026

Secretaría de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje “El Pozo”,
S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

RECONSTRUCCIÓN CRANEAL Y SEGMENTACIÓN ROBUSTA DE IMÁGENES MÉDICAS MEDIANTE APRENDIZAJE PROFUNDO

Victor Franco Matzkin

Lugar de Trabajo:

SINC(*I*)
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:
Dr. Enzo Ferrante — ICC, CONICET-UBA

Co-director:
Dr. Diego H. Milone — sinc(*i*), CONICET-UNL

Jurado Evaluador:
Dra. María Elena Buemi — FCEyN, UBA
Dr. José M. Massa — Fac. Cs. Exactas, UNICEN
Dr. Leandro Vignolo — UNL-CONICET

2026

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 10 de marzo de 2026

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Reconstrucción craneal y segmentación robusta de imágenes médicas mediante aprendizaje profundo"***, desarrollada por el Ing. Víctor Franco MATZKIN, en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de la versión digital final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dra. María Elena Buemi	Dr. José Massa	Dr. Leandro Vignolo

Santa Fe, 10 de marzo de 2026

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

…….	…….
Dr. Diego Humberto Milone	Dr. Enzo Ferrante
Codirector de Tesis	Director de Tesis

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS · SANTA FE

Dr. José Luis Macor
Director de Posgrado
FICH - UNL

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

DECLARACIÓN DEL AUTOR

Esta disertación ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería ante la Universidad Nacional del Litoral y ha sido depositada en Repositorio Institucional de Acceso Abierto -RIAA- de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disponible a sus lectores bajo las condiciones estipuladas.

Citaciones breves de esta disertación son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para una citación extendida o para la reproducción parcial o total de este manuscrito serán concedidos por el portador legal del derecho de propiedad intelectual de la obra.

Victor Franco Matzkin

TESIS POR COMPILACIÓN

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución No 255/17 (Expte. No 888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

"En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión."

AGRADECIMIENTOS

A Enzo y Diego, mis directores, por la paciencia, la perseverancia y la confianza. Por abrirme puertas, aportar siempre y apoyarme en cada etapa de este camino.

A mis compañeros del **sinc(*i*)**: becarios, CPAs, investigadores y docentes. Por los mates, los almuerzos, los seminarios y tantas horas compartidas. Por el apoyo de siempre. Gracias por hacer del instituto mucho más que un lugar de trabajo.

Al CONICET y a la Universidad Nacional del Litoral, mi segunda casa durante todos estos años.

A mi familia y amigos, por el apoyo incondicional. Por estar siempre, incluso cuando todo se vuelve difícil.

Al jurado evaluador, por su tiempo, su predisposición y sus aportes durante el proceso de revisión.

A la educación pública argentina, sin la cual nada de esto hubiera sido posible.

Franco Matzkin
Santa Fe, 2026.

Índice general

Índice de figuras

Índice de tablas

RESUMEN

Esta tesis de doctorado se centra en el uso de técnicas de aprendizaje profundo para la reconstrucción de implantes craneales y la segmentación robusta de imágenes médicas, abordando las limitaciones de los datos y el desafío del cambio de dominio. Debido al costo y tiempo que implica el diseño manual de implantes craneales, se proponen enfoques novedosos basados en aprendizaje autosupervisado. En estos se simula la extracción del hueso en tomografías computarizadas de cráneos completos, generando datos sintéticos que compensan la escasez de datos reales de pacientes. Este procedimiento, denominado "craniectomía virtual", se ha mejorado mediante la incorporación de patrones de simulación más complejos y variables. Las estrategias planteadas superan en tiempo y desempeño a métodos alternativos como los modelos de forma estadística y otros enfoques de aprendizaje profundo. Además, se explora la utilización de restricciones o conocimiento anatómico previo sobre la forma esperada, que imita la variabilidad de los defectos craneales reales en la entrada del modelo, con el objetivo de mejorar su solidez al enfrentar implantes con formas fuera de distribución y de mayor resolución.

En la segunda parte, la tesis se enfoca en estudiar la robustez de los modelos de segmentación de imágenes médicas frente a variaciones en la adquisición de imágenes y las poblaciones de pacientes. Se estudia la segmentación de hiperintensidades de la sustancia blanca en resonancias magnéticas cerebrales, evaluando técnicas de regularización por entropía para mejorar la estimación de incertidumbre en las predicciones, especialmente bajo condiciones de cambio de dominio. Se analizó la correlación entre las estimaciones de incertidumbre y los errores de segmentación, encontrando que métodos de regularización por entropía, como la entropía máxima en predicciones erróneas, mejoran la capacidad del modelo para expresar incertidumbre en áreas de mayor ambigüedad. Los resultados demuestran que estas estrategias pueden detectar cambios de dominio y servir como indicadores de errores de segmentación en ausencia de anotaciones manuales, con mejoras significativas en la calibración del modelo tanto en datos dentro como fuera de distribución.

La investigación propuesta en esta tesis ofrece avances novedosos en términos metodológicos en el área del aprendizaje automático, con aplicaciones en dos problemas de análisis computacional de imágenes cerebrales. Las soluciones propuestas contribuyen a mejorar la precisión, eficiencia y robustez tanto en la reconstrucción de implantes craneales como en la segmentación de imágenes médicas, con un enfoque hacia su aplicación clínica. La validación experimental en múltiples conjuntos de datos, incluyendo casos clínicos reales y competencias internacionales, demuestra el potencial de estas metodologías para mejorar significativamente los procesos de planificación quirúrgica y diagnóstico asistido por computadora, por lo que se espera que resulte finalmente en una mejora del cuidado de pacientes.

ABSTRACT

This doctoral thesis focuses on the use of deep learning techniques for cranial implant reconstruction and robust medical image segmentation, addressing data limitations and the challenge of domain shift. Due to the cost and time involved in manual cranial implant design, novel approaches based on self-supervised learning are proposed. These simulate bone extraction from computed tomography scans of complete skulls, generating synthetic data that compensates for the scarcity of real patient data. This procedure, called "virtual craniectomy," has been improved through the incorporation of more complex and variable simulation patterns. The proposed strategies outperform alternative methods such as statistical shape models and other deep learning approaches in both time and performance. Additionally, the use of anatomical constraints or prior knowledge about the expected shape is explored, which mimics the variability of real cranial defects in the model input, with the aim of improving its robustness when facing out-of-distribution implants at higher resolutions.

In the second part, the thesis focuses on studying the robustness of medical image segmentation models against variations in image acquisition and patient populations. The segmentation of white matter hyperintensities in brain magnetic resonance images is studied, evaluating entropy regularization techniques to improve uncertainty estimation in predictions, especially under domain shift conditions. The correlation between uncertainty estimates and segmentation errors was analyzed, finding that entropy regularization methods, such as maximum entropy on erroneous predictions, improve the model's ability to express uncertainty in areas of greater ambiguity. The results demonstrate that these strategies can detect domain shifts and serve as indicators of segmentation errors in the absence of manual annotations, with significant improvements in model calibration on both in-distribution and out-of-distribution data.

The research proposed in this thesis offers novel methodological advances in the field of machine learning, with applications in two problems of computational brain image analysis. The proposed solutions contribute to improving the precision, efficiency, and robustness of both cranial implant reconstruction and medical image segmentation, with a focus on clinical application. Experimental validation on multiple datasets, including real clinical cases and international competitions, demonstrates the potential of these methodologies to significantly improve surgical planning and computer-aided diagnosis processes, which is expected to ultimately result in improved patient care.

Capítulo 1:
Introducción

Durante los últimos años, la adopción de neuroimágenes para el diagnóstico y tratamiento médico ha crecido ampliamente. Esta tecnología es utilizada en una gran variedad de contextos de aplicación, que abarcan desde evaluar el estado y realizar pronósticos de patologías neurológicas, hasta la planificación de procedimientos quirúrgicos. El análisis de dichas imágenes con el objetivo de extraer información de utilidad clínica es, en la actualidad, asistido por sistemas informáticos que utilizan métodos computacionales, en su mayoría basados en aprendizaje profundo [29]. Estos métodos permiten realizar tareas de interés como la clasificación, segmentación, reconstrucción o registración de imágenes. En esta tesis doctoral, motivada por la necesidad de mejorar la precisión y eficiencia en la planificación de cirugías craneales y el diagnóstico de lesiones cerebrales, el foco estará puesto en nuevos aportes metodológicos para las tareas de reconstrucción y segmentación, con aplicación en distintos contextos clínicos.

El primer problema a abordar será la reconstrucción de implantes para cirugías craneales en el contexto de pacientes con traumatismo craneoencefálico (TCE). El TCE es una de las causas más comunes de muerte en la población adulta, con casi 50 millones de personas que lo sufren anualmente [5]. Cuando se produce dicha lesión, algunos pacientes se someten a una cirugía craneal para evitar posibles daños asociados al aumento de la presión intracraneal. En esta cirugía, conocida como craniectomía descompresiva (CD), se extrae un trozo de hueso del cráneo (denominado "colgajo óseo") para aliviar la presión y, la mayoría de las veces, se reinserta, dependiendo del riesgo de infección considerado. En muchos casos, cuando el colgajo no puede reinsertarse debido a los riesgos de infección, es necesario fabricar implantes para sustituir el hueso extraído (ver Figura 1.1). La mayoría de las veces, esta reconstrucción es realizada manualmente por diseñadores expertos. Los enfoques automáticos existentes emplean técnicas asistidas por computadora, en las que se aprovecha la simetría de la cabeza para completar la parte del cráneo que falta [16]. Sin embargo, esta técnica tiene la restricción de admitir solo craniectomías unilaterales, es decir, que no se podría aplicar cuando el defecto craneal esté presente en ambos hemisferios a la vez. Otra alternativa simple y efectiva a considerar consiste en la sustracción de las imágenes pre y postoperatorias alineadas. Por supuesto, esto no se puede hacer si los datos proporcionados solo contienen imágenes postoperatorias, lo que suele ser la situación más común en escenarios clínicos reales. La complejidad de este proceso radica en la variabilidad anatómica individual y la necesidad de lograr una restauración precisa de la geometría craneal original. Los avances en capacidad de cómputo y su creciente accesibilidad han facilitado la adopción de nuevas aproximaciones basadas en inteligencia artificial, que presentan un potencial significativo para alcanzar altos niveles de precisión en la reconstrucción mientras reducen los tiempos de

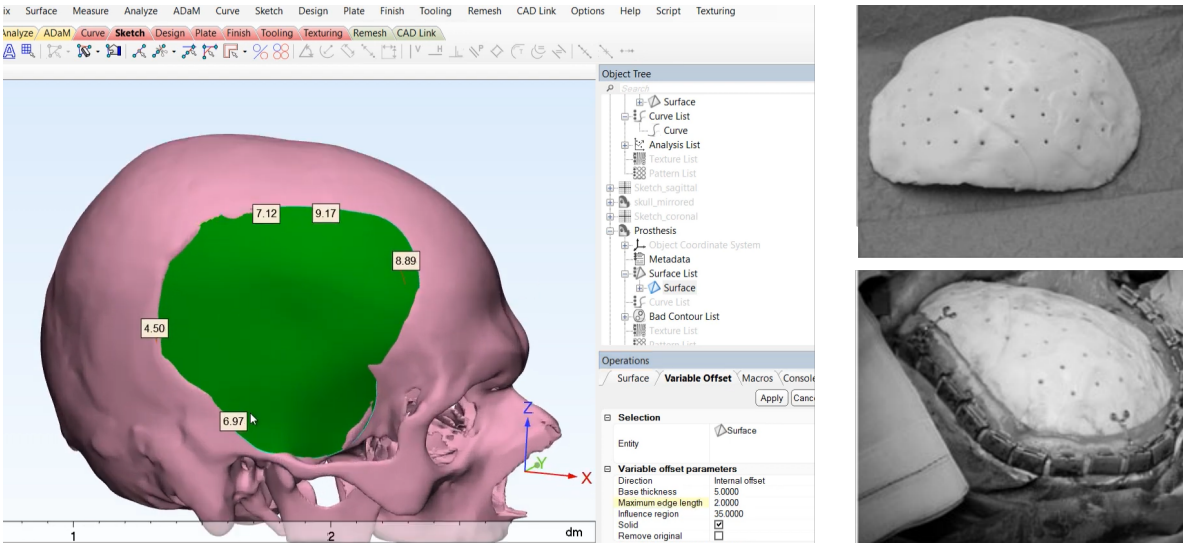


Figura 1.1: Diseño manual de los implantes craneales (izquierda). Implante específico para el paciente después de la impresión y después de ser colocado en el paciente (derecha). Imágenes tomadas de [49] y [46].

diseño [10].

En esta tesis se propone abordar esta problemática utilizando técnicas basadas en aprendizaje profundo. En particular, se explorará el uso de métodos autosupervisados [17], que no requieren anotaciones manuales de referencia para supervisar el proceso de aprendizaje, lo cual es especialmente relevante dado el alto costo y complejidad asociados a obtener dichas anotaciones en el contexto médico. Estos enfoques tienen el potencial de superar las limitaciones de las técnicas tradicionales, al permitir la reconstrucción automática de implantes craneales complejos y variados, sin depender de suposiciones de simetría o disponibilidad de imágenes preoperatorias.

Otro de los problemas en que se indagará en esta tesis doctoral, de fundamental importancia para el campo de las neuroimágenes, es el de segmentación de lesiones cerebrales. En este caso, el problema consiste en diferenciar los píxeles que corresponden a una lesión de aquellos que corresponden al tejido sano, con el objetivo de cuantificar su volumen. En particular, se abordará la segmentación de hiperintensidades de la sustancia blanca (HSB) [51]. Estas lesiones son particularmente visibles en imágenes de resonancia magnética adquiridas con la secuencia FLAIR (del inglés, Fluid-Attenuated Inversion Recovery). Esta técnica es crucial en neuroimágenes porque anula la señal del líquido cefalorraquídeo, lo que provoca que las HSB, que son brillantes (hiperintensas), resalten con un alto contraste respecto al tejido sano y otras estructuras cerebrales, facilitando así su detección y análisis. La evaluación precisa del volumen de HSB es de vital importancia para estudios sobre diversas patologías neurológicas (por ejemplo, la escleorosis múltiple y la demencia), para determinar la asociación entre dichas lesiones y los datos clínico-cognitivos, así como también sus causas y los efectos de nuevos tratamientos en ensayos clínicos [14]. Si bien se han propuesto diversos

métodos basados en redes neuronales profundas para abordar este problema [4], uno de los aspectos metodológicos que ha recibido poca atención hasta el momento es la calibración de los modelos de segmentación [34].

La calibración en el área de aprendizaje automático, entendida como la relación entre las probabilidades predichas y la frecuencia real de aciertos, ha recibido mucha atención en los últimos años debido a su estrecha relación con el concepto de incertidumbre asociada a las predicciones. La incertidumbre en las predicciones es crucial en el contexto médico, ya que permite evaluar la confiabilidad de las segmentaciones generadas automáticamente y brindar una mejor información para la toma de decisiones clínicas. Los métodos existentes a menudo se centran en maximizar la precisión de la segmentación, pero no proporcionan información sobre la certeza del modelo en cada predicción. Esto puede llevar a una falsa sensación de confianza en las salidas del modelo, especialmente en presencia de casos atípicos o ambiguos.

Cuando un modelo no se encuentra bien calibrado, sus predicciones probabilísticas suelen tener un exceso de confianza [15], en el sentido que asignan altas probabilidades a las predicciones positivas, mientras que las negativas tienen una probabilidad muy baja, perdiendo la capacidad de indicar con qué grado de certeza realizó cada predicción (ver Figura 1.2). Peor aún, en ocasiones los modelos con exceso de confianza suelen asignar probabilidades muy altas a predicciones completamente erroneas. La idea de contar con modelos de segmentación bien calibrados resulta de gran interés en el contexto médico, dado que permite estimar la confiabilidad de las predicciones asignadas a las clases de interés (por ejemplo, lesión o tejido sano).

En particular, el foco de esta parte de la tesis estará puesto en el estudio de la calibración de modelos de segmentación frente a cambios de dominio. En el caso de las imágenes médicas, uno de los escenarios más comunes de cambio de dominio ocurre al querer utilizar modelos que fueron entrenados con datos provenientes de un centro médico en otro distinto (lo que se conoce como datos multi-céntricos). En estos casos, los cambios en las condiciones de captura de las imágenes, en los dispositivos de adquisición, en los protocolos o incluso en la demografía de los pacientes, suelen producir que modelos previamente eficaces dejen de funcionar correctamente. Aún así, con una caída importante en el desempeño, estos modelos siguen brindando salidas con exceso de confianza, es decir, descalibradas. En este sentido, el interés de esta tesis radica en explorar cómo los estimadores de incertidumbre en redes neuronales pueden actuar como indicadores para detectar tempranamente potenciales fallos en los modelos de segmentación, especialmente en escenarios donde no se dispone de anotaciones manuales. Para ello, se analizarán estrategias para mejorar la calidad de la estimación de incerteza mediante métodos de regularización por entropía, y se realizará una validación experimental en imágenes de resonancia magnética multi-centricas.

Objetivos

Objetivo General

El objetivo general de esta tesis consiste en desarrollar técnicas innovadoras para la reconstrucción craneal mediante aprendizaje profundo auto-supervisado, y estudiar la estima-

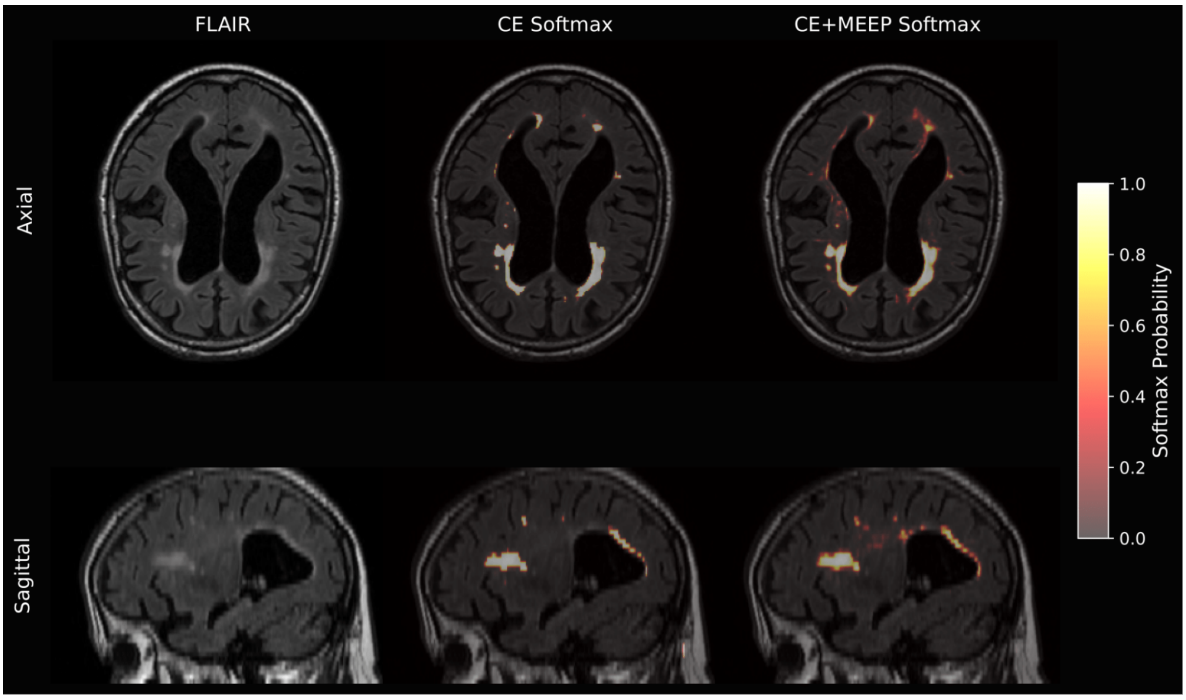


Figura 1.2: Comparación de la segmentación de hiperintensidades de la sustancia blanca (HSB) en una resonancia magnética FLAIR de un paciente con esclerosis múltiple. Se contrasta la imagen de entrada (izquierda) y la salida con exceso de confianza de un modelo CE Softmax (centro) con el resultado de un modelo CE+MEEP Softmax (derecha). Este último enfoque produce segmentaciones probabilísticas más detalladas que capturan mejor la incertidumbre mediante valores intermedios, sobre todo en los contornos de las lesiones y en las regiones pequeñas de HSB.

ción de incertidumbre en modelos de segmentación de imágenes médicas, con el fin de mejorar la precisión y eficiencia en la planificación de cirugías craneales y el diagnóstico de patologías neurológicas.

Objetivos Específicos

1. Desarrollar técnicas basadas en aprendizaje profundo auto-supervisado para la reconstrucción craneal en el espacio volumétrico, que no requieran anotaciones manuales.

2. Implementar metodologías de generación de datos sintéticos en el contexto de la reconstrucción craneal, que superen las limitaciones de las técnicas tradicionales basadas en simetría.

3. Estudiar y evaluar estrategias de regularización por entropía para mejorar la estimación de incertidumbre en modelos de segmentación de imágenes, analizando su comportamiento frente a cambios de dominio.

4. Evaluar la capacidad de los modelos de segmentación para generar estimaciones de incertidumbre que sirvan como indicadores confiables de errores de predicción, especialmente en escenarios de cambio de dominio donde la validación manual no está disponible.

5. Evaluar experimentalmente las metodologías propuestas en conjuntos de datos reales, comparando su rendimiento con métodos existentes y analizando su aplicabilidad práctica.

Organización de la tesis

La estructura de esta tesis está organizada de la siguiente manera:

- El **Capítulo 2** describe la metodología propuesta para la reconstrucción de implantes craneales mediante métodos autosupervisados. Primero se introducen los conjuntos de datos utilizados y las técnicas de preprocesamiento aplicadas. Luego se detalla la metodología de craniectomía virtual, que permite generar datos de entrenamiento sin necesidad de anotaciones manuales. A continuación, se presenta la arquitectura de la red neuronal convolucional 3D diseñada para la tarea de reconstrucción, seguida de una descripción de los procedimientos de entrenamiento y las estrategias de validación. Finalmente, se presentan los resultados experimentales obtenidos, incluyendo una evaluación cuantitativa y cualitativa de los métodos propuestos en ambos contextos.

- El **Capítulo 3** introduce las técnicas implementadas para la calibración de modelos de segmentación de lesiones de HSB, y, a su vez, propone cómo utilizar la incertidumbre resultante para la estimación no supervisada de errores, especialmente frente a cambios de dominio. Se describen los conjuntos de datos utilizados, las técnicas de preprocesamiento y la metodología de calibración propuesta. También se presentan los resultados experimentales obtenidos, incluyendo una evaluación cuantitativa y cualitativa de los métodos propuestos.

- Finalmente, en el **Capítulo 4** se presentan las conclusiones generales del trabajo realizado y se proponen líneas futuras de investigación.

Capítulo 2:
Reconstrucción craneal mediante técnicas de aprendizaje profundo

La reconstrucción craneal representa un desafío crítico en el ámbito de la neurocirugía, con implicaciones que trascienden la mera recuperación anatómica. En este capítulo abordamos dos aspectos fundamentales de este proceso: la estimación de indicadores clínicos relevantes para el seguimiento post-operatorio y el diseño automatizado de prototipos de implantes para craneoplastía, el procedimiento quirúrgico que repara el defecto óseo para restaurar la integridad estructural del cráneo y proteger el cerebro [2]. La estimación precisa del volumen del colgajo óseo extraído durante una craniectomía descompresiva constituye un indicador fundamental para evaluar el esfuerzo descompresivo y su potencial impacto en la recuperación del paciente. Los métodos tradicionales para esta estimación, basados en mediciones manuales o aproximaciones geométricas simples, pueden presentar limitaciones significativas en términos de precisión y reproducibilidad. Nuestra propuesta busca superar estas limitaciones mediante el desarrollo de técnicas automáticas que aprovechan el potencial del aprendizaje profundo.

Por otro lado, el diseño de implantes craneales específicos para cada paciente requiere una comprensión detallada de la geometría craneal original y sus particularidades anatómicas. Los métodos convencionales basados en software CAD, si bien efectivos, demandan un tiempo considerable y *expertise* específica. Nuestra aproximación propone automatizar parte de este proceso mediante redes neuronales convolucionales 3D, capaces de aprender patrones anatómicos complejos y generar reconstrucciones precisas.

Un aspecto innovador de nuestra metodología es la incorporación de técnicas de aprendizaje autosupervisado. Este paradigma permite entrenar modelos robustos sin necesidad de extensas bases de datos anotadas manualmente. La estrategia consiste en generar automáticamente las propias etiquetas de entrenamiento a partir de los datos de entrada mediante una tarea de pretexto. En nuestro caso, esto se logra mediante un procedimiento de "craniectomía virtual" que simula defectos craneales en cráneos completos, generando así los pares de entrenamiento (cráneo con su defecto) que guían el aprendizaje del modelo. Esta aproximación no solo elimina la dependencia de datos etiquetados manualmente, sino que también permite una mejor generalización a diferentes tipos y tamaños de defectos craneales.

La metodología propuesta integra varias innovaciones técnicas, incluyendo el uso de atlas anatómicos como información a priori, técnicas de registro de imágenes para alineación espacial y estrategias de aumento de datos específicamente diseñadas para la tarea de reconstrucción craneal. Estos desarrollos se evaluaron exhaustivamente tanto en escenarios simulados como en casos clínicos reales, demostrando su potencial para mejorar la precisión y eficiencia en la planificación quirúrgica.

2.1 Antecedentes

Debido a que la reimplantación del colgajo óseo extraído previamente puede conllevar complicaciones en el paciente (desde leves como retracciones en la piel o problemas en la cicatrización, hasta graves como infecciones óseas) [45], existen varias alternativas a la reinserción que buscan restablecer un estado anatómico normal y cumplir con la misma función protectora que ofrece el cráneo. Los implantes genéricos por lo general consisten en injertos óseos, cemento óseo o mallas metálicas maleables, que si bien son de bajo costo relativo y permiten recuperar la función protectora del cerebro, no garantizan el mejor resultado estético [8]. Una solución superadora consiste en diseñar implantes específicos por paciente que dependan de su anatomía individual. Tradicionalmente, el proceso de fabricación de estos implantes puede requerir un tiempo de espera considerable, ya que el diseño de los mismos debe realizarse manualmente por especialistas (como ya se mostró en la Figura 1.1), previo a la impresión 3D o fabricación por otros métodos.

A partir de estas necesidades, el proceso de diseño automático del implante para la reconstrucción craneal puede plantearse inicialmente como un problema de segmentación de imágenes. Por la naturaleza de las tomografías computarizadas (TC), el tejido óseo puede ser extraído fácilmente para crear una nueva imagen mediante técnicas simples de umbralamiento utilizando los valores más altos en la escala de Hounsfield [43]. De esta forma, la complejidad del problema disminuye (al tomar una imagen cuyos valores pasan a ser binarios: fondo y hueso) y pueden emplearse modelos de segmentación profundos estándar de tipo codificador-decodificador, como U-Net [42] o SegNet [3], buscando que, dada una imagen de entrada con CD, la salida del modelo sea una máscara tridimensional que represente el colgajo deseado o el cráneo reconstruido[1].

En paralelo a los enfoques de reconstrucción geométrica, la estimación del volumen del defecto craneal ha sido abordada mediante métodos manuales simplicados para su uso en la práctica clínica. Entre ellos, destaca el método ABC, propuesto por Xiao et al. [52], que se consolidó como una técnica de referencia. Este enfoque estima el volumen del colgajo óseo a partir de tres mediciones manuales simples (longitud, ancho y profundidad) realizadas directamente sobre las imágenes de TC, aplicando una fórmula geométrica para aproximar el volumen. Si bien su simplicidad lo hace útil para una estimación rápida, su precisión está limitada por la aproximación geométrica y la variabilidad del observador, lo que motiva el desarrollo de métodos automáticos más precisos y robustos.

Otra vertiente de investigación se basó en los Modelos Estadísticos de Forma para reconstruir el cráneo a partir de una forma media y sus principales modos de variación anatómica [11]. En estos modelos, una forma craneal específica (generalmente representada como una malla 3D) se describe como una deformación de una forma promedio, la cual es aprendida a partir de una población de entrenamiento. Matemáticamente, una nueva forma $S(w)$ se puede generar a partir de un vector de pesos w de la siguiente manera

[1] A diferencia de los métodos tradicionales de segmentación de imágenes, en este tipo de métodos profundos podemos aprovechar la capacidad de comprimir la información a un espacio latente (conservando la estructura anatómica principal) y, a la vez, recuperar los detalles de alta resolución gracias a las conexiones por saltos (skip-connections en inglés).

$$S(w) = \bar{S} + \sum_{i=1}^{N} w_i p_i, \tag{2.1}$$

donde $S(w)$ es la forma craneal reconstruida, generada por el vector de pesos $w = (w_1, ..., w_n)$, $\bar{S}$ representa la forma craneal promedio de la población de entrenamiento, como un único vector de coordenadas; p_i es el i-ésimo modo principal de variación, que captura un patrón de deformación anatómica y es obtenido mediante Análisis de Componentes Principales sobre formas anatómicas previamente alineadas[2]; y w_i es el peso escalar que determina la contribución del i-ésimo modo de variación a la forma final. La sumatoria se realiza sobre los N modos de variación más significativos considerados por el modelo. El objetivo durante la reconstrucción es encontrar el conjunto de pesos w^* que mejor ajuste la forma del modelo a la porción sana del cráneo del paciente. Aunque esta estrategia puede aplicarse para ajustar la anatomía de un paciente con defecto craneal, depende fuertemente del espacio de formas observado en la población de entrenamiento y puede resultar costosa en términos de cómputo, requiriendo varios minutos por caso [28].

Simultáneamente, se desarrollaron métodos que aprovechaban la naturaleza bilateral del cráneo, empleando el lado contralateral como guía anatómica para reconstruir el área faltante. Sin embargo, esta alternativa se ve limitada en presencia de defectos bilaterales o de asimetrías significativas. De igual forma, la sustracción entre TC pre y postoperatorias se ha utilizado como recurso para estimar la región faltante, aunque su aplicabilidad clínica se encuentra restringida por la falta de datos preoperatorios disponibles o debidamente alineados.

Como posible enfoque alternativo, la reconstrucción craneal podría inscribirse en el campo del completado de formas tridimensionales, un área muy estudiada en visión por computadora y gráficos computacionales [6]. Dichos métodos se apoyan en la restauración de geometrías incompletas a partir de nubes de puntos, mallas poligonales o vóxeles, y ofrecen conceptos potencialmente útiles para la reconstrucción de defectos craneales. No obstante, su adaptación al ámbito de las imágenes médicas con segmentaciones puramente volumétricas abre nuevas líneas de investigación que exceden el alcance principal de esta tesis.

Finalmente, la irrupción de las técnicas de aprendizaje profundo supuso un cambio radical en la reconstrucción craneal. Los primeros trabajos incluyeron la utilización de autocodificadores de limpieza de ruido para el completado de cráneos [36], aunque inicialmente estaban restringidos a imágenes de baja resolución. El incremento de la capacidad de cómputo (especialmente gracias a las GPU) permitió la implementación de arquitecturas neuronales más complejas, capaces de procesar imágenes de mayor resolución y de aprender patrones anatómicos detallados a partir de grandes volúmenes de datos.

Las metodologías basadas en aprendizaje profundo ofrecen la posibilidad de automatizar buena parte de la reconstrucción craneal, incorporar mecanismos de autosupervisión para disminuir la dependencia de datos manualmente etiquetados y capturar de manera más robusta la variabilidad anatómica. A continuación, se describirá la aproximación propuesta en esta te-

[2]En este contexto, un "modo" es un vector con la misma dimensión que la forma promedio $\bar{S}$. No representa una forma en sí, sino un patrón de deformación. Por ejemplo, un modo podría describir la transición de un cráneo alargado a uno más redondeado.

sis, detallando los procedimientos de preprocesamiento, las arquitecturas de redes neuronales empleadas y los criterios de evaluación que se aplican a lo largo del proceso.

2.2 Materiales y métodos

La presente sección describe el flujo de trabajo propuesto para abordar la reconstrucción craneal mediante técnicas de aprendizaje profundo. En primer lugar, se detallan las operaciones de preprocesamiento necesarias para obtener datos consistentes y homogéneos, tanto desde el punto de vista de la intensidad (escala de Hounsfield) como de la geometría (alineación y normalización espacial). A continuación, se presenta la estrategia de craniectomía virtual, la cual permite la generación de pares de entrenamiento de forma autosupervisada. Finalmente, se detallan las arquitecturas de redes neuronales implementadas para la tarea de reconstrucción.

2.2.1 Preprocesamiento

El preprocesamiento de las imágenes de TC constituye una etapa crítica para asegurar la calidad y consistencia de los datos de entrada. El objetivo principal es obtener una representación binaria normalizada del cráneo que permita el análisis posterior. Este proceso se estructura en las siguientes tres etapas fundamentales.

Registro de imágenes

La primera etapa consiste en el registro de todas las imágenes a un espacio común definido por un atlas craneal de referencia. Este paso es fundamental por varios motivos:

- Normaliza la posición y orientación de todas las imágenes, reduciendo la variabilidad no relacionada con la morfología craneal
- Facilita el aprendizaje al permitir que la red se concentre en las variaciones anatómicas relevantes
- Simplifica la comparación entre casos al tener todos los cráneos en un sistema de coordenadas común

El proceso de registro se realiza mediante SimpleElastix [32], utilizando transformaciones rígidas que preservan la anatomía original. Para cada imagen I, se obtiene una transformación $\mathscr{T}$ y su inversa $\mathscr{T}^{-1}$, que permiten mapear la imagen al espacio del atlas y viceversa según

$$I_{reg} = \mathscr{T} \circ I. \tag{2.2}$$

Remuestreo y normalización

Tras el registro, las imágenes se remuestrean a una resolución isotrópica común:

- Espaciado uniforme de 2 mm entre vóxeles en todas las direcciones

- Dimensiones estandarizadas para facilitar el procesamiento por *batches*.

Este paso es especialmente importante dado que las imágenes originales pueden tener diferentes resoluciones y espaciados entre cortes, particularmente en la dirección z. La resolución elegida permite preservar la información anatómica relevante para la tarea, al mismo tiempo que se reduce la carga computacional de las etapas posteriores.

Extracción de máscaras binarias

La última etapa consiste en la obtención de máscaras binarias del cráneo. Esta segmentación se realiza mediante umbralización global, aprovechando que el tejido óseo presenta valores característicamente altos en la escala de Hounsfield. El umbralado se realiza mediante

$$M(x) = \begin{cases} 1, & \text{si } I(x) \geq 90 \text{ HU}, \\ 0, & \text{en otro caso.} \end{cases} \tag{2.3}$$

En este caso, $I(x)$ representa el valor de intensidad en la posición x y el umbral de 90 unidades Hounsfield (HU) fue determinado empíricamente para optimizar la segmentación del tejido óseo [43, 12]. El resultado final de este proceso es una representación binaria normalizada del cráneo, donde:

- Valor 1 representa tejido óseo
- Valor 0 representa tejido blando o fondo
- Todas las imágenes están en el mismo espacio de coordenadas
- La resolución es uniforme e isotrópica

Este preprocesamiento es fundamental para el éxito de las etapas posteriores, ya que proporciona una base consistente y normalizada para el entrenamiento de los modelos de reconstrucción.

2.2.2 Metodología de craniectomía virtual

En esta tesis se propuso el concepto de Craniectomía Virtual (CV) para abordar uno de los principales desafíos en la reconstrucción craneal asistida por computadora: la escasez de datos etiquetados. Esta técnica nos permite generar de manera sintética pares de entrenamiento consistentes en cráneos con defectos simulados y sus correspondientes colgajos óseos faltantes. El proceso, ilustrado en la Figura 2.1, facilita el entrenamiento de modelos de aprendizaje profundo utilizando técnicas autosupervisadas, sin depender de anotaciones manuales costosas.

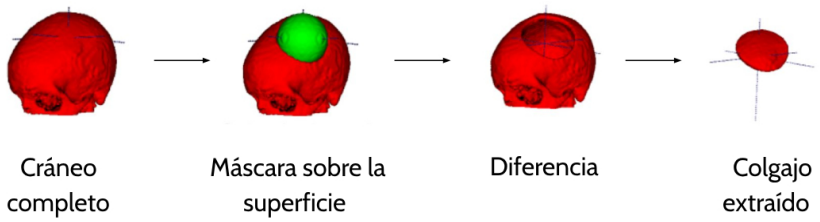


Figura 2.1: Ilustración del proceso de Craniectomía Virtual (CV). A partir de un cráneo completo (C_o), mostrado a la izquierda, se define una máscara de extracción esférica (M_e) (en verde). La intersección de ambos da como resultado un cráneo con un defecto simulado y su correspondiente colgajo óseo extraído (C_v), mostrado a la derecha, que se utiliza como referencia.

Aprendizaje autosupervisado y CV

En el método de aprendizaje autosupervisado que proponemos, la CV actúa generando automáticamente las señales de supervisión necesarias para el aprendizaje a partir de los propios datos no etiquetados.

La clave de este enfoque reside en que, al simular defectos craneales de manera controlada, podemos generar automáticamente pares de entrenamiento donde conocemos con exactitud tanto la forma y ubicación del defecto como la anatomía original. Esto permite explotar grandes volúmenes de TC craneales rutinarias donde cada cráneo completo actúa como su propia señal de supervisión.

Evolución de las máscaras de extracción en la CV

El proceso de CV ha evolucionado significativamente a lo largo de nuestra investigación, pasando de formas geométricas simples a patrones más sofisticados que buscan emular con mayor fidelidad los defectos quirúrgicos reales observados en la práctica clínica. Inicialmente, implementamos un sistema basado en máscaras esféricas simples. Esta aproximación se fundamentó en la observación empírica de que, para imágenes de baja resolución, las formas esféricas proporcionan una aproximación razonable de los defectos quirúrgicos. Como se ilustra en la Figura 2.1, el proceso consiste en una operación de intersección booleana. La extracción del colgajo virtual (C_v) se define como la intersección ($\cap$) entre el cráneo original intacto (C_o) y una máscara esférica de extracción (M_e):

$$C_v = C_o \cap M_e \tag{2.4}$$

Sin embargo, a medida que avanzamos hacia aplicaciones que requieren mayor precisión geométrica, como la reconstrucción detallada de implantes, exploramos un conjunto más amplio de formas geométricas para las máscaras de extracción. Esto incluyó no solo esferas, sino también cubos, prismas, cilindros y combinaciones de estas formas básicas. El objetivo era encontrar patrones de extracción que capturen mejor las características geométricas de los defectos craneales reales, como bordes irregulares y áreas de trepanación. Por ejemplo, una

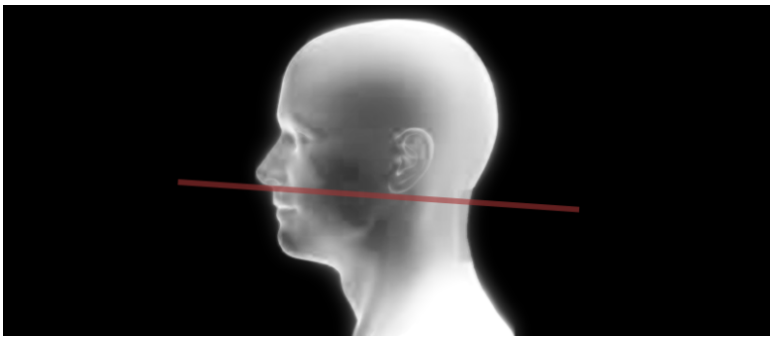


Figura 2.2: Plano de referencia anatómico definido para el control espacial y la parametrización de la cavidad virtual. Este plano se establece utilizando la base nasal en la parte anterior y la unión craneovertebral en la parte posterior, garantizando la plausibilidad anatómica de los defectos generados

máscara compuesta por la unión de un prisma rectangular y varios cilindros puede aproximar la forma de un defecto quirúrgico con múltiples orificios de trepanación mediante

$$M_{co} = (M_p \cup M_c) \cap M_s, \tag{2.5}$$

donde M_{co} es la máscara de extracción compuesta resultante, M_p es un prisma rectangular básico, M_c son uno o más cilindros posicionados para simular orificios de trepanación, y M_s es una máscara que restringe la extracción a la superficie exterior del cráneo.

La exploración de diferentes formas geométricas para las máscaras de extracción nos permitió generar una mayor variedad de defectos virtuales, abarcando desde formas simples hasta patrones más complejos e irregulares. Esta flexibilidad en la generación de defectos fue clave para entrenar modelos robustos capaces de manejar la variabilidad anatómica observada en escenarios clínicos reales. Es importante tener en cuenta que la elección de la forma geométrica óptima para las máscaras de extracción depende en gran medida de la aplicación específica y las características de los defectos craneales de interés. Nuestra metodología de CV proporciona un marco flexible para adaptar las máscaras de extracción según las necesidades particulares de cada caso, permitiendo una simulación más realista y adaptada a la tarea en cuestión.

Control espacial y parametrización de la CV

Para garantizar la plausibilidad anatómica de los defectos generados, desarrollamos un sistema de control espacial basado en puntos de referencia anatómicos clave, definiendo un plano de referencia utilizando la base nasal anteriormente y la unión craneovertebral posteriormente (Figura 2.2).

Las extracciones virtuales se restringen a la región craneal superior a este plano, asegurando que los defectos simulados correspondan a ubicaciones anatómicamente plausibles para craniectomías reales. Además, parametrizamos el tamaño de los defectos generados para replicar la variabilidad observada en casos clínicos reales asegurando $V_d \in [V_{min}, V_{max}]$, donde

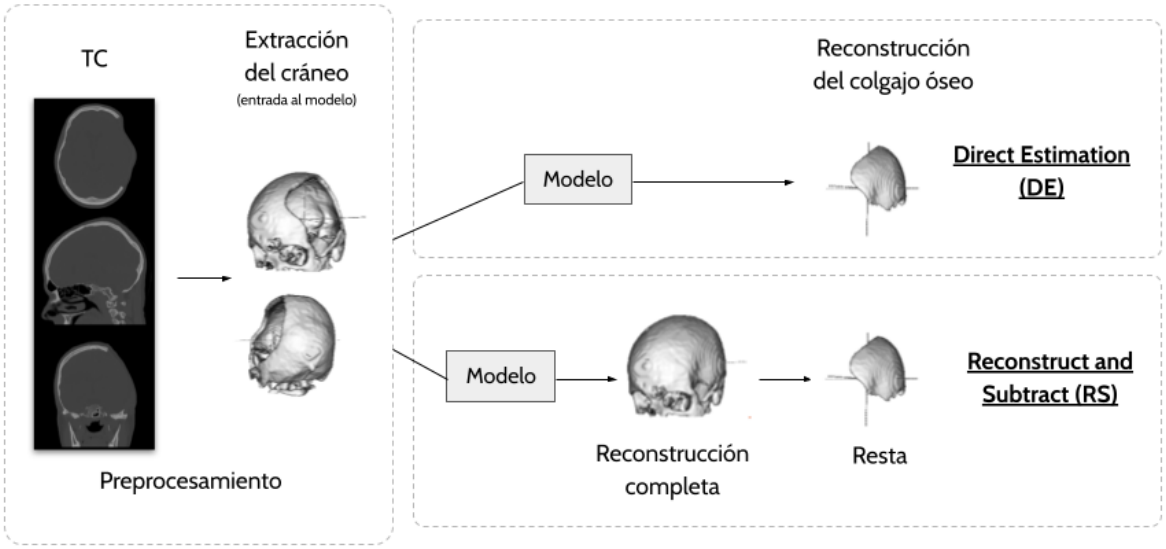


Figura 2.3: Esquema general de la metodología propuesta. A partir de una TC de entrada, una etapa de preprocesamiento extrae el cráneo binario. Posteriormente, se comparan dos estrategias para la reconstrucción del colgajo óseo: la estimación directa (DE), donde un modelo predice directamente el implante faltante; y la reconstrucción y resta (RS), donde un modelo primero reconstruye el cráneo completo para luego obtener el implante mediante una sustracción con el cráneo de entrada.

V_d es el volumen del defecto generado, y $[V_{min}, V_{max}]$ es el rango de volúmenes permitidos. Establecimos empíricamente este rango entre 0.7 cm³ y 350 cm³ basándonos en un análisis estadístico detallado de casos reales del conjunto de datos CENTER-TBI (ver Sección 2.3.2).

Implementación y optimización de la CV

El proceso de CV fue integrado en el flujo de entrenamiento de nuestros modelos. Para ello, generamos los defectos virtuales de manera dinámica durante el entrenamiento con una probabilidad p_{cv} de 0.8. Esto permite que el modelo observe tanto cráneos intactos como cráneos con defectos durante el aprendizaje, mejorando la robustez y previniendo el sobreajuste a patrones específicos de defectos. Complementamos la CV con técnicas estándar de aumentación de datos, aplicando transformaciones geométricas aleatorias y añadiendo ruido de tipo "sal y pimienta",

$$X_a = \mathscr{A}(X_{cv}) + \eta \tag{2.6}$$

donde X_{cv} es un cráneo con un defecto generado por CV, $\mathscr{A}$ representa transformaciones como rotaciones, traslaciones y escalamientos aleatorios, y η es el ruido añadido. Estas técnicas de aumentación ayudan a mejorar la robustez y la capacidad de generalización de los modelos entrenados.

Las arquitecturas desarrolladas en este trabajo abordan la reconstrucción del colgajo óseo siguiendo dos estrategias principales, las cuales se resumen visualmente en la Figura 2.3.

La primera es la Estimación Directa (DE), que predice el implante directamente a partir del cráneo con defecto. La segunda es la de Reconstrucción y Resta (RS), que primero reconstruye el cráneo completo para luego obtener el implante por sustracción. A continuación, se detallan los modelos implementados para cada estrategia, los cuales se agrupan según dos objetivos principales: la estimación de volumen y la reconstrucción precisa de implantes.

Modelos para estimación de volumen

Para la estimación del volumen del colgajo óseo se evaluaron diversos modelos que implementan las estrategias de estimación directa (DE) y reconstrucción y resta (RS). Las arquitecturas de aprendizaje profundo se basan en tres variantes: dos que siguen el enfoque RS (RS-AE y RS-UNET) y una que implementa la estrategia DE (DE-UNET). Adicionalmente, se incluyó como método de referencia un modelo basado en análisis de componentes principales que también utiliza la estrategia RS (RS-PCA), para así proporcionar un punto de comparación.

Estimación directa con U-Net Esta primera arquitectura busca estimar directamente el colgajo óseo, eliminando los pasos intermedios de reconstrucción y sustracción que podrían introducir errores. Basada en la arquitectura U-Net 3D con conexiones de salto, este modelo aprende a reconstruir directamente el colgajo óseo removido durante la craniectomía virtual. Requiere únicamente cráneos completos, lo que permite preservar el principio de aprendizaje autosupervisado y elimina la dependencia de anotaciones manuales.

Reconstrucción y resta con autocodificador La segunda arquitectura consiste en un autocodificador (AE, por sus siglas en inglés) completamente convolucional entrenado para reconstruir cráneos completos. Siguiendo la metodología propuesta por Larrazabal et al. [23], se implementa un AE de eliminación de ruido (denoising autoencoder en inglés) que integra craniectomías virtuales en el proceso de entrenamiento[3]. El modelo opera únicamente con cráneos completos, aplicando CV aleatoria previo al ingreso de la imagen. Durante la fase de prueba, el colgajo óseo se estima mediante la sustracción entre el cráneo original dañado y su versión reconstruida.

Reconstrucción y resta con U-Net Este tercer enfoque explora la utilización de U-Net siguiendo la estrategia de reconstrucción y resta, con el objetivo de evaluar el impacto de las conexiones de salto en el proceso de reconstrucción. La Figura 2.4 presenta un ejemplo comparativo de reconstrucción de colgajos óseos obtenidos mediante estos diferentes métodos, aplicados a un caso real de craniectomía descompresiva de nuestro conjunto de datos de prueba.

2.2.3 Incorporación de restricciones anatómicas explícitas

Extendiendo las arquitecturas base presentadas en la Sección 2.2.2, se desarrollaron variantes con restricciones anatómicas (a las que también denominas *shape priors* en inglés) para abordar la reconstrucción precisa de implantes denominadas RS-AE-Shape, DE-UNET-Shape y RS-UNET-Shape. La motivación para crear estas variantes surge del desafío de procesar imágenes de mayor resolución, lo cual presenta un gran desafío: al aumentar la resolución de la imagen de entrada sin modificar la arquitectura, el campo receptivo del modelo cubre un área anatómica proporcionalmente más pequeña, y en defectos craneales extensos, esto puede resultar en predicciones inconsistentes o con agujeros.

Para abordar esta limitación, la solución adoptada en estas variantes consiste en incorporar una guía anatómica explícita, proporcionando un atlas craneal (previamente alineado) junto con la imagen de entrada, en un canal adicional (siguiendo el enfoque propuesto en [25]), usando la misma arquitectura de base. Esta guía anatómica global complementa la información local del campo receptivo, actuando como una guía anatómica que le proporciona al modelo un contexto estructural, especialmente en regiones donde la señal de la imagen es ambigua o inexistente. Esto permite que los modelos procesen imágenes de mayor resolución y generen reconstrucciones coherentes y anatómicamente plausibles, tal como se ilustra en la Figura 2.5.

2.3 Experimentos

2.3.1 Configuración experimental

Para la tarea de estimación del volumen del colgajo óseo, se implementaron los modelos base (RS-AE, DE-UNET, RS-UNET) utilizando imágenes submuestreadas a una resolución de $64 \times 64 \times 64$ vóxeles. Los modelos se entrenaron durante 100 épocas utilizando el optimizador Adam, con una tasa de aprendizaje de 1×10^{-4} y un *batch size* de 8. Para la tarea de reconstrucción precisa de implantes, se evaluaron las variantes con restricciones de forma (RS-AE-Shape, DE-UNET-Shape, RS-UNET-Shape), sobre el conjunto de datos AutoImplant [27]. Estas variantes trabajaron con resolución nativa de $128 \times 128 \times 128$ vóxeles para preservar detalles geométricos finos. El entrenamiento se realizó durante 200 épocas, manteniendo los hiperparámetros del optimizador pero reduciendo el *batch size* a 4 debido a restricciones de memoria.

La evaluación del rendimiento se basó en métricas geométricas estándar para segmentación médica: el coeficiente Dice, la distancia de Hausdorff (en mm) y el error medio de superficie [48].

El coeficiente Dice cuantifica la superposición entre la segmentación predicha y la de referencia, siendo ampliamente utilizado en tareas de segmentación médica debido a su robustez ante desbalances de clases. Se calcula como:

$$\text{Dice} = \frac{2|G \cap P|}{|G| + |P|}, \tag{2.7}$$

[3]La arquitectura resultante es idéntica al RS-UNet, aunque quitando las conexiones de salto.

donde $G \cap P$ representa el número de elementos comunes entre el conjunto de la segmentación de referencia G y el conjunto predicho P, y $|\cdot|$ denota el número de elementos en el conjunto.

Para evaluar la precisión en la localización de los bordes, se empleó la distancia de Hausdorff, que mide la máxima distancia entre cualquier punto de una superficie y el punto más cercano en la otra superficie. Para dos conjuntos de puntos A y B, se define como:

$$H(A,B) = \max\{h(A,B), h(B,A)\}, \tag{2.8}$$

donde $h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$ es la distancia de Hausdorff direccional.

2.3.2 Conjuntos de datos

La evaluación y validación de los métodos propuestos se llevó a cabo empleando dos conjuntos de datos complementarios, cada uno con características específicas que permitieron analizar distintos aspectos de la metodología. En particular, se utilizó el conjunto de datos CENTER-TBI [31], que incluye casos clínicos reales, y el conjunto de datos AutoImplant Challenge [27], conformado por cráneos simulados diseñados para la evaluación de métodos de reconstrucción craneal. A continuación, se procederá a describir ambos conjuntos de datos.

Conjunto de datos CENTER-TBI

El primer conjunto de datos fue proporcionado por la División de Anestesia del Departamento de Medicina de la Universidad de Cambridge e incluye TC de pacientes con TCE. Su relevancia radica en que está compuesto por casos clínicos reales que abarcan distintos tipos y grados de lesiones, lo cual aporta una perspectiva más cercana a la práctica clínica. En total, se dispone de 98 estudios de TC cerebral correspondientes a 27 pacientes. Dichas imágenes se encuentran distribuidas en 31 casos postoperatorios con CD y 67 casos con el cráneo completo (ya sean preoperatorios o de seguimiento). Para asegurar una evaluación rigurosa y minimizar sesgos, se definió una estrategia de división de datos con los siguientes criterios. En primer lugar, se conformó un **conjunto de entrenamiento** integrado por 52 imágenes (pertenecientes a 17 pacientes distintos) que incluyeran únicamente cráneos completos. Con el fin de evitar contaminación cruzada, se excluyeron aquellos pacientes que tuvieran una imagen asociada con CD. En segundo lugar, se estableció un **conjunto de prueba** conformado por 10 pacientes para los cuales se disponía de estudios pre y postoperatorios. Esta disponibilidad simultánea de ambas imágenes hizo posible emplear la diferencia entre ellas como referencia para la evaluación cuantitativa de la reconstrucción del colgajo óseo. Finalmente, las 36 imágenes restantes (incluyendo casos preoperatorios del conjunto de prueba o postoperatorios sin su correspondiente imagen preoperatoria) no fueron consideradas para la evaluación formal, aunque se aprovecharon para realizar análisis cualitativos adicionales.

Conjunto de datos AutoImplant Challenge

El segundo conjunto de datos corresponde al AutoImplant Challenge, una iniciativa presentada en la conferencia MICCAI 2020 [27, 28], orientada específicamente a la reconstrucción craneal automatizada. Este recurso ofrece un conjunto de datos para evaluar de manera

controlada el desempeño de distintos métodos de reconstrucción. Para los experimentos presentados en esta tesis, se utilizaron 100 imágenes para entrenamiento y 110 para prueba.

Los cráneos poseen resolución estandarizada de 512×512 píxeles en el plano axial, y varían en la cantidad de cortes axiales (dimensión Z). Cada caso incluye:

1. El cráneo completo original.

2. Una versión del cráneo con un defecto quirúrgico simulado (cuya ubicación y tamaño varían).

3. Una máscara del defecto que funciona como referencia para el implante.

Una característica destacada de este conjunto de datos es el realismo de sus defectos simulados. Para emular cirugías complejas, muchos casos incluyen orificios de trepanación en los bordes del defecto (ver Figura 2.6). Estos patrones fueron replicados utilizando la metodología de máscaras compuestas por prismas y cilindros descrita en la sección anterior, lo cual enriquece la complejidad de la tarea y la asemeja a escenarios clínicos reales.

Es importante destacar la composición del conjunto de prueba, que fue dividido por los organizadores en dos subconjuntos para evaluar la robustez de los modelos. El primer subconjunto, que aquí denominamos $\mathscr{D}_{\text{test}}$, contiene 100 casos cuyos defectos siguen la misma distribución que los del conjunto de entrenamiento. El segundo, denominado $\mathscr{D}_{\text{test}-\text{extra}}$, está compuesto por 10 casos con defectos atípicos o fuera de distribución, diseñados para ser más desafiantes.

La disponibilidad de estos dos conjuntos de datos (uno con casos clínicos reales y otro con cráneos simulados de alta variabilidad) garantiza una evaluación exhaustiva de las metodologías propuestas. Mientras el conjunto de datos CENTER-TBI permite comprobar el desempeño de los algoritmos en un entorno clínico real, el AutoImplant Challenge facilita experimentos controlados y comparaciones cuantitativas directas.

2.4 Resultados

A continuación, se presentan los resultados experimentales obtenidos al aplicar la metodología propuesta en los conjuntos de datos descritos, tanto reales como simulados. El objetivo es validar la capacidad del modelo para estimar con precisión el volumen del colgajo óseo y para generar reconstrucciones craneales que se ajusten correctamente a la anatomía original. Se incluyen comparaciones con métodos tradicionales, enfoques propuestos en la literatura y soluciones de otros participantes en el desafío AutoImplant, lo cual permite evaluar de forma integral el desempeño y la robustez de la solución planteada bajo diversos escenarios y métricas de evaluación.

2.4.1 Evaluación volumétrica en el conjunto de datos CENTER-TBI

En este primer experimento se buscó evaluar la calidad de la estimación del volumen del colgajo óseo. Los resultados muestran que el modelo de estimación directa (DE-UNet)

Tabla 2.1: Resultados cuantitativos obtenidos para los dos métodos propuestos (DE-UNet y DE-Shape-UNet) comparados con los dos métodos de referencia reportados por los organizadores del desafío en [27]. Reportamos los valores medios de Dice y HD, y la desviación estándar entre paréntesis.

Método	$\mathscr{D}_{\text{test}}$ (100)		$\mathscr{D}_{\text{test-extra}}$ (10)		General	
	Dice	HD (mm)	Dice	HD (mm)	Dice	HD (mm)
Referencia N1 [27]	0.809	5.440	-	-	-	-
Referencia N2 [27]	0.855	5.182	-	-	-	-
DE-UNet	**0.913 (0.038)**	**4.067 (1.762)**	0.769 (0.126)	8.585 (5.128)	**0.900 (0.067)**	**4.477 (2.626)**
DE-Shape-UNet	0.845 (0.107)	6.414 (9.060)	**0.816 (0.078)**	**5.952 (1.258)**	0.842 (0.105)	6.372 (8.648)

supera consistentemente a los otros métodos en términos de coeficiente Dice y distancia Hausdorff. Como se muestra en la Figura 2.7, el modelo obtiene mejores resultados tanto en escenarios reales como simulados. Los gráficos de dispersión de la Figura 2.8 comparan el volumen predicho con el real para todos los métodos, incluyendo el método manual ABC. La proximidad de los puntos a la línea de identidad indica que DE-UNet logra las estimaciones más precisas del volumen del colgajo óseo.

2.4.2 Evaluación en el conjunto de datos AutoImplant

Posteriormente, se procedió a evaluar los métodos propuestos en el conjunto de datos AutoImplant. Para el subconjunto $\mathscr{D}_{\text{test}}$ (100 casos en distribución), el mejor de los métodos propuestos fue DE-UNet, que logró un Dice de 0.913 y una distancia Hausdorff de 4.067 mm, superando a los métodos de referencia ($N1$ y $N2$) del desafío [27], los cuales consisten en dos enfoques de aprendizaje profundo: un modelo en cascada que refina la predicción por parches y otro que reconstruye directamente el cráneo completo.

Para el subconjunto $\mathscr{D}_{\text{test-extra}}$ (que contiene 10 casos fuera de distribución), el mejor rendimiento entre nuestros métodos lo obtuvo DE-UNet-Shape, que alcanzó un Dice de 0.816 y una distancia Hausdorff de 5.952 mm. Cabe destacar que para este conjunto no se disponía de resultados de los métodos $N1$ y $N2$ proporcionados por los organizadores. Finalmente, la evaluación global sobre los 110 casos de prueba, DE-UNet nuevamente mostró el mejor desempeño entre los modelos propuestos con un Dice de 0.900 y una distancia Hausdorff de 4.477 mm (ver Tabla 2.1).

2.4.3 Comparativa con métodos de otros participantes del AutoImplant Challenge

Para contextualizar el rendimiento de nuestros enfoques DE-UNet y DE-Shape-UNet, se los comparó con los diversos métodos presentados en la competencia. Las arquitecturas más relevantes incluyeron:

- SSM-GAN: Una aproximación híbrida que combina Modelos Estadísticos de Forma (SSM) con una Red Generativa Antagónica (GAN) para el refinamiento [41].
- SE-CNN: Arquitecturas codificador-decodificador que utilizan bloques de *Squeeze-and-Excitation* (SE) para mejorar la captura de características [44].

Tabla 2.2: Resultados cuantitativos (media de Dice y HD) de los algoritmos participantes en los conjuntos $D_{test}100$ y $D_{test}10$. Nuestros métodos son DE-UNet y su variante con prior de forma, DE-UNet-Shape.

Métrica\Algoritmo	SSM-GAN	SE-CNN	**DE-UNet**	**DE-UNet-Shape**	Res-UNet	Cascaded-CNN	Reg-UNet	SC-UNet	VAE-Net	VAE-Net(re)	Patch-UNet(r)	Patch-UNet(p)	Baseline(r)	Baseline(bbox)
Dice (100)	0.917	0.931	0.913	0.845	0.944	0.920	0.907	0.896	0.887	0.891	0.735	0.889	0.810	0.856
Dice (10)	0.919	0.924	0.769	0.816	0.932	0.910	0.870	-	0.351	0.473	-	-	-	-
HD (100)	4.336	3.660	4.067	6.414	3.564	4.137	4.180	4.602	7.017	6.909	7.243	5.534	5.440	5.183
HD (10)	3.987	4.090	8.585	5.952	3.934	4.707	4.760	-	29.476	21.049	-	-	-	-

- Res-UNet: Una variante de U-Net que incorpora bloques residuales para aumentar la precisión [13].
- Cascaded-CNN: Un modelo en cascada que refina la predicción en dos etapas, procesando la imagen primero a baja y luego a alta resolución [19].
- VAE-Net: Una red que integra un Auto-Codificador Variacional (VAE) para regularizar la forma del implante y asegurar su plausibilidad anatómica [50].

La Tabla 2.2 muestra una comparación detallada con todos los participantes, evidenciando el rendimiento diferenciado de nuestras dos propuestas. Nuestro modelo DE-UNet presenta resultados altamente competitivos en el conjunto $D_{test}100$, mientras que la variante DE-UNet-Shape demuestra una mayor robustez en los casos más complejos de $D_{test}10$. En el Anexo D se puede profundizar sobre los demás métodos participantes y la comparativa.

Los resultados aquí presentados evidencian el potencial de las arquitecturas basadas en aprendizaje profundo para la reconstrucción craneal, incluso bajo escenarios clínicamente exigentes. Si bien se observan algunas limitaciones ante casos complejos, el enfoque propuesto logra una alta precisión en la mayoría de los escenarios evaluados, sentando las bases para futuras mejoras e integraciones. Las conclusiones completas de esta línea de trabajo se discutirán en el Capítulo 4, donde se profundizará en las implicaciones clínicas y técnicas de la metodología.

En el siguiente capítulo, se abordará otro problema de gran relevancia en el ámbito de las neuroimágenes: la estimación de la incertidumbre en la segmentación de hiperintensidades de sustencia blanca, cuyo desarrollo representa un paso adicional hacia la implementación de sistemas de inteligencia artificial más confiables en entornos médicos.

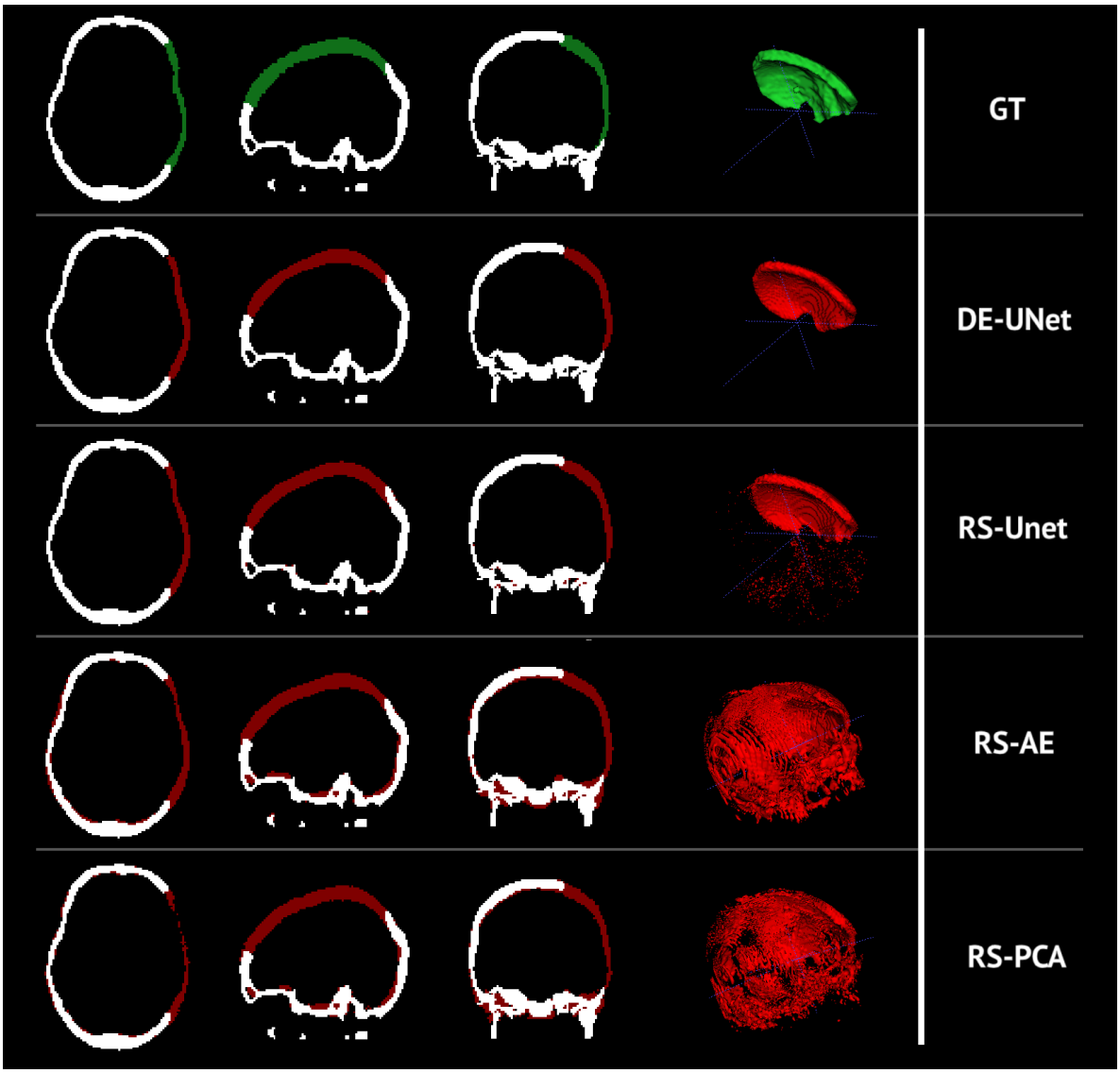


Figura 2.4: Reconstrucción del colgajo óseo (en rojo o verde) obtenida con los diferentes enfoques comparados en este trabajo para un caso real de craniectomía descompresiva de nuestro conjunto de datos de prueba. La reconstrucción de referencia se muestra en verde.

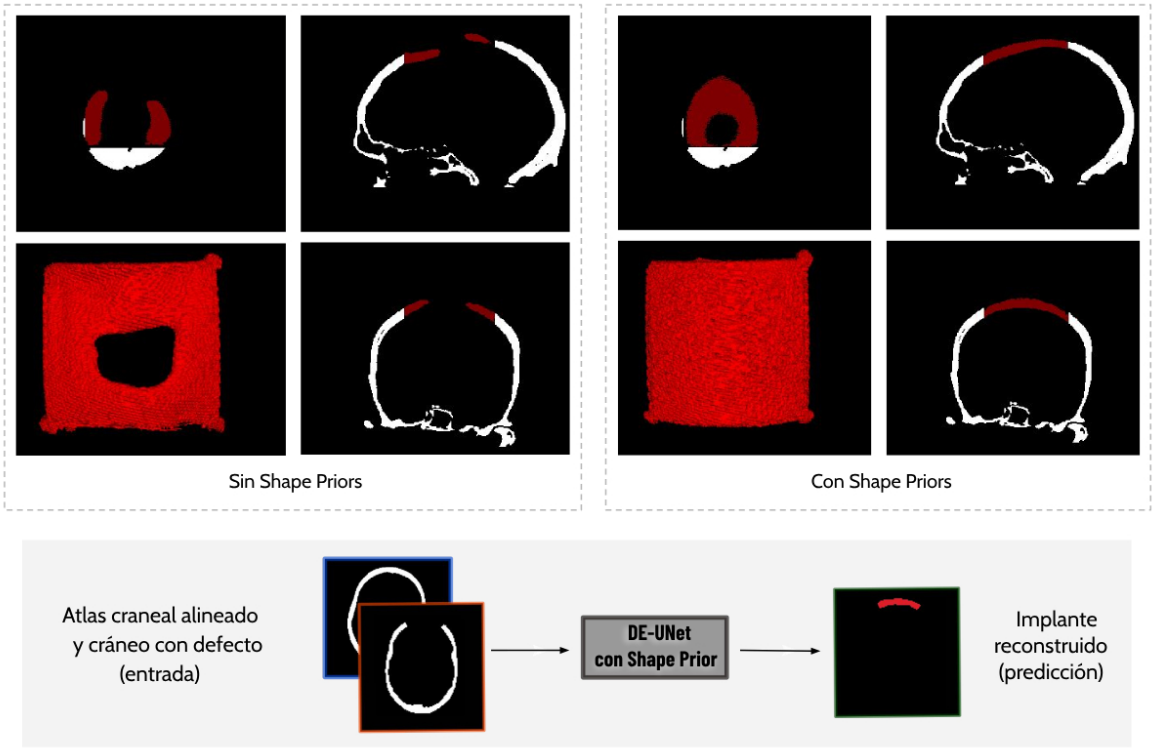


Figura 2.5: Comparación cualitativa y esquema metodológico del enfoque con restricciones anatómicas. **Arriba:** Ejemplos de reconstrucción de implantes para casos con defectos extensos. El modelo sin restricciones anatómicas (izquierda) genera predicciones con agujeros e inconsistencias debido al campo receptivo limitado. El modelo que incorpora el atlas como restricción anatómica (derecha) produce reconstrucciones completas y anatómicamente coherentes. **Abajo:** Diagrama simplificado de la arquitectura, que utiliza como entrada un cráneo con defecto y el atlas craneal alineado para predecir el implante.

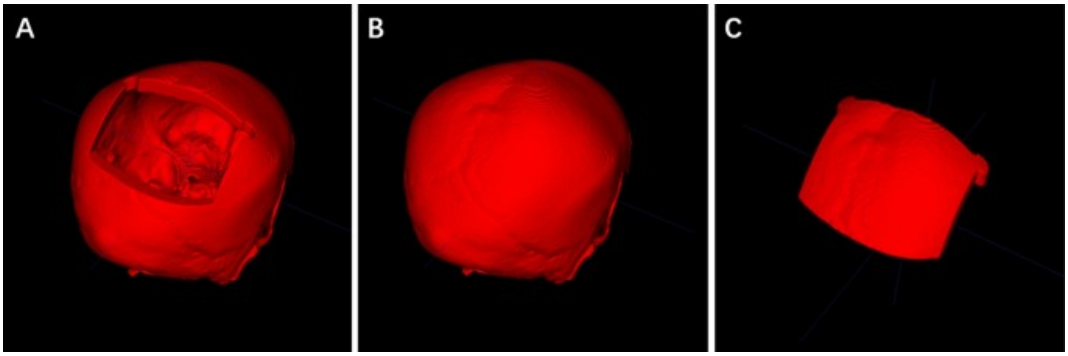


Figura 2.6: Ejemplo de un caso del conjunto de datos AutoImplant Challenge. Se muestra (A) el cráneo con el defecto simulado, (B) el cráneo completo original del cual se generó, y (C) el implante correspondiente que funciona como referencia. Este tipo de datos permite el entrenamiento y la evaluación supervisada de los métodos de reconstrucción. Imagen tomada de [27].

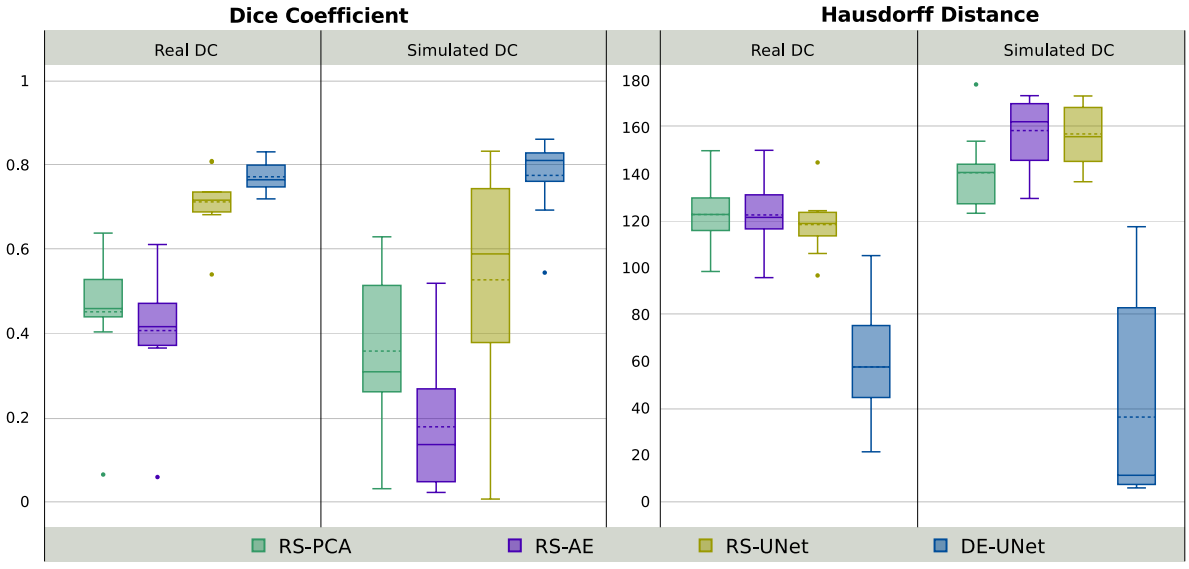


Figura 2.7: Coeficiente Dice y Distancia Hausdorff (en mm) de los métodos propuestos comparados con la referencia. La línea punteada indica el valor medio. Se puede observar que el modelo DE-UNet supera a los otros métodos evaluados.

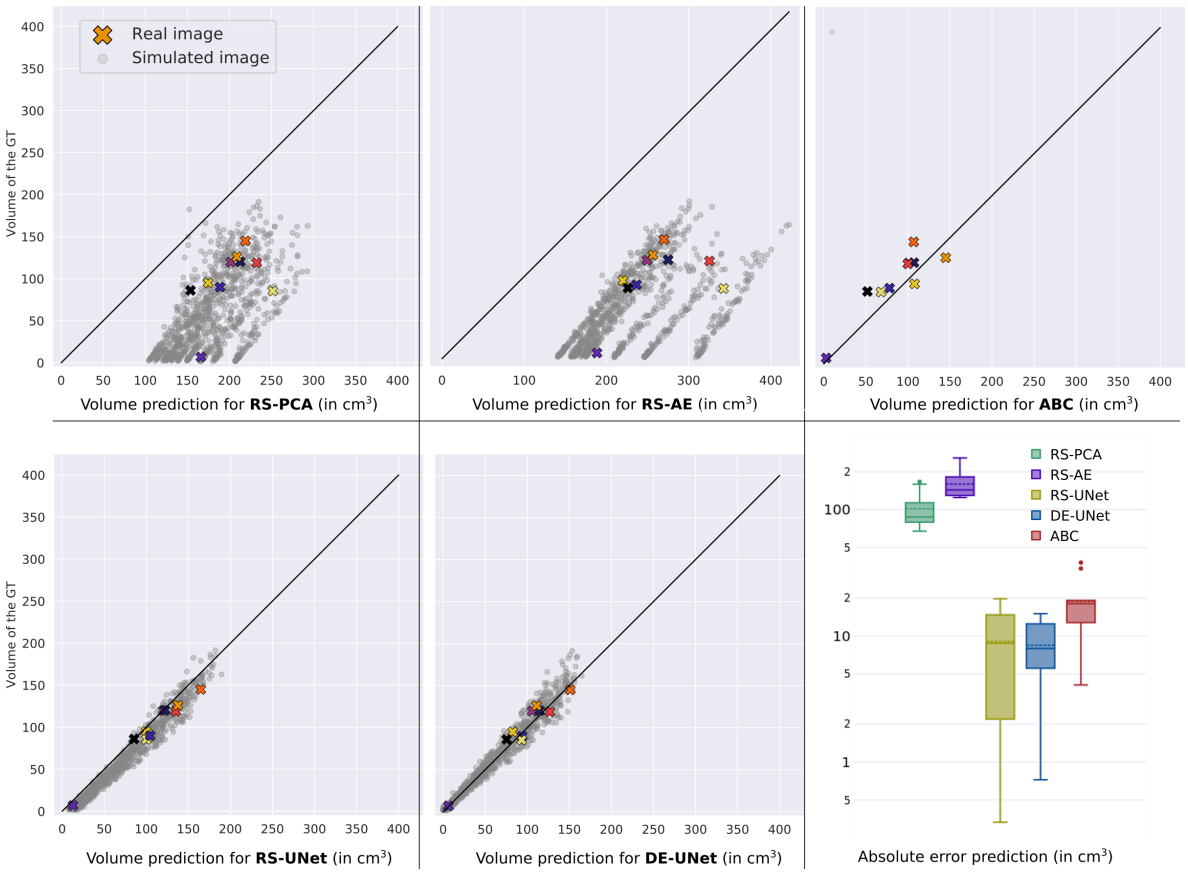


Figura 2.8: Comparación cuantitativa para la estimación del volumen del colgajo óseo con los diferentes métodos implementados. Los gráficos de dispersión muestran el volumen estimado (eje x) frente al volumen de referencia del colgajo óseo. Nótese que RS-PCA, RS-AE, RS-UNet y DE-UNet muestran resultados tanto para casos reales (marcadores en cruz de color) como para casos simulados (círculos en gris). Para ABC, solo mostramos resultados en casos reales, ya que se requiere la imagen TC real para la anotación manual.

Capítulo 3:
Estimación de incerteza en la segmentación de hiperintensidades de la sustancia blanca frente a cambio de dominio

La segmentación precisa de hiperintensidades de la sustancia blanca (HSB) en imágenes de resonancia magnética (RM) es de gran importancia tanto para la investigación como para la práctica clínica. La cuantificación del volumen de HSB resulta esencial en estudios que buscan comprender la progresión de enfermedades neurodegenerativas y su relación con factores clínicos y cognitivos [14].

Sin embargo, aunque los modelos de segmentación automática han alcanzado altos niveles de precisión, suelen generar predicciones con niveles de confianza elevados incluso en regiones que son intrínsecamente ambiguas. En el contexto médico, la estimación de la incertidumbre de un modelo es crucial, ya que permite detectar posibles errores en las segmentaciones y brinda herramientas para evaluar la fiabilidad de las predicciones sin necesidad de contar con anotaciones manuales en cada caso. Un modelo que pueda estimar adecuadamente su incertidumbre puede servir como una herramienta de apoyo en la toma de decisiones clínicas, señalando regiones donde la predicción es menos confiable y podría requerir una revisión experta.

En este capítulo se estudia el impacto del cambio de dominio en la estimación de incertidumbre en modelos de segmentación de HSB. El cambio de dominio es un problema recurrente en imágenes médicas, ya que las diferencias en los equipos de RM, protocolos de adquisición y características poblacionales pueden afectar la distribución de los datos, degradando el rendimiento del modelo. En este contexto, se busca determinar si la incertidumbre de las predicciones puede servir como un indicador de potenciales errores en la segmentación cuando el modelo es aplicado en datos provenientes de distintos centros de adquisición.

Para mejorar la estimación de incertidumbre en este contexto, se exploran técnicas de regularización por entropía, que buscan evitar predicciones excesivamente confiadas, fomentando distribuciones de probabilidad más representativas de la incertidumbre real del modelo. Como se ilustra en la Figura 1.2, los modelos tradicionales entrenados con entropía cruzada estándar tienden a producir predicciones con exceso de confianza, mientras que las técnicas de regularización por entropía pueden generar segmentaciones que capturan mejor la incertidumbre inherente en regiones ambiguas. A través de este enfoque, se investiga si modelos entrenados con este tipo de regularización pueden generar mapas de incertidumbre más informativos, reflejando con mayor precisión las regiones donde el modelo tiene menor certeza sobre su predicción. Como consecuencia de esta mejora en la estimación de incertidumbre, se analiza el efecto sobre la calibración del modelo, entendida como la relación entre las probabilidades predichas y la frecuencia real de aciertos. El objetivo central de este capítulo es evaluar si la incertidumbre basada en entropía puede utilizarse como una medida confia-

ble para anticipar errores en la segmentación de HSB, tanto en datos del mismo dominio de entrenamiento como en datos de distribución diferente. Para ello, se comparan diversas estrategias de regularización y se mide su impacto en métricas de segmentación, incertidumbre y calibración en escenarios de cambio de dominio.

3.1 Antecedentes

La segmentación de HSB en imágenes de RM ha sido objeto de intenso estudio durante las últimas décadas, dado su valor para la investigación clínica y el seguimiento de patologías neurológicas [14]. Inicialmente, gran parte de los métodos propuestos se fundamentaron en enfoques clásicos de procesamiento de imágenes y modelos estadísticos que buscaban detectar regiones hiperintensas en modalidades particulares de estas imágenes (como secuencias T2/FLAIR), empleando umbrales fijos o técnicas de agrupamiento. Con la popularización de las técnicas de aprendizaje automático, surgieron métodos basados en modelos probabilísticos (por ejemplo regresión logística o bosques aleatorios), que aprovechaban características de intensidad y textura para mejorar la discriminación entre lesión y tejido sano.

El advenimiento de las redes neuronales profundas marcó un punto de inflexión en la precisión y robustez de la segmentación de HSB. Arquitecturas de tipo U-Net, V-Net o variantes codificador-decodificador [4] se consolidaron como el estándar de facto para la tarea, mostrando mejoras considerables en métricas de superposición (como el Dice) con respecto a métodos tradicionales. No obstante, la mayoría de estos trabajos se enfocaron en maximizar la exactitud de la segmentación sin prestar especial atención a la incertidumbre de las predicciones. En el ámbito clínico, la falta de un indicador claro sobre la *confiabilidad* de la salida de la red puede derivar en una excesiva confianza en regiones dudosas, impactando potencialmente en la toma de decisiones médicas.

En los últimos años, se ha subrayado la importancia de calibrar los modelos de segmentación, entendiendo la calibración como la correspondencia entre las probabilidades predichas y la frecuencia real de aciertos [15, 34]. Un modelo bien calibrado otorga probabilidades altas cuando está seguro de su predicción y reduce el exceso de confianza en casos ambiguos o con características atípicas. En el ámbito de las HSB, este aspecto cobra especial relevancia cuando se enfrentan cambios de dominio, como la aplicación de un modelo entrenado en un centro hospitalario a imágenes adquiridas en otro, con distintas configuraciones de resonador o pacientes con diferentes características demográficas [37, 20]. La literatura reciente muestra que en estos escenarios la calidad de la segmentación puede verse afectada drásticamente, y la calibración del modelo tiende a degradarse de forma pronunciada [39].

Es posible cuantificar la dispersión en las salidas del modelo mediante diversas técnicas, como la entropía predictiva, el ensamblado de múltiples redes o el uso de aproximaciones bayesianas (por ejemplo, Monte Carlo Dropout, o redes U-Net probabilísticas). Estos enfoques han sido ampliamente explorados por la comunidad científica, abarcando desde la descomposición teórica de la incertidumbre y los métodos de ensamblado, hasta aproximaciones basadas en la regularización del modelo y análisis críticos sobre su fiabilidad en la práctica [18, 22, 33, 30, 1]. Sin embargo, su aplicación en contextos específicos como la calibración y el cambio de dominio para HSB sigue siendo un área poco explorada.

Aunque existen múltiples métodos de regularización que buscan mejorar la estimación de la incertidumbre en modelos profundos, como el ajuste de la entropía de las predicciones, la penalización de la confianza en errores o el suavizado de las etiquetas de referencia para evitar una certeza excesiva [34, 24, 40, 47], gran parte de los estudios previos se han limitado a escenarios con datos del mismo dominio de entrenamiento, sin explorar en profundidad su comportamiento bajo cambios de dominio. En el contexto de esta tesis, cuando se aplican estos modelos a poblaciones o centros de adquisición distintos, las variaciones en protocolos de adquisición, demografía de los pacientes o distribución de intensidades suelen provocar una degradación tanto en la precisión de la segmentación como en la calibración de las probabilidades [15, 34].

3.2 Materiales y métodos

Esta sección describe el flujo de trabajo propuesto para abordar la estimación de incertidumbre en la segmentación de HSB mediante técnicas de regularización por entropía. En primer lugar, se introduce la formulación del problema y la notación empleada para entrenar un modelo de segmentación con funciones de costo habituales. A continuación, se presenta la *estimación de incertidumbre por entropía* como la métrica central para cuantificar la confianza en cada vóxel. Seguidamente, se explican las *estrategias de regularización por entropía*, diseñadas para reducir el exceso de confianza de las predicciones y mejorar la capacidad del modelo de reflejar adecuadamente la incertidumbre ante distribuciones de datos no contempladas en el entrenamiento. Finalmente, se describen los procedimientos de evaluación, tanto cuantitativos como cualitativos, que garantizan la validación rigurosa de cada componente de la metodología.

Sea $S : X \to Y$ un modelo de segmentación que, dada una imagen X, produce un mapa de segmentación probabilística a nivel de vóxel Y. Para cada vóxel i, el mapa Y asigna una probabilidad y_i a la clase de lesión HSB y $1 - y_i$ a la clase de tejido sano. S puede implementarse como una red neuronal convolucional de tipo codificador-decodificador basada en U-Net [42], aunque el enfoque es aplicable a otras arquitecturas que produzcan mapas de probabilidad a nivel de vóxel.

La red se entrena utilizando la entropía cruzada a nivel de vóxel, una función de costo que, en comparación con otras alternativas como Soft-Dice [35], tiende a producir modelos mejor calibrados [34, 53].Para cuantificar la confianza de las predicciones, este trabajo se centra en la entropía de la predicción, una métrica ampliamente utilizada [38] que mide la dispersión de la distribución de probabilidad en la salida del modelo. El objetivo es optimizar estas estimaciones por entropía para que sirvan como indicadores fiables de error frente a cambios de dominio.

3.2.1 Estimación de incertidumbre por entropía

La incertidumbre en las predicciones del modelo puede estimarse utilizando su entropía. Para la segmentación binaria, se ha empleado la entropía binaria de la región segmentada para proporcionar información sobre los niveles de confianza asociados con las predicciones.

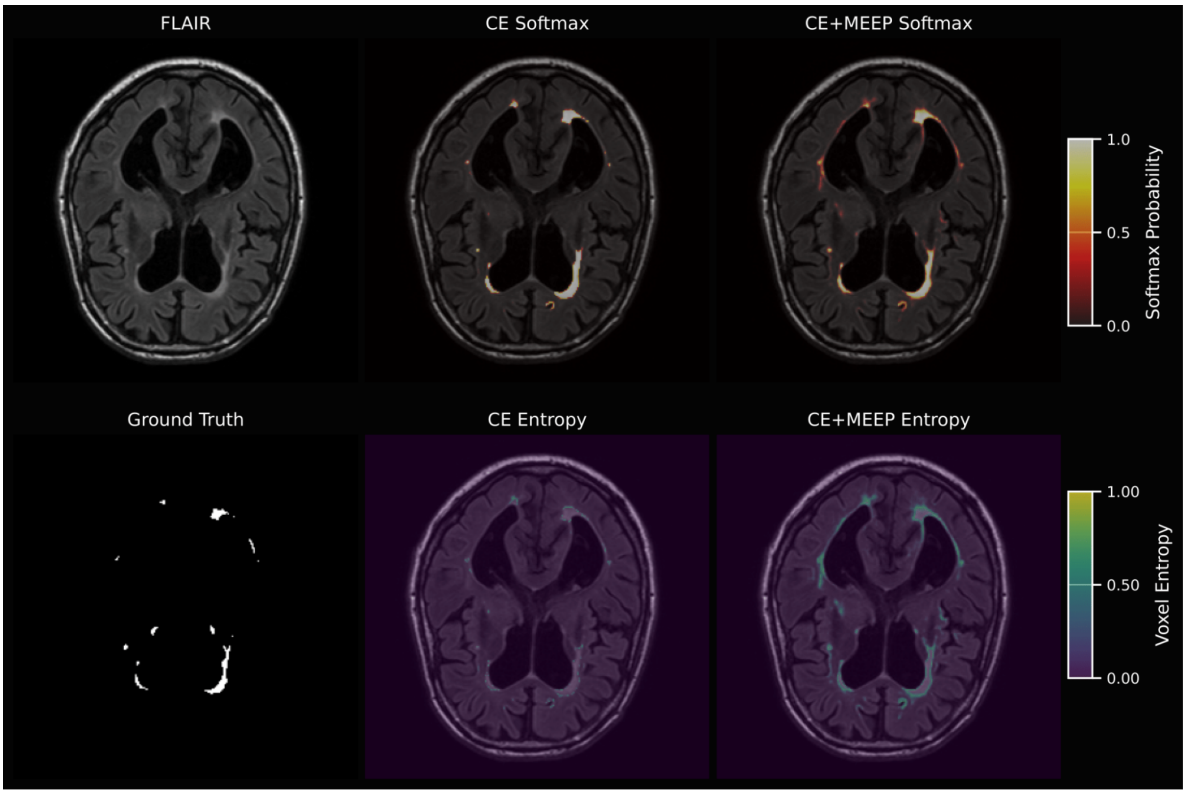


Figura 3.1: Secuencia FLAIR de RM de entrada (arriba izquierda) y segmentación de referencia (abajo izquierda) para HSB. Estas se muestran junto con las salidas de probabilidad softmax de CE Softmax (arriba centro) y CE_{MEEP} Softmax (arriba derecha), y sus respectivos mapas de entropía de vóxeles: CE Entropy (abajo centro) y CE_{MEEP} Entropy (abajo derecha). Notablemente, los mapas de entropía de CE_{MEEP} destacan más distintamente la incertidumbre en pequeñas HSB visibles en la segmentación de referencia, comparado con el mapa de entropía CE.

Como se ilustra en la Figura 3.1, los mapas de entropía permiten visualizar las regiones donde el modelo expresa mayor incertidumbre, siendo especialmente útiles para identificar áreas de difícil segmentación.

Dada una distribución de probabilidad de Bernoulli parametrizada por p, su entropía binaria se define como

$$H_b(p) = -p\log_2(p) - (1-p)\log_2(1-p),$$

donde p representa la probabilidad de que un vóxel o punto de datos pertenezca a la clase de interés (en este caso, la lesión). La entropía binaria H_b varía de 0 (certeza total) a 1 (incertidumbre máxima), lo que la convierte en una métrica directa para evaluar la confianza de las predicciones a nivel de vóxel y mejorar la interpretabilidad del modelo.

3.2.2 Estrategias de regularización por entropía para mejorar la estimación de incertidumbre

Para estudiar el comportamiento de la estimación de incertidumbre bajo cambio de dominio, se analizan tres estrategias de regularización por entropía que fomentan distribuciones de probabilidad mejor calibradas. Dichas estrategias se incorporan como un término adicional en la función de pérdida, con el fin de promover mayor entropía en las predicciones y evaluar su efectividad en conjuntos de datos con distribuciones distintas a las de entrenamiento. El objetivo principal es analizar cómo estas técnicas afectan la estimación de incertidumbre cuando el modelo se enfrenta a datos fuera de distribución (out-of-distribution, OOD), es decir, datos que provienen de una distribución estadística diferente a la utilizada durante el entrenamiento. En estos escenarios, donde tradicionalmente las redes neuronales tienden a producir predicciones con exceso de confianza sin reflejar adecuadamente la incertidumbre del sistema, es crucial que el modelo pueda expresar su incertidumbre de manera apropiada.

Para implementar estas estrategias, se utilizó la Entropía Cruzada (Cross-Entropy, CE) como función de pérdida base, a la cual se le añadieron los distintos términos de regularización, dando lugar a las variantes CE_{MEALL}, CE_{MEEP} y CE_{KL} que se detallan a continuación.

En términos generales, el modelo se entrena utilizando la función de pérdida

$$L = L_{\text{seg}}(Y, \hat{Y}) + \lambda L_{\text{reg}}(Y),$$

donde L_{seg} es el término de datos correspondiente a la entropía cruzada, calculado al comparar la máscara de segmentación predicha Y con la etiqueta real $\hat{Y}$. El término L_{reg} es la regularización por entropía, ponderada por un factor λ.

Penalización de la confianza general

La primera alternativa, que denominaremos regularización por máxima entropía en todas las predicciones (Maximum Entropy on ALL predictions, MEALL) busca penalizar el exceso de confianza de forma global, promoviendo alta entropía en todas las predicciones de vóxel. Se parte de la idea introducida por [40] en el contexto de clasificación y se adapta a la segmentación de imágenes. Así, la entropía de todas las predicciones y_i se calcula como:

$$\mathscr{L}_a(Y) = -H_b(Y) = -\sum_i y_i \log_2(y_i) - (1-y_i)\log_2(1-y_i),$$

y se incorpora a la pérdida general:

$$L(Y, \hat{Y}) = L_{\text{seg}}(Y, \hat{Y}) + \lambda \mathscr{L}_a(Y).$$

De esta forma, se imponen distribuciones de probabilidad más suaves incluso en predicciones correctas.

Entropía máxima en las predicciones erróneas

El segundo enfoque busca abordar esta limitación al penalizar la baja entropía únicamente en aquellas regiones donde el modelo comete errores. Para ello, se define el regularizador de

entropía en las predicciones erróneas (Maximum Entropy on Erroneous Predictions, MEEP), $L_m(Y_w)$, que actúa sobre el subconjunto de vóxeles mal clasificados Y_w,

$$L_m(Y_w) = -H_b(Y_w) = -\sum_{i \in Y_w} y_i \log_2(y_i) - (1-y_i)\log_2(1-y_i).$$

Este regularizador se añade al término de datos mediante

$$L(Y,\hat{Y}) = L_{\text{seg}}(Y,\hat{Y}) + L_m(Y_w),$$

evitando que el modelo sea excesivamente confiado justo en aquellos vóxeles donde falla, lo que resulta especialmente interesante para escenarios de cambio de dominio con mayor incertidumbre.

Entropía máxima en predicciones erróneas mediante divergencia de Kullback-Leibler

Finalmente, se propone un tercer enfoque, inspirado en los trabajos sobre penalización de la confianza [40], donde también se promueve la alta entropía en los vóxeles mal clasificados, pero minimizando la divergencia de Kullback-Leibler (KL) con respecto a una distribución uniforme. La divergencia KL, $D_{KL}(Q\|P)$, es una medida de la diferencia entre dos distribuciones de probabilidad P y Q [21]. Originalmente desarrollada en el contexto de la teoría de la información, la divergencia KL se utiliza ampliamente en aprendizaje automático para medir qué tan bien una distribución se aproxima a otra, siendo especialmente útil en tareas de regularización y optimización de modelos probabilísticos [7]. Dado que la distribución uniforme representa la máxima entropía, se busca que la distribución de las predicciones erróneas Y_w se asemeje a ella. Para ello, siendo Q una distribución de probabilidad uniforme, el término de regularización se define como

$$L_{KL}(Y_w) = -D_{KL}(Q\|Y_w),$$

quedando la función de pérdida

$$L(Y,\hat{Y}) = L_{\text{seg}}(Y,\hat{Y}) + L_{KL}(Y_w).$$

Aunque tanto $L_{KL}(Y_w)$ como $L_m(Y_w)$ incentivan distribuciones de mayor entropía en las regiones erróneas, la dinámica del gradiente es distinta en cada caso. Mientras que MEEP ajusta la entropía en función de la presencia de errores específicos, la regularización por KL impone una tendencia global hacia distribuciones más uniformes en las regiones mal clasificadas. Evaluar estas diferencias en escenarios de cambio de dominio permite entender mejor cómo cada regularizador afecta a la capacidad del modelo para expresar incertidumbre en datos fuera de distribución.

3.3 Experimentos

Para evaluar el desempeño de los métodos considerados, se diseñaron experimentos que analizan su capacidad para estimar la incertidumbre en la segmentación de HSB, particularmente en presencia de cambio de dominio. Se emplean métricas que permiten medir tanto

la calidad de la segmentación como la calibración del modelo, junto con análisis específicos sobre el comportamiento de la incertidumbre en distintos escenarios.

Para llevar a cabo esta evaluación, los experimentos se estructuraron en dos escenarios fundamentales: en distribución (In-Distribution, ID) y fuera de distribución (Out-of-Distribution, OOD). El escenario ID sirve como punto de referencia, donde el modelo se evalúa en datos con características similares a los de entrenamiento. Por el contrario, el escenario OOD simula el desafío del cambio de dominio, evaluando el modelo en un conjunto de datos completamente distinto, como se detallará en la siguiente sección.

3.3.1 Métricas y procedimientos de evaluación

Dado que los modelos de segmentación suelen ser entrenados en conjuntos de datos limitados en términos de diversidad de adquisición, un modelo que mantenga buenas propiedades de incertidumbre en escenarios fuera de distribución resulta crucial para la aplicabilidad clínica. Por ello, el análisis comparativo de estas técnicas se realiza considerando conjuntos de datos multicéntricos, donde los modelos pueden enfrentar variaciones en los escáneres, protocolos de adquisición y poblaciones de pacientes.

La evaluación del desempeño de los modelos de segmentación de HSB bajo cambio de dominio se centra en tres aspectos fundamentales: la capacidad discriminativa, la calibración de las predicciones y el uso de la entropía como indicador de incertidumbre. El análisis de estos factores permite obtener una visión integral sobre la robustez y fiabilidad del modelo en distintos escenarios.

Métricas de discriminación

Para evaluar la capacidad del modelo para distinguir correctamente entre las clases, se utilizó el Coeficiente Dice, (definido previamente en el Capítulo 2), que cuantifica la superposición entre segmentaciones predichas y de referencia.

Métricas de calibración

Las métricas de calibración son importantes para evaluar qué tan bien las probabilidades predichas de un modelo se alinean con los resultados reales, proporcionando información sobre la fiabilidad de las estimaciones de probabilidad. El error de calibración esperado (ECE) es una métrica adecuada en este sentido. Para calcularlo, primero asignamos cada predicción de vóxel a un intervalo o rango de intensidades, dependiendo del valor de probabilidad predicha. Consideramos una separación de intervalos de 0.1, resultando en $M = 10$ intervalos de la forma $\{B_0 = [0;0,1), B_1 = [0,1;0,2), \ldots, B_{10} = [0,9;1]\}$. El ECE se calcula como

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|,$$

donde $|B_m|$ es el número de muestras en el intervalo B_m, n es el número total de muestras, $\mathrm{acc}(B_m)$ es la exactitud promedio de los vóxeles en el intervalo B_m, y $\mathrm{conf}(B_m)$ es la confianza

promedio del intervalo B_m. Esta confianza se mide habitualmente como la relación entre la cantidad de muestras positivas (de referencia) y la cantidad total de muestras en el intervalo.

Otra herramienta esencial para evaluar la calibración es el gráfico de fiabilidad. Esta representación gráfica muestra la probabilidad predicha promedio, p, frente a la fracción real de positivos, f_p, para cada intervalo. Idealmente, los puntos en un gráfico de fiabilidad deberían estar sobre la línea $p = f_p$, lo que indicaría una calibración perfecta, donde la probabilidad predicha coincide con la frecuencia observada del evento. Esta visualización ayuda a identificar áreas donde el modelo sobreestima o subestima la confianza en sus predicciones.

3.3.2 Conjuntos de datos

Para validar los métodos propuestos de estimación de incertidumbre en la segmentación de HSB frente a cambios de dominio, se analizaron dos conjuntos de datos de resonancia magnética cerebral: el Desafío de segmentación de HSB (WMH Segmentation Challenge, WMH-SC) y la base de datos 3D-MR-MS, centrada en pacientes con esclerosis múltiple (EM), ambas con sus correspondientes segmentaciones de lesiones. Estos conjuntos de datos se describen en detalle a continuación.

Conjunto de datos WMH-SC

El primer conjunto de datos corresponde a una iniciativa asociada con la conferencia MICCAI 2017 en Quebec, que estuvo activa desde 2017 hasta 2022 [20]. Este desafío se diseñó específicamente para proporcionar una referencia para comparar métodos automatizados de segmentación de HSB, centrándose en lesiones que se cree que son de origen vascular.

El conjunto de datos está compuesto por imágenes de RM cerebrales, que incluyen imágenes ponderadas en T1 y FLAIR, acompañadas de anotaciones manuales de HSB realizadas por expertos. Se divide en dos subconjuntos principales: un conjunto de entrenamiento que consta de 60 pares de imágenes (T1 y FLAIR) recopiladas de tres instituciones diferentes, y un conjunto de prueba que contiene 110 pares de imágenes de cinco escáneres distintos. En este estudio se utilizan solo los datos de entrenamiento, dado que los de prueba no contienen segmentaciones manuales.

Conjunto de datos 3D-MR-MS

El segundo conjunto de datos, denominado **3D-MR-MS** [26], ofrece una perspectiva diferente, concentrándose en imágenes de RM adquiridas de pacientes diagnosticados con EM. Una característica clave de este conjunto de datos es la inclusión de segmentaciones de HSB basadas en un consenso, lo que proporciona una referencia fiable para el entrenamiento y la evaluación. Este conjunto de datos fue sometido a varios pasos de preprocesamiento, incluyendo corrección de sesgo y co-registro. Incluye múltiples modalidades de RM: ponderada en T1 (T1W), ponderada en T1 con contraste (T1WKS), ponderada en T2 (T2W) y FLAIR.

El uso combinado de estos dos conjuntos de datos, uno que representa un desafío multicéntrico centrado en la segmentación de HSB y el otro centrado en pacientes con EM, constituye un escenario ideal para evaluar casos de cambio de dominio. La variedad en los

protocolos de escaneo, las poblaciones de pacientes y las características de las lesiones permitirá evaluar el impacto de dichos cambios de dominio en la calidad de las estimaciones de incerteza.

Con estos dos conjuntos de datos definidos, la configuración experimental para los análisis ID y OOD es la siguiente: todos los modelos fueron entrenados utilizando el conjunto de datos WMH-SC. La evaluación ID se realizó sobre una parte de este mismo conjunto, mientras que la evaluación OOD se llevó a cabo utilizando la totalidad del conjunto de datos 3D-MR-MS.

3.4 Resultados

A continuación se presentan los resultados experimentales obtenidos al aplicar la metodología propuesta en los conjuntos de datos descritos. El objetivo es validar la capacidad del modelo para estimar la incertidumbre de manera efectiva y para generar mapas de confianza que se correlacionen con errores reales de segmentación, especialmente en escenarios de cambio de dominio. Se incluyen comparaciones entre diferentes estrategias de regularización por entropía, lo cual permite evaluar de forma integral el desempeño y la robustez de la solución planteada bajo diversos escenarios y métricas de evaluación.

3.4.1 La entropía como indicador de errores en escenarios con cambio de dominio

El objetivo principal de esta subsección es investigar si la entropía de las predicciones del modelo puede servir como un indicador confiable para anticipar errores de segmentación, especialmente cuando existe un cambio de dominio. Para ello, se analizó la correlación de Pearson entre la entropía promedio de los vóxeles clasificados como HSB y el coeficiente de Dice (que mide la calidad de la segmentación). La hipótesis de partida es que una mayor incertidumbre (mayor entropía) debería correlacionarse con una menor calidad de segmentación (menor Dice), y que esta correlación debería ser más pronunciada en presencia de un cambio de dominio.

Para simular una situación clínica realista, donde no se dispone de la segmentación de referencia, se calculó la entropía promedio solo para aquellos vóxeles que el modelo predijo como pertenecientes a la lesión (probabilidad $y_i > 0{,}5$). La Figura 3.2 muestra un diagrama de dispersión que compara la entropía promedio con el coeficiente de Dice para cada imagen, tanto para los datos ID como OOD, y para las cuatro estrategias de entrenamiento consideradas: entropía cruzada estándar (CE), *CE* regularizada con entropía máxima en predicciones erróneas (CE_{MEEP}), *CE* regularizada con divergencia de Kullback-Leibler (CE_{KL}), y *CE* con entropía máxima en todas las predicciones (CE_{MEALL}). Se ajustaron líneas de regresión lineal a cada conjunto de puntos, y se calculó el coeficiente de correlación de Pearson correspondiente.

Los resultados muestran una correlación negativa entre el Dice y la entropía para todas las funciones de pérdida, lo que confirma la hipótesis inicial: a mayor incertidumbre, menor calidad de segmentación. Sin embargo, la intensidad de esta correlación varía significativamente

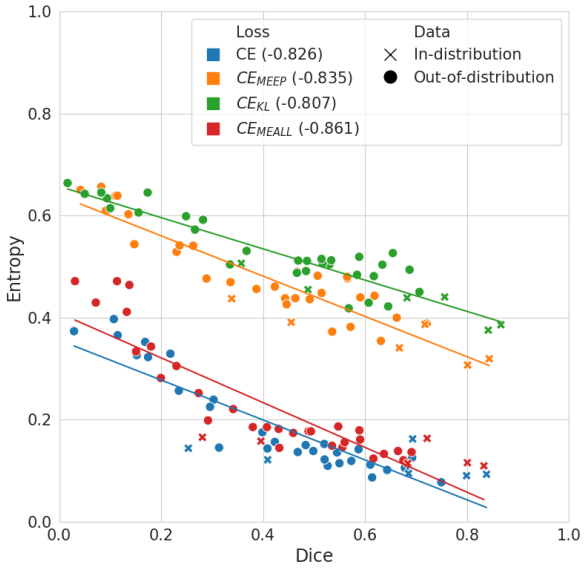


Figura 3.2: Diagrama de dispersión comparando la entropía de las predicciones del primer plano y el coeficiente de Dice, por imagen, para pacientes ID y OOD. El coeficiente de correlación de Pearson entre entropía y Dice se muestra entre paréntesis en el cuadro de la leyenda. Se puede observar que las estimaciones de entropía para los modelos MEEP y KL muestran una mejor anticorrelación, actuando así como mejores predictores de posibles fallos.

entre las diferentes estrategias. Las estrategias CE_{MEEP} y CE_{KL} presentan las correlaciones negativas más fuertes (más cercanas a -1), tanto para datos ID como OOD. Esto sugiere que estas estrategias proporcionan estimaciones de incertidumbre más informativas, ya que sus valores de entropía reflejan mejor el rendimiento real de la segmentación. En contraste, las estrategias CE y CE_{MEALL} muestran correlaciones más débiles.

La Figura 3.3 profundiza en el análisis de la incertidumbre, examinando su distribución para diferentes tipos de errores de predicción: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Cada punto en el diagrama de dispersión representa un vóxel, y se utilizan colores para distinguir entre datos ID (azul) y OOD (naranja).

Se observan los siguientes patrones:

- **TP:** Como era de esperar, los vóxeles correctamente clasificados como lesión (TP) presentan, en general, baja incertidumbre. Las estrategias CE y CE_{MEALL} muestran los valores de entropía más bajos (mayor confianza), mientras que CE_{MEEP} y CE_{KL} presentan valores superiores, aunque todavía relativamente bajos. Adicionalmente, para todas las estrategias, se observa que la incertidumbre de los TP es consistentemente mayor en los datos OOD en comparación con los datos ID, lo cual es un resultado deseable, ya que indica que el modelo es más inseguro al realizar predicciones correctas sobre datos de un dominio no visto.
- **TN:** Los vóxeles correctamente clasificados como tejido sano (TN) también presentan

baja incertidumbre, con medianas cercanas a cero para todas las estrategias.

- **FP:** Los falsos positivos (vóxeles clasificados incorrectamente como lesión) muestran, en promedio, una incertidumbre significativamente mayor que los TP y TN. Este es un resultado deseable, ya que indica que la entropía puede ayudar a identificar predicciones erróneas. Las estrategias CE_{MEEP} y CE_{KL} exhiben los valores de entropía más altos para los FP. Es notable, además, que estos valores son consistentemente más elevados en el escenario OOD que en el ID, lo cual es un comportamiento esperado y útil: el modelo no solo expresa duda en sus errores, sino que esta duda se intensifica al enfrentarse a datos de un dominio desconocido.
- **FN:** Los falsos negativos (vóxeles de lesión incorrectamente clasificados como tejido sano) también muestran, en general, una incertidumbre elevada. Nuevamente, CE_{MEEP} y CE_{KL} presentan los valores de entropía más altos, lo que sugiere que estas estrategias son más sensibles para indicar incertidumbre en los errores por omisión de lesiones.

La Figura 3.4 muestra tres comparaciones importantes: el panel izquierdo compara la entropía promedio de los vóxeles predichos como positivos (HSB) para los datos ID y OOD, utilizando diagramas de caja (boxplots). Se observa que las estrategias CE y CE_{MEALL} no muestran diferencias significativas en la entropía entre los datos ID y OOD (según una prueba de Mann-Whitney U). En cambio, CE_{MEEP} y CE_{KL} sí muestran diferencias significativas, con valores de entropía notablemente más altos para los datos OOD. Esto indica que estas estrategias son capaces de "detectar" el cambio de dominio a través de un aumento en la incertidumbre de las predicciones. El panel central muestra el rendimiento en términos del coeficiente de Dice para los datos ID y OOD. Como era de esperar, el rendimiento disminuye para los datos OOD en todas las estrategias, pero esta disminución es, en general, más pronunciada para CE y CE_{MEALL}. El panel derecho ilustra las distancias de Hausdorff, proporcionando información sobre la precisión en la localización de fronteras.

3.4.2 Análisis de incertidumbre en relación al tamaño de lesión

En estudios previos se ha demostrado que la segmentación de HSB es más difícil para lesiones pequeñas [9]. En consecuencia, es razonable esperar que la incertidumbre de las predicciones sea mayor para lesiones de menor tamaño.

Para investigar esta relación se agruparon las lesiones en tres categorías según su volumen: menores de 5 mL, entre 5 mL y 15 mL, y mayores de 15 mL. La Figura 3.5 muestra la entropía promedio de los vóxeles predichos como positivos para cada categoría de tamaño de lesión y para cada estrategia de entrenamiento, separando los datos ID y OOD. Se observa que las lesiones más pequeñas tienden a tener mayor entropía en todas las funciones de pérdida. Esta observación se alinea con la dificultad de alcanzar consenso experto en las etiquetas de referencia para lesiones pequeñas, ya que su apariencia sutil puede dificultar su identificación y delimitación. Las lesiones más grandes generalmente se asocian con valores de entropía más bajos, indicando mayor confianza del modelo, y esta tendencia se observa consistentemente tanto para casos ID como OOD.

Cuantitativamente, con CE la mediana de entropía para lesiones pequeñas (<5 mL) fue aproximadamente 0.58, comparada con 0.23 para lesiones grandes (>15 mL), ilustrando la disminución en la incertidumbre del modelo con el aumento del tamaño de la lesión. Notablemente, la estrategia de regularización CE_{MEEP} apunta específicamente a estas lesiones pequeñas al empujar los niveles de incertidumbre hacia el máximo, reflejando la ambigüedad inherente y el potencial de desacuerdo en estos casos. Este enfoque dirigido podría ser particularmente valioso en la práctica clínica, ya que permite al modelo marcar sus propias limitaciones y promover una mayor investigación o consulta para lesiones pequeñas inciertas.

3.4.3 Calibración del modelo en escenarios de cambio de dominio

Finalmente, para evaluar el impacto del cambio de dominio en la calibración del modelo, se analizan diagramas de fiabilidad y ECE para cada función de pérdida considerando tanto escenarios ID como OOD (Figura 3.6). En el escenario ID, CE_{MEEP} supera a otras pérdidas en términos de ECE, mientras que en el escenario OOD, todas las funciones de pérdida exhiben peor calibración, excepto por la pérdida basada en KL, que demuestra calibración superior y robustez al cambio de dominio.

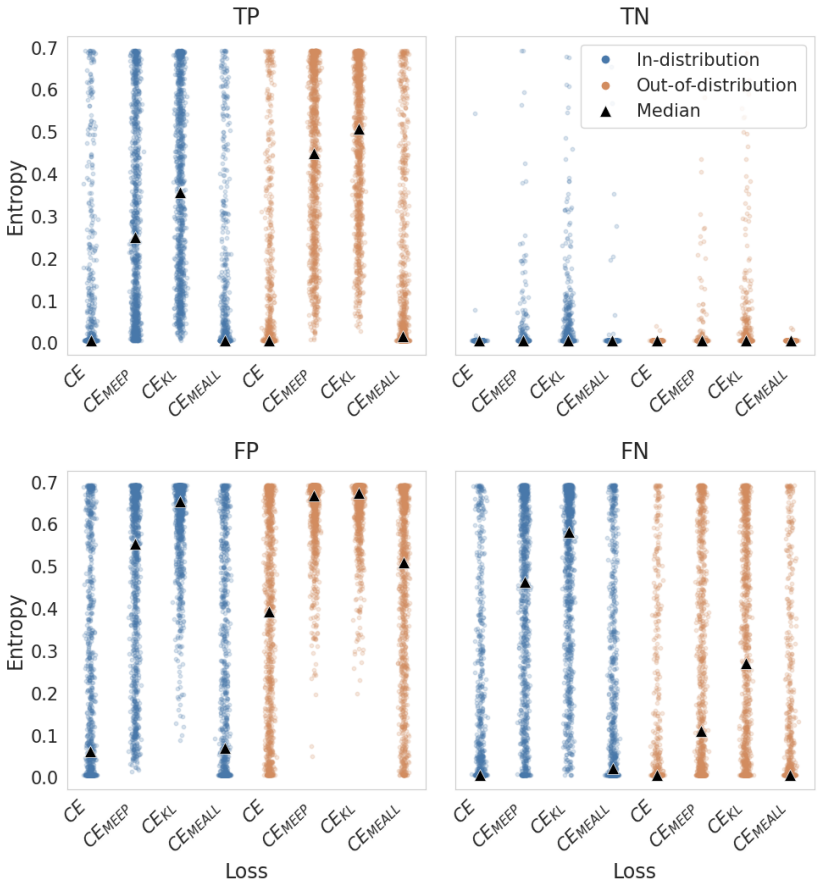


Figura 3.3: Distribución de estimaciones de incertidumbre a través de diferentes resultados de predicción (verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos) para varias estrategias de entrenamiento bajo escenarios ID y OOD. Cada punto representa un vóxel, con azul indicando datos ID y naranja representando datos OOD. El eje x muestra diferentes estrategias de entrenamiento, mientras que el eje y representa los valores de entropía. Los triángulos negros denotan los valores medianos de entropía. Esta visualización permite comparar comportamientos de incertidumbre a través de diferentes funciones de pérdida, revelando cómo métodos como CE_{MEEP} y CE_{KL} tienden a producir incertidumbres más altas, particularmente para falsos positivos y falsos negativos, tanto en configuraciones ID como OOD.

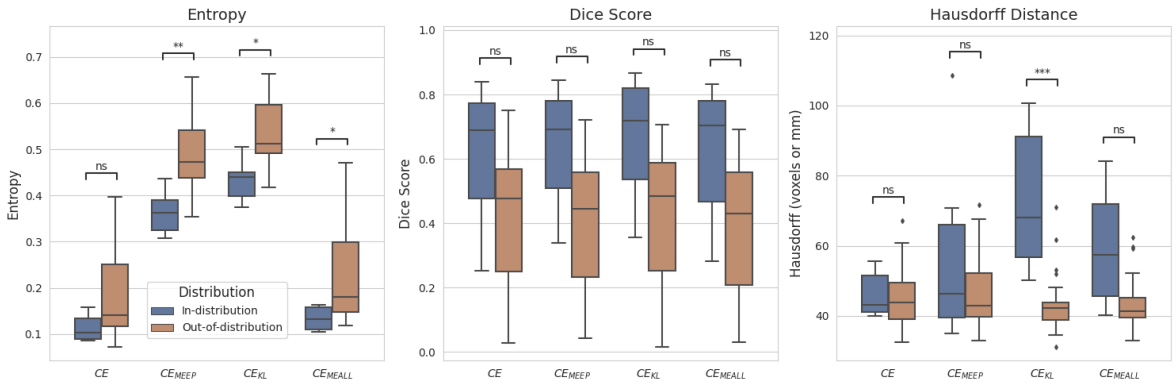


Figura 3.4: Boxplots comparando métricas a través de datos en distribución (ID) y fuera de distribución (OOD) para diferentes funciones de pérdida. (Izquierda): Entropía promedio para vóxeles predichos como positivos, mostrando un aumento general en la incertidumbre bajo cambio de dominio, especialmente para CE_{MEEP} y CE_{KL}. (Centro): Rendimiento del puntaje Dice a través de funciones de pérdida, con puntajes ID consistentemente más altos que los puntajes OOD. (Derecha): Distancias de Hausdorff ilustrando el rendimiento de localización de fronteras a través de casos ID y OOD. La significancia estadística se indica donde es aplicable según la prueba de Mann–Whitney U.

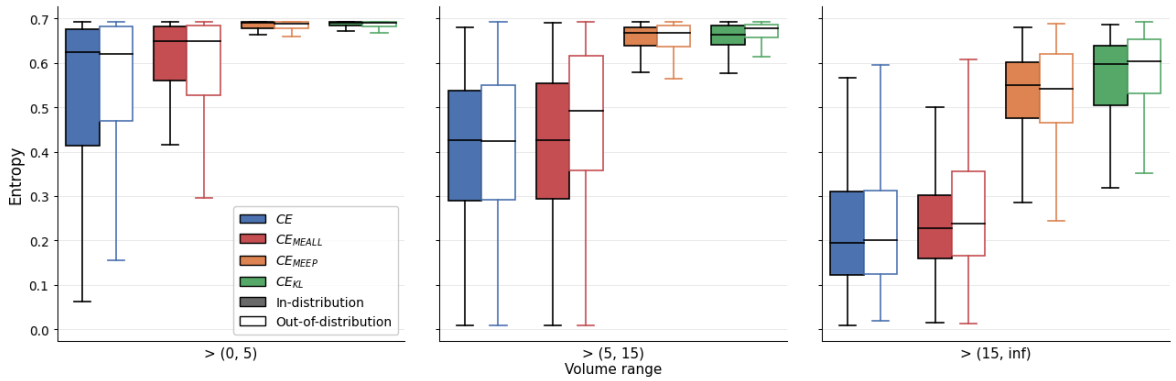


Figura 3.5: Boxplots comparando la entropía promedio para vóxeles predichos como positivos a través de diferentes estrategias en tres rangos de volumen de lesión. El gráfico distingue entre datos ID (cajas llenas) y OOD (cajas vacías). Observamos que volúmenes de lesión más grandes generalmente se asocian con menor entropía, confirmando que puede servir como indicador de incertidumbre del modelo. Notablemente, esta tendencia se conserva tanto para casos ID como OOD.

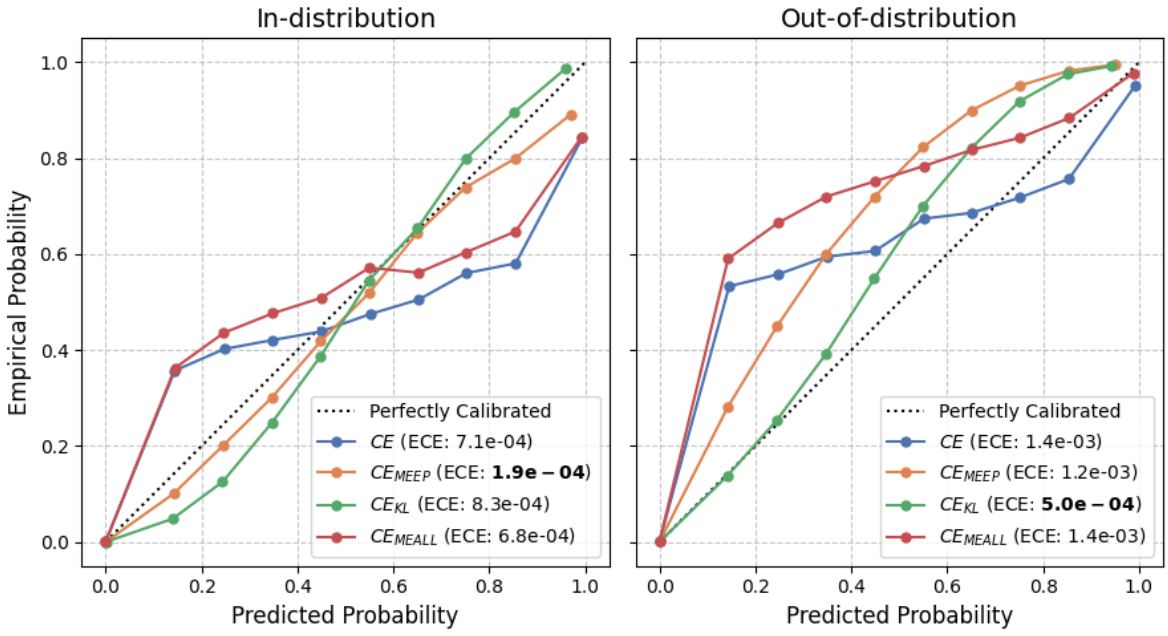


Figura 3.6: Gráficos de fiabilidad para diferentes funciones de pérdida en datos ID y OOD. Cada línea de color corresponde a una función de pérdida diferente, con el ECE mostrado entre paréntesis (los mejores se muestran en negrita). Los puntos por encima de la diagonal indican subconfianza, mientras que los puntos por debajo indican exceso de confianza. Un modelo bien calibrado debería aproximarse a la línea diagonal punteada (que representa la calibración perfecta).

Capítulo 4:
Conclusiones

En esta tesis se abordaron dos problemáticas relevantes en el ámbito de neuroimágenes y la neurocirugía mediante el uso de técnicas de aprendizaje profundo: la reconstrucción craneal y la segmentación de HSB. A continuación, se resumen las principales contribuciones y conclusiones de este trabajo, así como las perspectivas para futuras investigaciones.

En cuanto a la reconstrucción craneal, se desarrollaron y evaluaron modelos basados en redes neuronales convolucionales que lograron avances significativos en la automatización de este proceso. Se propuso una metodología de craniectomía virtual que permite generar datos sintéticos para entrenar modelos en un esquema autosupervisado, eliminando la necesidad de datos anotados manualmente, lo que representa una solución innovadora a una limitación clave en este campo. Asimismo, se presentaron arquitecturas avanzadas como DE-UNet y su variante con restricciones anatómicas, DE-UNet-Shape, las cuales demostraron ser efectivas en la reconstrucción de defectos craneales. Estos modelos no solo superaron a los métodos manuales en la estimación del volumen del colgajo óseo, sino que también mostraron un desempeño competitivo en el AutoImplant Challenge, especialmente en casos de defectos complejos y de gran tamaño. Además, la incorporación de restricciones de forma y otra información anatómica conocida a priori permitió mejorar significativamente la coherencia y precisión de las reconstrucciones, abordando con éxito escenarios donde los métodos tradicionales basados en simetría presentan limitaciones. Adicionalmente, la evaluación en el conjunto CENTER-TBI permitió validar el enfoque de reconstrucción y estimación volumétrica en un entorno real, complementando los resultados obtenidos con los métodos propuestos (por ejemplo RS-AE, RS-UNet y sus variantes con restricciones anatómicas) y demostrando la flexibilidad de la craniectomía virtual bajo distintos escenarios clínicos.

En lo que respecta a la segmentación de hiperintensidades de la sustancia blanca, este trabajo exploró técnicas innovadoras para mejorar tanto la robustez frente al cambio de dominio como la estimación de incertidumbre, aspectos críticos para su aplicación clínica. Se analizaron los efectos del cambio de dominio en la calibración de modelos de segmentación, evidenciando la necesidad de enfoques robustos frente a variaciones en los datos. Para abordar esta problemática, se propuso evaluar estrategias basadas en la regularización por entropía, como CE_{MEEP} y CE_{KL} bajo este contexto, que mejoraron la estimación de incertidumbre y redujeron las predicciones con exceso de confianza. Los resultados demostraron que las estimaciones de incertidumbre basadas en entropía pueden utilizarse como indicadores para anticipar errores de segmentación en dominios no vistos, mostrando su utilidad en aplicaciones prácticas. Además, se evidenció que la elección de la función de pérdida influye significativamente en la calidad de la cuantificación de la incertidumbre, así como en la calibración del modelo. La validación de estos métodos en el contexto específico de segmentación demostró la efectividad de las técnicas propuestas para mejorar la robustez de

los modelos ante cambios de dominio, lo que permitirá que los especialistas médicos puedan tomar decisiones más informadas basadas en la confiabilidad de las predicciones.

Este trabajo constituye un aporte significativo en la aplicación de técnicas de aprendizaje profundo a problemas específicos en neurocirugía y segmentación de lesiones cerebrales. Los resultados, así como las limitaciones y posibles mejoras expuestas, sientan las bases para la optimización de estos métodos y su validación futura en contextos clínicos más amplios, enfocándose en aplicaciones concretas con un potencial de impacto en la práctica neuroquirúrgica. A su vez, los enfoques desarrollados han ofrecido avances tangibles en la automatización de la reconstrucción craneal y la segmentación de lesiones en imágenes médicas. A través de la implementación de la craniectomía virtual y la incorporación de técnicas de regularización por entropía, se ha logrado superar limitaciones tradicionales como la escasez de datos reales para entrenar y las variaciones en los protocolos de adquisición de imágenes. No obstante, la transición de estos métodos a entornos clínicos reales exige una validación más rigurosa, especialmente en cuanto a la personalización y adaptabilidad de los modelos ante la diversidad de casos clínicos. Los futuros desarrollos deberían centrarse en la integración de características individuales de los pacientes, como la variabilidad anatómica, y en la creación de arquitecturas que trabajen directamente con representaciones más detalladas de la geometría craneal, como mallas 3D. La validación clínica prospectiva será esencial para evaluar el impacto de estos avances en la práctica neuroquirúrgica, mientras que la mejora continua en la estimación de incertidumbre y su integración con el juicio clínico podría potenciar aún más la robustez y utilidad de las herramientas propuestas.

Capítulo 5:
Publicaciones

A continuación se listan todos los trabajos relacionados con la tesis en los que se participó durante el desarrollo del doctorado.

Publicaciones en revistas

- **Matzkin, F.**, Larrazabal, A., Milone, D. H., Dolz, J., y Ferrante, E. (2025). Improving uncertainty estimates under domain shift in white matter hyperintensity segmentation via maximum-entropy regularization. Computers in Biology and Medicine, Elsevier (Q1, IF: 6.3)

- Li, J., Pimentel, P., Szengel, A., Ehlke, M., **Matzkin, F.**, et al. (2021). AutoImplant 2020-first MICCAI challenge on automatic cranial implant design. IEEE Transactions on Medical Imaging, 40(9), 2329-2342.

Trabajos en eventos científicos de alto impacto

- **Matzkin, F.**, Newcombe, V., Glocker, B., y Ferrante, E. (2020). Cranial implant design via virtual craniectomy with shape priors. In Towards the Automatization of Cranial Implant Design in Cranioplasty: First Challenge, AutoImplant 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 1 (pp. 37-46). Springer International Publishing.

- **Matzkin, F.**, Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., ... y Ferrante, E. (2020). Self-supervised skull reconstruction in brain CT images with decompressive craniectomy. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23 (pp. 390-399). Springer International Publishing.

Bibliografía

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (Dec 2021). https://doi.org/10.1016/j.inffus.2021.05.008, http://dx.doi.org/10.1016/j.inffus.2021.05.008

[2] Alkhaibary, A., Alharbi, A., Alnefaie, N., Oqalaa Almubarak, A., Aloraidi, A., Khairy, S.: Cranioplasty: A comprehensive review of the history, materials, surgical aspects, and complications. World Neurosurgery **139**, 445–452 (Jul 2020). https://doi.org/10.1016/j.wneu.2020.04.211, http://dx.doi.org/10.1016/j.wneu.2020.04.211

[3] Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)

[4] Balakrishnan, R., Hernández, M.d.C.V., Farrall, A.J.: Automatic segmentation of white matter hyperintensities from brain magnetic resonance images in the era of deep learning and big data–a systematic review. Computerized Medical Imaging and Graphics **88**, 101867 (2021)

[5] Bazarian, J.J., Biberthaler, P., Welch, R.D., Lewis, L.M., Barzo, P., Bogner-Flatz, V., Brolinson, P.G., Büki, A., Chen, J.Y., Christenson, R.H., et al.: Serum gfap and uch-l1 for prediction of absence of intracranial injuries on head ct (alert-tbi): a multicentre observational study. The Lancet Neurology **17**(9), 782–789 (2018)

[6] Berger, M., Tagliasacchi, A., Seversky, L., Alliez, P., Levine, J., Sharf, A., Silva, C.: State of the art in surface reconstruction from point clouds. Eurographics 2014-State of the Art Reports **1**(1), 161–185 (2014)

[7] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)

[8] Campe, G.v., Pistracher, K.: Patient specific implants (psi). In: Cranial Implant Design Challenge. pp. 1–9. Springer (2020)

[9] Chaves, H., Serra, M., Shalom, D., et al.: Assessing robustness and generalization of a deep neural network for brain ms lesion segmentation on real-world data. European Radiology **34**, 2024–2035 (2024). https://doi.org/10.1007/s00330-023-10093-5

[10] Chen, X., Xu, L., Li, X., Egger, J.: Computer-aided implant design for the restoration of cranial defects. Scientific Reports **7**(1) (Jun 2017). https://doi.org/10.1038/s41598-017-04454-6, http://dx.doi.org/10.1038/s41598-017-04454-6

[11] Cootes, T.F., Taylor, C.J.: Statistical models of appearance for medical image analysis and computer vision. In: Medical Imaging 2001: Image Processing. vol. 4322, pp. 236–248. SPIE (2001)

[12] van Eijnatten, M., van Dijk, R., Dobbe, J., Streekstra, G., Koivisto, J., Wolff, J.: CT image segmentation methods for bone used in medical additive manufacturing. Medical Engineering & Physics **51**, 6–16 (Jan 2018). https://doi.org/10.1016/j.medengphy.2017.10.008, https://doi.org/10.1016/j.medengphy.2017.10.008

[13] Ellis, D.G., Aizenberg, M.R.: Deep learning using augmentation via registration: 1st place solution to the autoimplant 2020 challenge. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 47–55. Springer (2020)

[14] Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., Wolz, R., Valdés-Hernández, M.d.C., Dickie, D.A., Wardlaw, J., et al.: White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical **17**, 918–934 (2018)

[15] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

[16] Huang, K.C., Liao, C.C., Xiao, F., Liu, C.C.H., Chiang, I.J., Wong, J.M.: Automated volumetry of postoperative skull defect on brain ct. Biomedical Engineering: Applications, Basis and Communications **25**(03), 1350033 (2013)

[17] Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4037–4058 (2020)

[18] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? (2017), https://arxiv.org/abs/1703.04977

[19] Kodym, O., Španěl, M., Herout, A.: Cranial defect reconstruction using cascaded cnn with alignment. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 56–64. Springer (2020)

[20] Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H.,

Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Transactions on Medical Imaging **38**(11), 2556–2568 (2019). https://doi.org/10.1109/TMI.2019.2905770

[21] Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1), 79–86 (1951)

[22] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (2017), `https://arxiv.org/abs/1612.01474`

[23] Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: Anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Transactions on Medical Imaging (2020). https://doi.org/10.1109/TMI.2020.3005297

[24] Larrazabal, A., Martínez, C., Dolz, J., Ferrante, E.: Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14222, pp. 273–283. Springer, Cham (2023). https://doi.org/`10.1007/978-3-031-43898-1_27`

[25] Lee, M.C.H., Oktay, O., Schuh, A., Schaap, M., Glocker, B.: Tetris: Template transformer networks for image segmentation with shape priors. IEEE Transactions on Medical Imaging **38**(9), 2106–2117 (2019)

[26] Lesjak, Z., Galimzianova, A., Koren, A., et al.: A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics **16**, 51–63 (2018). https://doi.org/`10.1007/s12021-017-9348-7`

[27] Li, J., Pimentel, P., Szengel, A., Ehlke, M., Lamecker, H., Zachow, S., Estacio, L., Doenitz, C., Ramm, H., Shi, H., et al.: Autoimplant 2020-first miccai challenge on automatic cranial implant design. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 1–15. Springer (2020)

[28] Li, J., Pimentel, P., Szengel, A., Ehlke, M., Lamecker, H., Zachow, S., Estacio, L., Doenitz, C., Ramm, H., Shi, H., et al.: Autoimplant 2020-first miccai challenge on automatic cranial implant design. IEEE transactions on medical imaging **40**(9), 2329–2342 (2021)

[29] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[30] Liu, Y., Pagliardini, M., Chavdarova, T., Stich, S.U.: The peril of popular deep learning uncertainty estimation methods (2021), `https://arxiv.org/abs/2112.05000`

[31] Maas, A.I.R., Menon, D.K., Steyerberg, E.W., Citerio, G., Lecky, F., Manley, G.T., Hill, S., Legrand, V., Sorgner, A., CENTER-TBI Participants and Investigators: Collaborative european NeuroTrauma effectiveness research in traumatic brain injury (CENTER-TBI): a prospective longitudinal observational study. Neurosurgery **76**(1), 67–80 (Jan 2015)

[32] Marstal, K., Berendsen, F., Staring, M., Klein, S.: Simpleelastix: A user-friendly, multilingual library for medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 134–142 (2016)

[33] McInerney, J., Kallus, N.: Variation due to regularization tractably recovers bayesian deep learning (2025), https://arxiv.org/abs/2403.10671

[34] Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020)

[35] Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016). https://doi.org/10.1109/3DV.2016.79

[36] Morais, A., Egger, J., Alves, V.: Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder. In: World conference on information systems and technologies. pp. 151–160. Springer (2019)

[37] Mozafari, A.S., Gomes, H.S., Leão, W., Gagné, C.: Unsupervised temperature scaling: An unsupervised post-processing calibration method of deep networks. arXiv preprint arXiv:1905.00174 (2019)

[38] Nair, T., Precup, D., Arnold, D., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Analysis **59**, 101557 (2020). https://doi.org/10.1016/j.media.2019.101557

[39] Pampari, A., Ermon, S.: Unsupervised calibration under covariate shift. arXiv preprint arXiv:2006.16405 (2020)

[40] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLR Workshop (2017), https://openreview.net/forum?id=HyhbYrGYe

[41] Pimentel, P., et al.: Automated virtual reconstruction of large skull defects using statistical shape models and generative adversarial networks. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 16–27. Springer (2020)

[42] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

[43] Seeram, E.: Computed Tomography - E-Book: Physical Principles, Clinical Applications, and Quality Control. Elsevier Health Sciences (2015), https://books.google.com.ar/books?id=DTCDCgAAQBAJ

[44] Shi, H., Chen, X.: Cranial implant design through multiaxial slice inpainting using deep learning. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 28–36. Springer (2020)

[45] Stieglitz, L.H., Fung, C., Murek, M., Fichtner, J., Raabe, A., Beck, J.: What happens to the bone flap? long-term outcome after reimplantation of cryoconserved bone flaps in a consecutive series of 92 patients. Acta neurochirurgica **157**(2), 275–280 (2015)

[46] Stieglitz, L.H., Gerber, N., Schmid, T., Mordasini, P., Fichtner, J., Fung, C., Murek, M., Weber, S., Raabe, A., Beck, J.: Intraoperative fabrication of patient-specific moulded implants for skull reconstruction: single-centre experience of 28 cases. Acta neurochirurgica **156**, 793–803 (2014)

[47] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)

[48] Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC medical imaging **15**, 1–28 (2015)

[49] Vancraen, F.: How to design a patient-specific cranial plate (2018), https://www.youtube.com/watch?v=23WZPl11HvE, disponible en: https://www.youtube.com/watch?v=23WZPl11HvE

[50] Wang, B., et al.: Cranial implant design using a deep learning method with anatomical regularization. In: Towards the Automatization of Cranial Implant Design in Cranioplasty. pp. 85–93. Springer (2020)

[51] Wardlaw, J.M., Smith, E.E., Biessels, G.J., Cordonnier, C., Fazekas, F., Frayne, R., Lindley, R.I., O'Brien, J.T., Barkhof, F., Benavente, O.R., et al.: Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. The Lancet Neurology **12**(8), 822–838 (2013)

[52] Xiao, F., Chiang, I.J., Hsieh, T.M.H., Huang, K.C., Tsai, Y.H., Wong, J.M., Ting, H.W., Liao, C.C.: Estimating postoperative skull defect volume from ct images using the abc method. Clinical Neurology and Neurosurgery **114**(3), 205 – 210 (2012). https://doi.org/https://doi.org/10.1016/j.clineuro.2011.10.003, http://www.sciencedirect.com/science/article/pii/S0303846711003076

[53] Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C., Yang, G.: Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. Journal of Digital Imaging **36**, 739–752 (2023). https://doi.org/10.1007/s10278-022-00735-3

Apéndices

Contribuciones

Self-supervised skull reconstruction in brain CT images with decompressive craniectomy

En este trabajo se presentaron y compararon distintas estrategias de entrenamiento de modelos profundos de segmentación de imágenes para el diseño automático de implantes craneales para la posterior obtención del indicador de uso clínico ABC, que constituye el volumen del colgajo óseo extraído. Además de elaborar los experimentos, participé en el análisis de los resultados, en la redacción del manuscrito y en su revisión.

Cranial implant design via virtual craniectomy with shape priors

Este trabajo, presentado dentro de la competencia "Autoimplant 2020", tuvo como objetivo adaptar las técnicas de mejor desempeño del trabajo anterior en el contexto de producir imágenes de alta resolución enfocado principalmente en el diseño de implantes. A la estrategia de estimación directa se le agregó también una nueva estrategia que incorpore priors de forma como entrada en la imagen. Además de elaborar los experimentos, participé en el análisis de los resultados, en la redacción del manuscrito y en su revisión. Este trabajo obtuvo el premio al mejor trabajo (Best Paper Award) dentro de la competencia.

AutoImplant 2020-First MICCAI Challenge on Automatic Cranial Implant Design

Desde la organización de la competencia se coordinó la publicación conjunta de un artículo en la revista IEEE Transactions on Medical Imaging (IEEE TMI). Para ello, se realizó un análisis y comparación de los métodos propuestos por los participantes, así como de los desafíos intrínsecos a la competencia en sí. Colaboré en la redacción del trabajo, y en las instancias de revisión.

Improving uncertainty estimates under domain shift in white matter hyperintensity segmentation via maximum-entropy regularization

En este trabajo se investigó el impacto del cambio de dominio en la calibración del modelo y la estimación de incertidumbre en la segmentación de HSB. Se evaluaron distintas técnicas de regularización por entropía para mejorar las estimaciones de incertidumbre. Este trabajo aporta conocimientos valiosos sobre la robustez de los modelos de segmentación en escenarios de cambio de dominio, crucial para su aplicación en entornos clínicos reales. Contribuí al diseño y realicé toda la implementación de los experimentos, aportando también en el análisis de los resultados. Además, participé activamente en la redacción del manuscrito, en la elaboración de las figuras y en el proceso de revisión.

Self-supervised skull reconstruction in brain CT images with decompressive craniectomy

Self-supervised Skull Reconstruction in Brain CT Images with Decompressive Craniectomy

Franco Matzkin[1], Virginia Newcombe[2], Susan Stevenson[2], Aneesh Khetani[2], Tom Newman[2], Richard Digby[2], Andrew Stevens[2], Ben Glocker[3], and Enzo Ferrante[1]

[1] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET, FICH-UNL (Argentina)
[2] Division of Anaesthesia, Department of Medicine, University of Cambridge (UK)
[3] BioMedIA, Imperial College London (UK)

Abstract. Decompressive craniectomy (DC) is a common surgical procedure consisting of the removal of a portion of the skull that is performed after incidents such as stroke, traumatic brain injury (TBI) or other events that could result in acute subdural hemorrhage and/or increasing intracranial pressure. In these cases, CT scans are obtained to diagnose and assess injuries, or guide a certain therapy and intervention. We propose a deep learning based method to reconstruct the skull defect removed during DC performed after TBI from post-operative CT images. This reconstruction is useful in multiple scenarios, e.g. to support the creation of cranioplasty plates, accurate measurements of bone flap volume and total intracranial volume, important for studies that aim to relate later atrophy to patient outcome. We propose and compare alternative self-supervised methods where an encoder-decoder convolutional neural network (CNN) estimates the missing bone flap on post-operative CTs. The self-supervised learning strategy only requires images with complete skulls and avoids the need for annotated DC images. For evaluation, we employ real and simulated images with DC, comparing the results with other state-of-the-art approaches. The experiments show that the proposed model outperforms current manual methods, enabling reconstruction even in highly challenging cases where big skull defects have been removed during surgery.

Keywords: Skull reconstruction · self-supervised learning · decompressive craniectomy

1 Introduction

Decompressive craniectomy (DC) is a surgical procedure performed for controlling the intracranial pressure (ICP) under some abnormal conditions which could be associated with brain lesions such as traumatic brain injury (TBI) [10]. In this procedure, a portion of the skull (bone flap) is removed, alleviating the risks associated with the presence of hematomas or contusions with a significant volume of blood [3]. In order to monitor the patient's condition and potential

complications from the injury, computed tomography scans (CTs) of the affected area are acquired before and after this intervention [2].

Previous works which study the complications that can emerge after DC suggest that the volume of the skull defect is an important parameter to evaluate the decompressive effort [16,14]. A manual method to estimate such volume was proposed by Xiao and co-workers [17]. The authors developed a simple equation relying on three basic manual measurements which are multiplied and provide a good approximation of the real skull defect size. However, this method requires manual intervention and its accuracy is limited by the geometric approximation which does not take into account specific details of the skull shape.

Alternatively, the extracted bone flap volume could be estimated from a 3D model of the defect, which may be also useful for estimating materials and dimensions of eventual cranioplasty custom-made implants [5]. These can be used instead of the stored bone flap after DC, which has shown to carry potential complications if reused [4]. Different methods can estimate such shapes: one strategy is to take advantage of the symmetry present in the images [6]. However, it has the restriction of handling only unilateral DCs. Another simple and effective alternative could be the subtraction of the aligned pre- and post-operative CT scans, highlighting the missing part of the skull. Of course, this cannot be done if the provided data only contains post-operative images, which tends to be a common situation in real clinical scenarios.

We propose a bone flap reconstruction method which directly operates on post-operative CT scans, can handle any type of DC (not only unilateral) and is more accurate than current state-of-the-art manual methods. Our model employs encoder-decoder convolutional neural networks (CNN) and is trained following a self-supervised strategy, in the sense that it only requires images with complete skull for training, and avoids the need for annotated DC images.

Contributions: Our contributions are 3-fold: (i) to our knowledge, this is the first deep learning based model to perform skull reconstruction from brain CT images, (ii) the method outperforms the accuracy of manual and automatic state-of-the-art algorithms both in real and simulated DC and (iii) we introduce a self-supervised training procedure which enables learning skull reconstruction using only complete skull images which are more common than images with DC.

2 Self-supervised skull reconstruction

Our reconstruction method consists of a CNN which operates on binary skull images obtained after pre-processing the CT. We designed a virtual craniectomy (VC) procedure where full skulls are used to simulate DC patients by randomly removing bone flaps from specific areas. We used the VC to train various CNN architectures which follow alternative strategies: reconstructing only the missing flap or reconstructing the full skull and then subtracting. In the following, we describe in detail every stage of the reconstruction method.

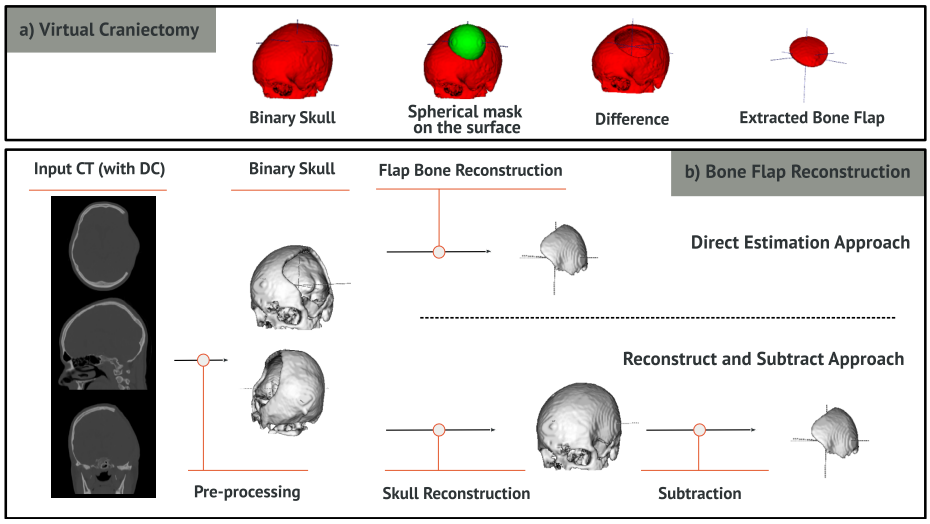


Fig. 1: a) Virtual Craniectomy process: given a skull, a spherical mask is applied in the surface for extracting a bone flap. b) In the direct estimation (DE) strategy, from the binary skull mask with DC, the bone flap is predicted by the network. In the reconstruct and subtract strategy (RS), the full skull is first reconstructed. Then, the binary mask with DC is subtracted from the complete skull, and the difference map is used as bone flap estimation.

2.1 Pre-processing

This stage extracts a binary skull mask from a CT and consists of three steps:

1. Registration: the images are registered to an atlas using rigid transformations, bringing all images into the same coordinate system. For registration, we use SimpleElastix [9], a state-of-the-art registration software publicly available. This pre-alignment encourages the model to focus on variations in the morphology of the skull, rather than learning features associated with its orientation and position.
2. Resampling: After registration, images are resampled to isotropic resolution (2mm).
3. Thresholding: In CT scans, global thresholding [1] can be employed to extract the bones due to their high values in terms of Hounsfield units [15]. We used a threshold value of 90HU. As we can observe in Figure 1b) a binary mask of the skull is obtained after pre-processing.

2.2 Virtual Craniectomy

We designed a virtual craniectomy procedure to simulate the effect of DC on full skulls. This enables the use of head CTs with the complete skull to self-supervise

the learning process, avoiding the need of manually annotated DC images where the flap is segmented. This process implies extracting the intersection of the input skull with a spherical-shaped binary mask, which can be located in its upper part and have a variable size, and use such intersection as the ground truth during training.

We remove skull flaps from random locations, excluding the zone corresponding to the lower part (containing the bones between the jaw and the spine), where a craniectomy would not occur. The radius of the sphere was established so that the volume of the extracted bone flaps would match with standard surgeries. We defined a radius between 5 and 53 voxels to simulate craniectomies of 0.7 to 350 cm^3 of flap volume. This process is depicted in Figure 1a).

2.3 Network Architectures

We implemented alternative encoder-decoder CNN architectures to address the flap reconstruction problem which are based on fully convolutional neural networks, but follow different reconstruction strategies, illustrated in Figure 1b). Note that our contributions are not related to novel CNN architectures (we employ standard autoencoders and U-Net), but regarding the VC-based self-supervised strategy and its application to a new problem (i.e. skull reconstruction) where deep learning approaches have not been explored to date.

a) Reconstruct and subtract with autoencoder (RS-AE): The first model is a fully convolutional autoencoder (AE) trained to reconstruct the complete version of a DC skull (see the Supplementary Material for a detailed description of the AE architecture). Following an approach similar to that of Larrazabal et al. [8,7], we employ a denoising AE where the training process does not only include noise for data augmentation, but also virtual craniectomies. During training, we employ only full skulls: a random VC is applied before the skull enters the AE, which is trained to output its full version. Similar to previous strategies initially developed for unsupervised lesion detection [12], at test time, we reconstruct the bone flap by subtracting the original DC and its reconstructed full version to generate a difference map. The difference map constitutes the final bone flap 3D estimation, from which we can compute features like volume, etc.

b) Direct estimation with U-Net (DE-UNET): The second model directly estimates the bone flap, avoiding the full skull reconstruction and subtraction steps, which may introduce errors in the process. We employ the same encoder-decoder architecture used for the AE, but including skip connections, resulting in a 3D version of the standard U-Net architecture [13] (a detailed description of the architecture is given in the Supplementary Material). For training, instead of aiming to reconstruct the full skull, we learn to reconstruct the bone flap removed during the VC. Note that, similar to the previous model, we only require full skulls for training, enabling self-supervised learning without bone flap annotations.

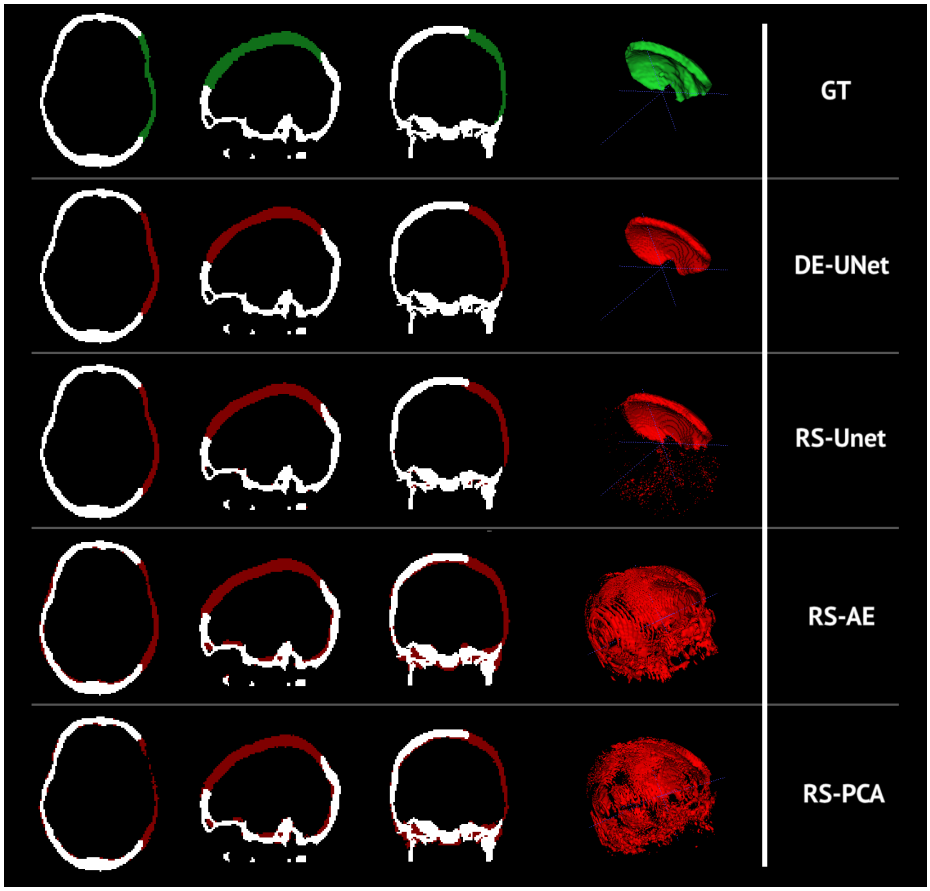


Fig. 2: Bone flap reconstruction (in red) obtained with the approaches compared in this work for a real decompressive craniectomy case from our test dataset.

c) Reconstruct and subtract with U-Net (RS-UNET): For completeness, we also explore the use of the U-Net following the reconstruct-and-subtract strategy, to evaluate the impact of the skip-connections in the resulting reconstruction.

2.4 Training and Implementation

The CNN architectures were implemented in PyTorch 1.4 and trained on an NVIDIA TITAN Xp GPU [4]. During training, the images are fed to the network by adding salt and pepper noise and performing VC (with probability of 0.8), allowing the networks to see both intact and VC skulls. For all models the loss

[4] The source code of our project is publicly available at: `http://gitlab.com/matzkin/deep-brain-extractor`

function L consists in a combination of the Dice loss L_{Dice} and the Binary Cross Entropy (BCE) Loss L_{BCE}. While cross-entropy loss optimizes for pixel level accuracy, the Dice loss function enhances the segmentation quality [11]. The compound loss function is defined as:

$$L = L_{Dice} + \lambda L_{BCE} \tag{1}$$

where the parameter $\lambda = 1$ was chosen by grid search. To improve generalization we incorporated dropout layers and use early stopping on validation data.

3 Experiments and Discussion

3.1 Database

The images used for this work were provided by the University of Cambridge (Division of Anaesthesia, Department of Medicine). They consist in 98 head CT images of 27 patients with Traumatic Brain Injury (TBI), including 31 images with DC and 67 cases with full skull. For training, we used full skull images only, excluding those patients who also have associated an image with DC. Patients which include pre and post-operative CT images were used for testing, since the difference between both images after registration was employed as ground-truth for the evaluation of bone flap estimation (an example is shown in green in Figure 2). In this context, we employed 52 images for training (corresponding to 17 different patients) and 10 for testing (since there are only 10 patients with pre and post DC studies). The 36 images not included in the study were either pre-operative images of patients from the test split or post-operative without their corresponding pre-operative.

3.2 Baseline models

We implemented a baseline model based on principal component analysis (PCA) for the task of flap bone estimation which follows the reconstruct and subtract strategy (RS-PCA). The principal components (see the Supplementary Material for visualizations of these components) were obtained by applying PCA to the vectorized version of the pre-processed complete skulls from the training fold. Similar to the RS-AE approach, the learnt latent representation provides a base for the space of complete skulls. Therefore, for reconstruction, we take the incomplete skull and project it to the learnt space to obtain its full version.

For the task of flap bone volume estimation, we also compared our methods with the manual state-of-the-art ABC approach [17]. The ABC method requires to annotate manual measurements on the DC images (see the Supplementary Material for an example) and estimates the flap volume following simple geometric rules (a complete description of ABC can be found in the original publication [17]).

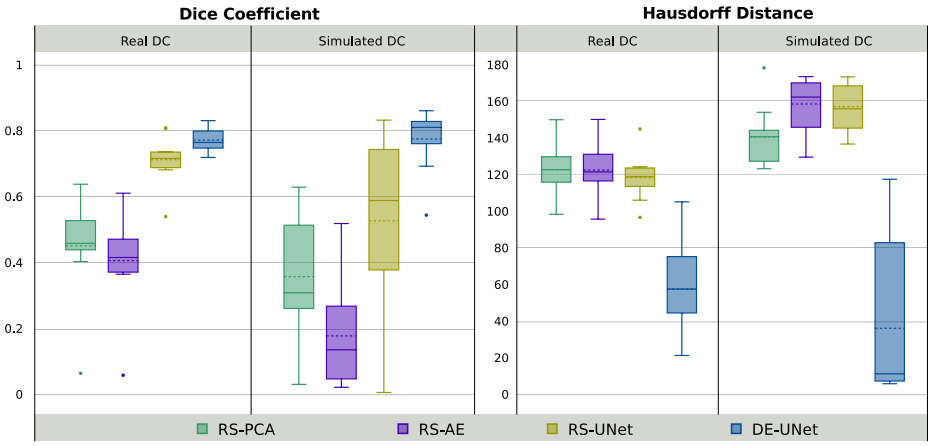


Fig. 3: Dice Coefficient and Hausdorff Distance (in mm) of the proposed methods output compared with the ground-truth (dashed line indicates the mean value). It can be seen that the DE-UNet outperforms the other discussed methods.

3.3 Experiments and results

We performed experiments for bone flap reconstruction and volume estimation in real and simulated craniectomies. The simulations were done by performing 100 random virtual craniectomes to every complete skull from the test fold, resulting in a total of 1000 simulations for test. Figure 2 provides a qualitative comparison of the reconstructions (in red) obtained using the different approaches in a real DC. It can be observed that those based on the reconstruct and subtract strategy using AE and PCA produce spurious segmentations in areas far from the flap. The best reconstructions are achieved using the DE-UNet and RS-UNet, highlighting the importance of the skip connections.

The quantitative analysis is summarized in Figures 3 and 4. Figure 3 shows Dice coefficient and Hausdorff distance between the ground-truth and reconstructed bone flaps for all the methods in real and simulated scenarios. Figure 4 includes scatter plots showing the accuracy of the bone flap volume estimation: we compare the predicted volume (x-axis) with the expected volume (y-axis). The closer the points to the identity, the more accurate are the predictions. For volume estimation we also include the manual ABC method. From these results, we observe that DE-UNet outperforms the other methods in both tasks, producing even better volume estimations than the manual ABC approach. We observed that Reconstruct and Subtract methods usually generate spurious pixels as prediction (as can be seen in Figure 2) and a post-processing step may be needed after subtracting the pre and post-operative images (e.g. taking the biggest connected component, or applying morphological operations in the prediction). This does not tend to happen with Direct Estimation, what explains the gain in performance.

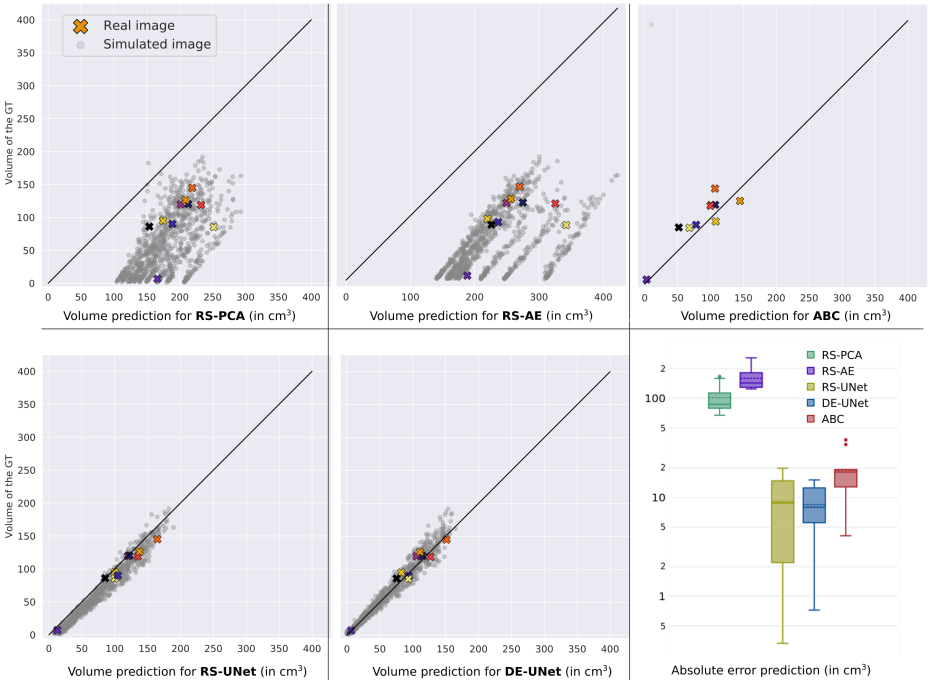


Fig. 4: Quantitative comparison for bone flap volume estimation with the different methods implemented in this study. The scatter plots show the estimated (x-axis) vs ground-truth bone flap volume. Note that RS-PCA, RS-AE, RS-UNet and DE-UNet show results for both real (cross markers in color) and simulated cases (circles in grey). For ABC, we only show results in real cases since the actual CT image is required for manual annotation (and virtual craniectomies for simulations are perform directly on the binary skull mask).

4 Conclusions

In this work, we propose and compare alternative self-supervised methods to estimate the missing bone flap on post-operative CTs with decompressive craniectomy. To our knowledge, this is the first study that tackles skull reconstruction and bone flap estimation using deep learning. We introduced a self-supervised training strategy which employs virtual craniectomy to generate training data from complete skulls, avoiding the need for annotated DC images.

We studied two different reconstruction strategies: direct estimation (DE) and reconstruct and subtract (RS). We found that DE outperforms RS strategies, since the last ones tend to generate spurious segmentations in areas far from the missing bone flap. The proposed methods were also compared with a PCA-based implementation of the RS reconstruction process and a state-of-the-art method (ABC) used in the clinical practise which requires manual measurements and

relies on a geometric approximation. The proposed direct estimation method based on the U-Net architecture (DE-UNet) outperforms all the other strategies.

The performance of our method was measured in real cases (TBI patients who underwent decompressive craniectomy) as well as simulated scenarios. In the future, we plan to explore the use of the bone flap features to improve patient treatment. In this sense, we are interested in studying specific features in terms of volume and shape of a craniectomy that leads to fewer complications and improves patient outcome after TBI.

Acknowledgments. The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research, and the support of UNL (CAID-PIC-50220140100084LI) and ANPCyT (PICT 2018-03907).

References

1. van Eijnatten, M., van Dijk, R., Dobbe, J., Streekstra, G., Koivisto, J., Wolff, J.: CT image segmentation methods for bone used in medical additive manufacturing. Medical Engineering & Physics 51, 6–16 (Jan 2018), https://doi.org/10.1016/j.medengphy.2017.10.008
2. Freyschlag, C.F., Gruber, R., Bauer, M., Grams, A.E., Thomé, C.: Routine postoperative computed tomography is not helpful after elective craniotomy. World Neurosurgery (2018), http://www.sciencedirect.com/science/article/pii/S1878875018326299
3. Galgano, M., Toshkezi, G., Qiu, X., Russell, T., Chin, L., Zhao, L.R.: Traumatic brain injury: Current treatment strategies and future endeavors. Cell Transplantation 26(7), 1118–1130 (2017), https://doi.org/10.1177/0963689717714102, pMID: 28933211
4. Herteleer, M., Ectors, N., Duflou, J., Calenbergh, F.V.: Complications of skull reconstruction after decompressive craniectomy. Acta Chirurgica Belgica 117(3), 149–156 (Dec 2016), https://doi.org/10.1080/00015458.2016.1264730
5. Hieu, L., Bohez, E., Sloten, J.V., Phien, H., Vatcharaporn, E., Binh, P., An, P., Oris, P.: Design for medical rapid prototyping of cranioplasty implants. Rapid Prototyping Journal 9(3), 175–186 (Aug 2003), https://doi.org/10.1108/13552540310477481
6. Huang, K.C., Liao, C.C., Xiao, F., Liu, C.C.H., Chiang, I.J., Wong, J.M.: Automated volumetry of postoperative skull defect on brain CT. Biomedical Engineering: Applications, Basis and Communications 25(03), 1350033 (May 2013), https://doi.org/10.4015/s1016237213500336
7. Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: Anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Transactions on Medical Imaging (2020)
8. Larrazabal, A.J., Martinez, C., Ferrante, E.: Anatomical priors for image segmentation via post-processing with denoising autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–593. Springer (2019)
9. Marstal, K., Berendsen, F., Staring, M., Klein, S.: Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 134–142 (2016)

10. Moon, J.W., Hyun, D.K.: Decompressive craniectomy in traumatic brain injury: A review article. Korean Journal of Neurotrauma 13(1), 1 (2017), https://doi.org/10.13004/kjnt.2017.13.1.1
11. Patravali, J., Jain, S., Chilamkurthy, S.: 2d-3d fully convolutional neural networks for cardiac mr segmentation. In: Pop, M., Sermesant, M., Jodoin, P.M., Lalande, A., Zhuang, X., Yang, G., Young, A., Bernard, O. (eds.) Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. pp. 130–139. Springer International Publishing, Cham (2018)
12. Pawlowski, N., Lee, M.C., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., et al.: Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders (2018)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
14. Sedney, C., Julien, T., Manon, J., Wilson, A.: The effect of craniectomy size on mortality, outcome, and complications after decompressive craniectomy at a rural trauma center. Journal of Neurosciences in Rural Practice 5(3), 212 (2014), https://doi.org/10.4103/0976-3147.133555
15. Seeram, E.: Computed Tomography - E-Book: Physical Principles, Clinical Applications, and Quality Control. Elsevier Health Sciences (2015), https://books.google.com.ar/books?id=DTCDCgAAQBAJ
16. Tanrikulu, L., Oez-Tanrikulu, A., Weiss, C., Scholz, T., Schiefer, J., Clusmann, H., Schubert, G.: The bigger, the better? about the size of decompressive hemicraniectomies. Clinical Neurology and Neurosurgery 135, 15–21 (Aug 2015), https://doi.org/10.1016/j.clineuro.2015.04.019
17. Xiao, F., Chiang, I.J., Hsieh, T.M.H., Huang, K.C., Tsai, Y.H., Wong, J.M., Ting, H.W., Liao, C.C.: Estimating postoperative skull defect volume from ct images using the abc method. Clinical Neurology and Neurosurgery 114(3), 205 – 210 (2012), http://www.sciencedirect.com/science/article/pii/S0303846711003076

Cranial implant design via virtual craniectomy with shape priors

Cranial Implant Design via Virtual Craniectomy with Shape Priors

Franco Matzkin[1], Virginia Newcombe[2], Ben Glocker[3], and Enzo Ferrante[1]

[1] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET, FICH-UNL (Argentina)
[2] Division of Anaesthesia, Department of Medicine, University of Cambridge (UK)
[3] BioMedIA, Imperial College London (UK)

Abstract. Cranial implant design is a challenging task, whose accuracy is crucial in the context of cranioplasty procedures. This task is usually performed manually by experts using computer-assisted design software. In this work, we propose and evaluate alternative automatic deep learning models for cranial implant reconstruction from CT images. The models are trained and evaluated using the database released by the AutoImplant challenge, and compared to a baseline implemented by the organizers. We employ a simulated virtual craniectomy to train our models using complete skulls, and compare two different approaches trained with this procedure. The first one is a direct estimation method based on the UNet architecture. The second method incorporates shape priors to increase the robustness when dealing with out-of-distribution implant shapes. Our direct estimation method outperforms the baselines provided by the organizers, while the model with shape priors shows superior performance when dealing with out-of-distribution cases. Overall, our methods show promising results in the difficult task of cranial implant design.

Keywords: Skull reconstruction · self-supervised learning · decompressive craniectomy · shape priors

1 Introduction

Crainoplasty is a surgical procedure aimed at repairing a skull vault defect by insertion of a bone or nonbiological implant (e.g. metal or plastic) [1]. Such skull defect may exist due to different reasons, like a brain tumor removal procedure or a decompressive craniectomy surgery following a traumatic brain injury [12]. Cranial implant design is usually performed by experts using computer-aided design software specifically tailored for this task [2]. The AutoImplant challenge, organized for the first time at MICCAI 2020, aims at bench-marking the latest developments in computational methods for cranial implant reconstruction. In this work, we propose and evaluate two approaches to solve this task using deep learning models.

Previous works on skull and cranial implant reconstruction suggest that deep learning models are good candidates to solve this task. In [13] a denoising autoencoder was used to perform skull reconstruction, following an approach similar to the recently proposed Post-DAE method [7, 6]. In this case, a denoising autoencoder is trained to reconstruct full skulls from corrupted versions. However, the model proposed in [13] works with skulls extracted from magnetic resonance images, can only handle low resolution images and was evaluated on the full-skull reconstruction task. Here we focus on reconstructing the flap only, on skulls extracted from high resolution and anisotropic computed tomography (CT) images. Other approaches rely on a head symmetry assumption and propose to take advantage of it to reconstruct the missing parts by mirroring the complete side of the skull [4]. However, this is not a realistic assumption since missing flaps may occur in both sides simultaneously. Another alternative could be the subtraction of the aligned pre- and post-operative CT scans. Unfortunately, this requires to have access to the pre-operative image, which may not be the case in real clinical scenarios.

Recently, we have proposed [11] a simple virtual craniectomy procedure which enables training different deep learning models in a self-supervised way, given a dataset composed of full skulls. In this work, we compared two different approaches: direct estimation of the implant, or reconstruct-and-subtract (RS) strategies where the full skull is first reconstructed, and then the original image is subtracted from it to generate a difference map. We evaluated different architectures and concluded that direct estimation produces more accurate estimates than RS strategies, since the latter one tends to generate noise in areas far from the flap. A different approach has been introduced by the AutoImplant challenge organizers [9] which also employs deep learning models, but it works in two steps. First, a low resolution version of the image is reconstructed to localize the area where the defected region is located. Then, they extract a 3D patch from the high resolution image and process it using a second neural network trained for fine implant prediction.

In this work, based on the conclusions from [11], we employ a direct estimation method that operates on full skulls which are rigidly registered to an atlas and resampled to an intermediate resolution. Aligning the images allow us to work in a common space which simplifies the reconstruction task. We adapt the virtual craniectomy procedure to account for more realistic flap shapes, similar to the ones introduced in the AutoImplant challenge. Moreover, we propose to incorporate anatomical priors into the standard direct estimation model introduced in [11] by feeding the registered skull atlas as an extra image channel. Previous works [8] have shown that incorporating approximate shape priors as additional image channels is a simple yet effective way to increase the anatomical plausibility of the segmentations, since it provides supplementary context information to the network. We compare the results of our two methods with those obtained by the baseline benchmark model introduced in [9], showing the superiority of our approach.

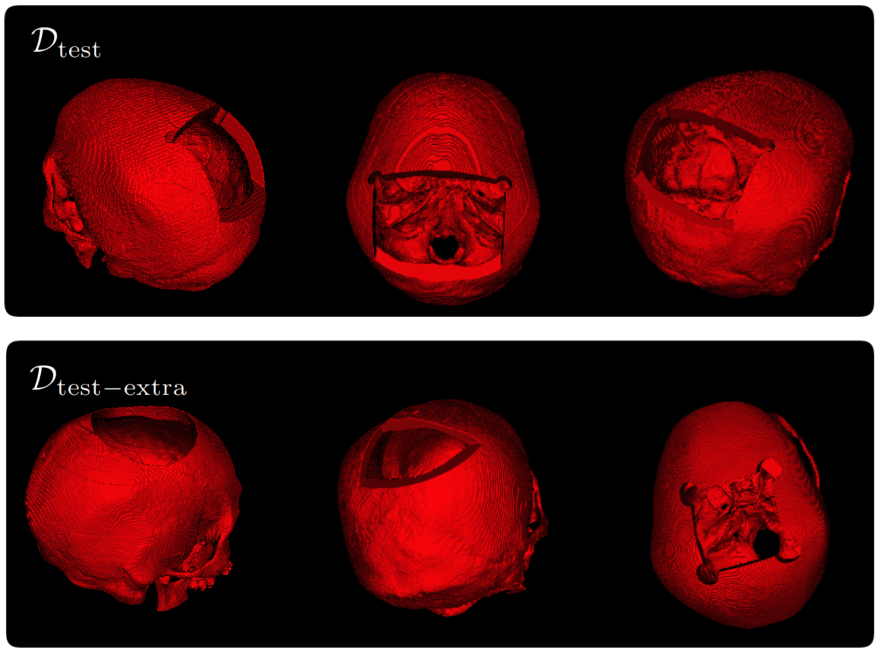


Fig. 1. Examples of images from the $\mathcal{D}_{\text{test}}$ set and $\mathcal{D}_{\text{test}-\text{extra}}$ (out-of-distribution cases). As it can be observed, images from $\mathcal{D}_{\text{test}}$ follow a common pattern, while those in $\mathcal{D}_{\text{test}-\text{extra}}$ present different defects with various shapes.

2 Challenge description and database

The AutoImplant challenge organizers provided 100 images for training ($\mathcal{D}_{\text{train}}$) and 110 images for testing. From the 110 test images, 100 of them (denoted here as $\mathcal{D}_{\text{test}}$) have simulated surgical defects which follow the same distribution as the ones on the training images, while the remaining 10 (denoted as $\mathcal{D}_{\text{test}-\text{extra}}$) have defects which do not follow the same distribution (see Figure 1). The images were selected from the CQ500 public database[4] [3]. They have fixed image dimension in the axial plane (512 x 512) and a variable number of axial slices Z.

The training dataset ($\mathcal{D}_{\text{train}}$) is composed of triplets $(\mathcal{X}^{\text{full}}, \mathcal{X}^{\text{defected}}, \mathcal{Y})$, where $\mathcal{X}^{\text{full}}$ is the full skull, $\mathcal{X}^{\text{defected}}$ corresponds to the defected skull and $\mathcal{Y}$ to the removed defect that we aim at reconstructing. For the test images, only the $\mathcal{X}^{\text{defected}}$ images were released. We evaluated the proposed methods in the test images and submitted the results to the organizers, who computed the metrics reported in this paper. It is important to note that, in order to avoid overfitting to the test data, we could submit our results a maximum of 5 times.

[4] The database can be accessed at: http://headctstudy.qure.ai/dataset

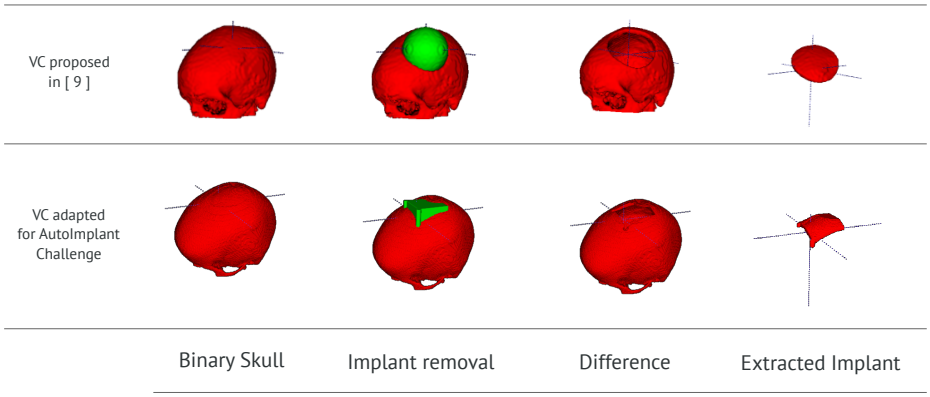


Fig. 2. Modified virtual craniectomy procedure. We incorporated new template shapes for the virtual craniectomy to account for the pattern found in the AutoImplant challenge dataset.

3 Methods

The proposed cranial implant reconstruction methods operate on the space of binary volumetric masks. Such binary skull can be obtained by simply thresholding a brain CT image according to the Hounsfield scale, or applying more sophisticated methods. In the AutoImplant challenge, the skulls were already provided as binary volumes, extracted from the CT images using thresholding and additional post-processing steps (for further details we refer to [9]). Since the training data includes the full skulls, we leveraged the virtual craniectomy procedure proposed in [11] to train our models.

3.1 Virtual Craniectomy and Data Augmentation

Given a full skull, we designed a virtual craniectomy procedure which consists in removing a bone flap using a template located in a random position along its upper part. In [11], spherical template shapes were used. By visual inspection of the AutoImplant training data, we observed that defects tend to follow a pattern given by the intersection of the skull with a cube with two cylinders over the edges perpendicular to the axial planes. So, we designed a variable-size template shape which produces similar defects, as shown in Figure 2. To increase the diversity of our training procedure, we also included spherical and cubic templates of random sizes (all of the three shapes were selected with equal probability).

The virtual craniectomy was used as a data augmentation mechanism to generate a variety of training samples from a limited amount of full skulls, resulting in a self-supervised learning approach, where no annotated skull defects are required for training. We also included salt and pepper noise in the input images with probability 0.01. Moreover, we also considered the defective skulls provided

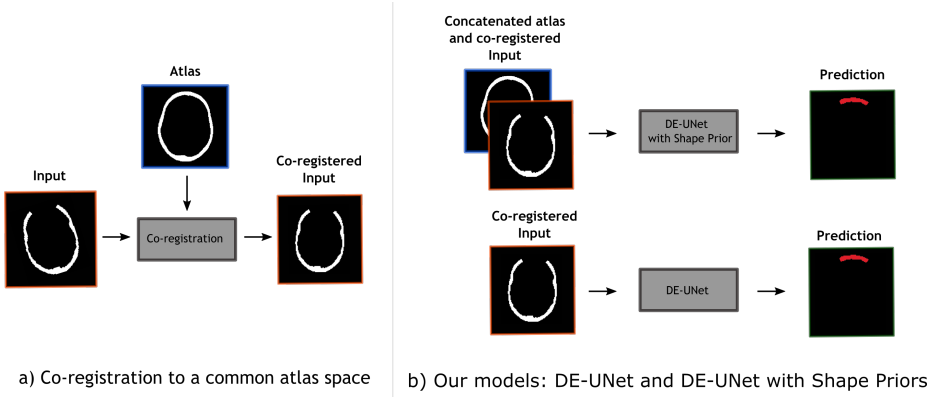


Fig. 3. (a) The images are first registered to an atlas space, and resampled to a common resolution. We store the resulting transform $\mathcal{T}$ and its inverse $\mathcal{T}^{-1}$. **(b)** We compare two different approaches for the implant reconstruction task. The first one is a standard DE-UNet model. The second one incorporates a shape prior by considering the atlas as an extra input channel to the network. After prediction, the segmentation mask is mapped-back to the original image space using the inverse transform $\mathcal{T}^{-1}$.

by the organizers as part of our datasets (in these cases, virtual craniectomy was not performed). During training, we sampled images coming from both sources: simulated virtual craniectomies and defective skulls provided by the organizers.

3.2 Common space alignment

Before training, all the images were rigidly registered to a common space determined by a full skull atlas. It consists in a thresholded version of a full-skull head CT atlas constructed by averaging several healthy head CT images. Such atlas allowed us to normalize the images by resampling them to an intermediate resolution. We chose this resolution to be 0.695 x 0.695 x 0.715 mm (resulting in a volume of 304 x 304 x 224 voxels) because it was the maximum size we managed to fit in GPU memory. Moreover, aligning the images in a common space simplifies the reconstruction task for the neural network, since it can focus on shape variations which are more relevant to the reconstruction task than translations and rotations. We used the FLIRT software package [5] for rigid registration. At test time, given a test defective image $\mathcal{X}_i^{\text{defected}}$, we apply the same registration procedure which returns a transformation $\mathcal{T}$ and its inverse $\mathcal{T}^{-1}$. The transformation is applied to the original image $\mathcal{T} \circ \mathcal{X}_i^{\text{defected}}$. The estimated skull defect $\hat{\mathcal{Y}}_i$ is reconstructed in the common space, and the final estimate in the original space is recovered by applying the inverse transformation $\mathcal{T}^{-1} \circ \hat{\mathcal{Y}}_i$.

3.3 Direct estimation

Our first method is a direct estimation model which follows the same architecture as the DE-UNet used in [11]. It is a standard 3D UNet encoder-decoder archi-

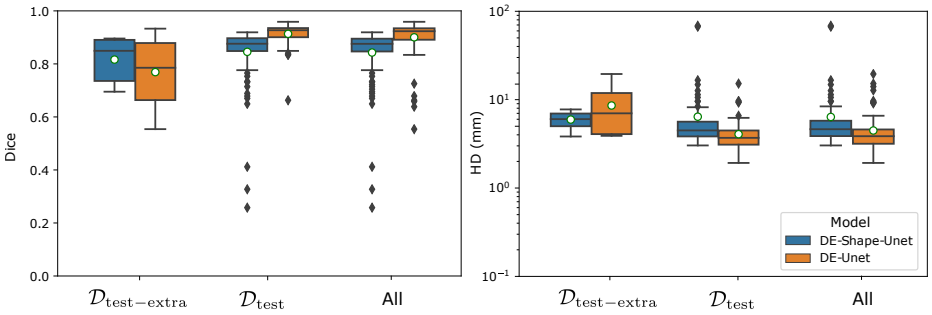


Fig. 4. Comparison of the results for the proposed methods in terms of Dice and Hausdorff Distance (HD). HD is shown in log scale for better visualization.

tecture with skip connections, trained using a compound loss which combines Dice and cross-entropy terms [14] (for more details, we refer to our work in [11]). After reconstruction, the segmentation is re-mapped to the original resolution using the inverse transform $\mathcal{T}^{-1}$ as previously discussed.

The model is trained using batches with full volume images, pre-aligned in the common space and resampled to an intermediate resolution as previously discussed.

3.4 Direct estimation with shape priors

Since the DE-UNet model is a fully convolutional architecture, the receptive field of the model is mainly determined by the amount of layers and parameters of the pooling and convolution operations. In other words, the local support of the output predictions is restricted to a certain area in the input image. When we have to reconstruct big or out-of-distribution skull defects, it may happen that most of the image support for certain parts of it are background, so the network may have no context to infer the implant shape. To overcome this limitation and make our model robust, we propose to incorporate context via shape priors given as an extra channel to the segmentation network. Previous works [8] have shown that this simple extension can boost the robustness of existing state-of-the-art pixel-wise approaches in medical image segmentation tasks.

We take advantage of the fact that images are co-registered to a common space, and use the same skull atlas as shape prior. After registration, we concatenate the resampled image with the atlas as a extra input channel, and train the network following the same strategy discussed before. In this case, the shape prior acts as a kind of initialization for the network's output, providing additional context that will be useful specially to reconstruct out-of-distribution defects. We refer to this model as DE-Shape-UNet.

Table 1. Quantitative results obtained for the two proposed methods (DE-UNet and DE-Shape-UNet) compared with the two baselines reported by the challenge organizers in [9]. We report the mean Dice and HD values, and the standard deviation in parentheses.

Method	$\mathcal{D}_{\text{test}}$ (100)		$\mathcal{D}_{\text{test-extra}}$ (10)		Overall	
	Dice	HD (mm)	Dice	HD (mm)	Dice	HD (mm)
Baseline N1 [9]	0.809	5.440	-	-	-	-
Baseline N2 [9]	0.855	5.182	-	-	-	-
DE-UNet	**0.913 (0.038)**	**4.067 (1.762)**	0.769 (0.126)	8.585 (5.128)	**0.900 (0.067)**	**4.477 (2.626)**
DE-Shape-UNet	0.845 (0.107)	6.414 (9.060)	**0.816 (0.078)**	**5.952 (1.258)**	0.842 (0.105)	6.372 (8.648)

3.5 Implementation details

The models were implemented in Pyhton, using the PyTorch 1.4 library. We trained and evaluated the CNNs using an NVIDIA TITAN Xp GPU with 12GB of RAM. The same virtual craniectomy and data augmentation procedure was used to train both models. In both cases we used a compound loss function which combines Dice loss and Binary Cross Entropy (BCE) as $L = L_{Dice} + \lambda L_{BCE}$ (parameter λ was set to $\lambda = 1$ by grid search). Both models followed the DE-UNet architecture described in [11]; the only difference between them was that we concatenated the atlas as an extra input channel in the DE-Shape-UNet model. For optimization, we used Adam with initial learning rate of 1e-4. The batch-size was set to 1 for memory restrictions. The models were trained for 50 epochs. The 100 training images were split in 95 images for training and 5 for validation. After 50 epochs, we kept the model that achieved best accuracy in the validation fold.

4 Results

Figure 4 and Table 1 include a quantitative comparison of the results. We report Dice coefficient and Hausdorff distance measured in the $\mathcal{D}_{\text{test}}$ (100 images), $\mathcal{D}_{\text{test-extra}}$ (10 images) and the whole test dataset. We observe that DE-Shape-UNet presents better performance for out-of-distribution cases ($\mathcal{D}_{\text{test-extra}}$), while DE-UNet outperforms the other model in the $\mathcal{D}_{\text{test}}$ set. Since the whole test dataset is composed of 100 images from $\mathcal{D}_{\text{test}}$ and only 10 images from $\mathcal{D}_{\text{test-extra}}$, the DE-UNet model shows better performance in the overall comparison. Moreover, DE-UNet model outperforms the two baseline models (N1 and N2) reported by the organizers in [9]. Figure 5 provides some visual examples for reconstructions obtained with both methods in samples from $\mathcal{D}_{\text{test}}$ and $\mathcal{D}_{\text{test-extra}}$.

5 Conclusions

In this work, we evaluated two different approaches for cranial implant reconstruction based on deep learning: a direct estimation method and an alterna-

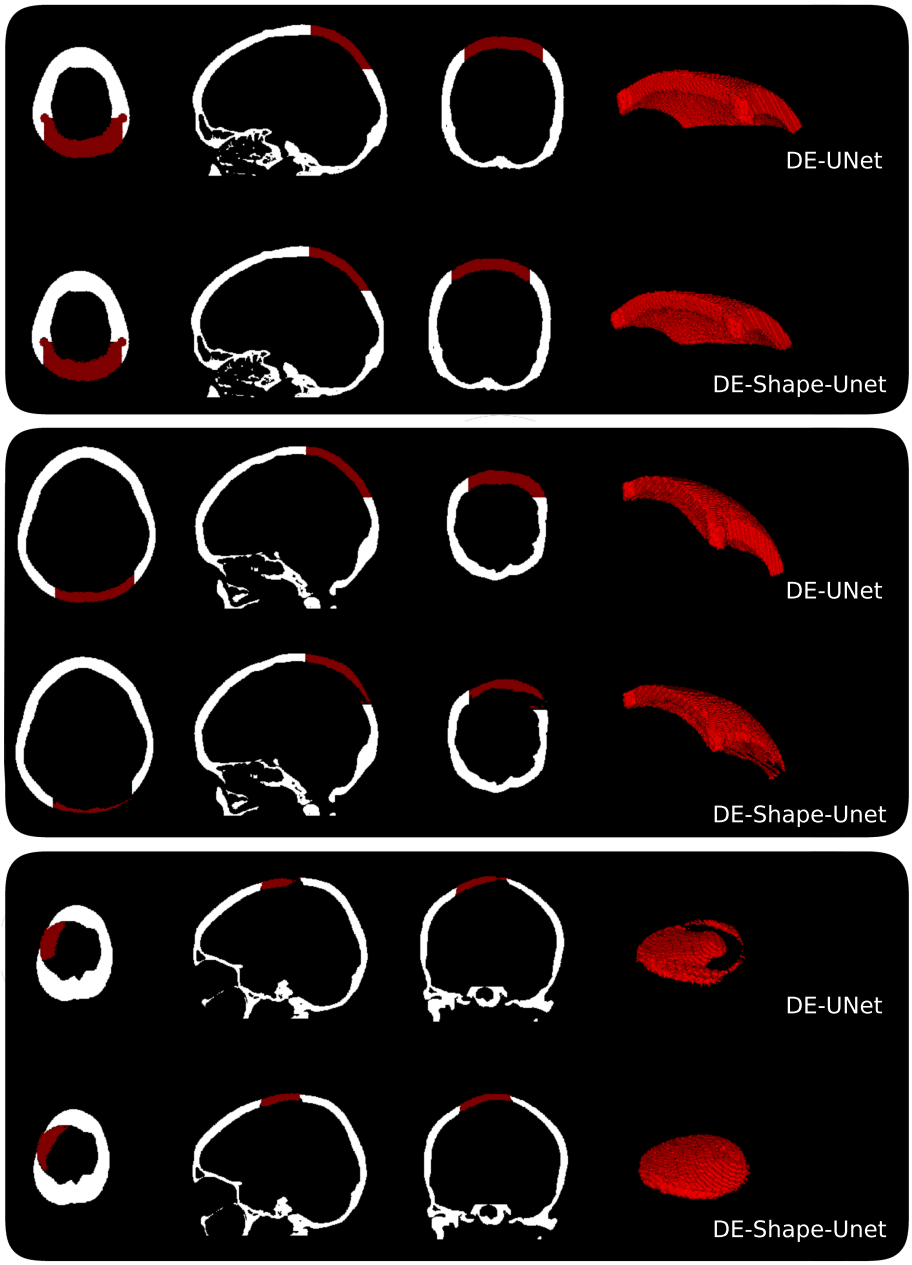


Fig. 5. Examples of different reconstructions from $\mathcal{D}_{\text{test}}$ (cases which follow the same pattern than the training dataset, shown in rows 1 and 2) and $\mathcal{D}_{\text{test}-\text{extra}}$ (out-of-distribution case, shown in row 3). As we can observe, both methods performed well in the image depicted in row 1. For the case in the 2nd row, even if the DE-Shape-UNet model managed to reconstruct the implant, the quality of the reconstruction is lower than that of the DE-UNet. The opposite happened with the image in row 3 (an out-of-distribution case from $\mathcal{D}_{\text{test}-\text{extra}}$) where the model which incorporated shape priors managed to reconstruct the implant, while the DE-UNet failed in this task.

tive strategy which incorporates shape priors. We adapted the virtual craniectomy procedure proposed in [11] to the defect distribution of the AutoImplant challenge. We found that the simple DE-UNet method produces more accurate results for the skull defects which follow the same distribution as those in the training dataset. However, for out-of-distribution cases where the DE-UNet model tends to fail, the use of shape priors increases the robustness of the model, providing additional context to the network. In our implementation, this gain in robustness for out-of-distribution cases was achieved to the detriment of the overall accuracy. In future work, we plan to study alternative ways to introduce shape priors, e.g. considering deformable registration with anatomical constraints [10] to the atlas space instead of rigid transformations, or incorporating shape priors in a co-registration and segmentation process [15].

6 Acknowledgments

The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research, and the support of UNL (CAID-PIC-50220140100084LI) and ANPCyT (PICT 2018-03907).

References

1. Andrabi, S.M., Sarmast, A.H., Kirmani, A.R., Bhat, A.R.: Cranioplasty: Indications, procedures, and outcome–an institutional experience. Surgical neurology international **8** (2017)
2. Chen, X., Xu, L., Li, X., Egger, J.: Computer-aided implant design for the restoration of cranial defects. Scientific reports **7**(1), 1–10 (2017)
3. Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., Warier, P.: Development and validation of deep learning algorithms for detection of critical findings in head ct scans. arXiv preprint arXiv:1803.05854 (2018)
4. Hieu, L., Bohez, E., Sloten, J.V., Phien, H., Vatcharaporn, E., Binh, P., An, P., Oris, P.: Design for medical rapid prototyping of cranioplasty implants. Rapid Prototyping Journal **9**(3), 175–186 (Aug 2003). https://doi.org/10.1108/13552540310477481, https://doi.org/10.1108/13552540310477481
5. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage **17**(2), 825–841 (2002)
6. Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: Anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE Transactions on Medical Imaging (2020). https://doi.org/10.1109/TMI.2020.3005297
7. Larrazabal, A.J., Martinez, C., Ferrante, E.: Anatomical priors for image segmentation via post-processing with denoising autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–593. Springer (2019)

8. Lee, M.C.H., Petersen, K., Pawlowski, N., Glocker, B., Schaap, M.: Tetris: template transformer networks for image segmentation with shape priors. IEEE transactions on medical imaging **38**(11), 2596–2606 (2019)
9. Li, J., Pepe, A., Gsaxner, C., von Campe, G., Egger, J.: A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. arXiv preprint arXiv:2006.12449 (2020)
10. Mansilla, L., Milone, D.H., Ferrante, E.: Learning deformable registration of medical images with anatomical constraints. Neural Networks **124**, 269–279 (2020)
11. Matzkin, F., Newcombe, V., Stevenson, S., Khetani, A., Newman, T., Digby, R., Stevens, A., Glocker, B., Ferrante, E.: Self-supervised skull reconstruction in brain ct images with decompressive craniectomy. MICCAI 2020 (2020)
12. Monteiro, M., Newcombe, V.F., Mathieu, F., Adatia, K., Kamnitsas, K., Ferrante, E., Das, T., Whitehouse, D., Rueckert, D., Menon, D.K., et al.: Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. The Lancet Digital Health (2020)
13. Morais, A., Egger, J., Alves, V.: Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder. In: World Conference on Information Systems and Technologies. pp. 151–160. Springer (2019)
14. Patravali, J., Jain, S., Chilamkurthy, S.: 2d-3d fully convolutional neural networks for cardiac mr segmentation. In: Pop, M., Sermesant, M., Jodoin, P.M., Lalande, A., Zhuang, X., Yang, G., Young, A., Bernard, O. (eds.) Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. pp. 130–139. Springer International Publishing, Cham (2018)
15. Shakeri, M., Ferrante, E., Tsogkas, S., Lippe, S., Kadoury, S., Kokkinos, I., Paragios, N.: Prior-based coregistration and cosegmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 529–537. Springer (2016)

AutoImplant 2020-First MICCAI Challenge on Automatic Cranial Implant Design

AutoImplant 2020-First MICCAI Challenge on Automatic Cranial Implant Design

Jianning Li, Pedro Pimentel, Angelika Szengel, Moritz Ehlke, Hans Lamecker, Stefan Zachow, Franco Matzkin, Enzo Ferrante, and et al.

Institution details omitted for brevity

Abstract. The aim of this paper is to provide a comprehensive overview of the MICCAI 2020 AutoImplant Challenge. The approaches and publications submitted and accepted within the challenge will be summarized and reported, highlighting common algorithmic trends and algorithmic diversity. Furthermore, the evaluation results will be presented, compared and discussed in regard to the challenge aim: seeking for low cost, fast and fully automated solutions for cranial implant design. Based on feedback from collaborating neurosurgeons, this paper concludes by stating open issues and post-challenge requirements for intra-operative use. The codes can be found at https://github.com/Jianningli/tmi.

I Introduction

Cranioplasty is a reconstructive surgery to repair skull damages resulting from brain tumor surgeries or head trauma, where a part of the skull bone (mainly in the neurocranium area) has to be removed. Increased use of decompressive craniectomies resulted in more reconstructions of cranial defects in the past 15 years, around 25 patients per one million inhabitants per year for Europe, the Middle East, and Africa [1, 2]. However, complications, like brain swelling and infections after decompressive craniectomies and cranioplasties, are frequent and can even be life-threatening events [3]. A systematic review revealed that one in 10 patients undergoing a decompressive craniectomy suffers a complication, which makes an additional medical or surgical intervention necessary [4]. Hence, a tailor-made patient-specific implant (PSI) of the cranium is needed in such surgery to optimally restore the protective, mechanical, and anatomical functions of the human skull [5].

The design of a PSI remains a bottleneck [6] for cranioplasty, since the reconstructive surgery can be performed only after the implant has been designed, manufactured, and delivered to the hospital, which may take weeks or even months. If cranioplasty could be performed immediately after the primary surgery that removes the skull bone, the overall duration of surgery can be reduced substantially. To achieve this goal, a fast, fully automatic, and in-operating-room (in-OR) manufacturing of PSI is required. Additive manufacturing or 3D printing enables fast manufacturing of 3D medical implants directly in the surgery room, given the corresponding 3D models.

Currently, the patient's head is scanned by computed tomography (CT) after primary surgery. The bone structures are extracted from the CT, converted into a 3D model, and used to guide the computer-aided design (CAD) of the implant [7–10]. Symmetry is often assumed in CAD procedures, which use a mirrored copy of the healthy skull side as a template. However, symmetry cannot be used when the skull is deformed or when the defect crosses the symmetry plane.

Inspired by the clinical practice of relying on a post-operative head CT for cranial implant design, the AutoImplant 2020 challenge encouraged the development of automated implant design by providing both pre- and post-operative skulls for supervised training and evaluation. Unlike the clinical practice, which models implants as meshes, the challenge encouraged participants to predict the binary implant masks directly from binary skull images (voxel grids). Ten full papers were accepted by the challenge. They cover a variety of data-driven methods, including classical statistical approaches, such as statistical shape models (SSM) [11], and deep learning approaches, such as generative adversarial networks (GAN) [12], variational auto-encoders (VAE) [13], and variants of U-Net [14], which are novel in neurosurgery. From a technical perspective, the processing of high-dimensional skull data and the generalization to varied skull defects are key considerations for the development and evaluation of the algorithms.

II Related Work

Prior to the challenge, automatic cranial implant design has been an under-researched area, especially concerning data-driven approaches, due to a lack of public datasets suitable for the task. This section summarizes the algorithms published online prior to the conclusion of the challenge, which have been used for automatic reconstruction of medical implants, including cranial implants. A review of general shape completion algorithms will also be covered in this section. An early study casts cranial implant design as a surface interpolation problem, smoothly interpolating the missing surface using radial basis functions [15].

A. Statistical Shape Model

Prior to the challenge, SSM is among the most widely used methods for reconstructing skull bones, including the facial area [16], [17], the cranium area [18] and other bone structures on the skull [19], [20]. A statistical model of the skull $S(w)$ represents the average shape $S \in \mathbb{R}^{3m}$ (where m is the number of vertices of the skull mesh) as well as a set of shape variations $p_i \in \mathbb{R}^{3m}$ of a given skull population:

$$S(w) = \bar{S} + \sum_{i=1}^{N} w_i p_i \tag{1}$$

where w_i is the shape weight of each mode of shape variation p_i, and its value is confined to the scope of the training skull population. Reconstructing a complete skull given a defective skull D is the task of finding the set of weight parameters

w^* such that $S(w^*)$ best matches the shape of D, except in the defective region. The cranial implant can then be obtained by taking the difference (logical XOR) of the reconstructed skull and D. Finding $S(w^*)$ is usually an iterative process.

B. Deep Learning

Recently, deep learning solutions have emerged. Morais et al. [21] were the first to demonstrate a denoising auto-encoder for skull shape completion on very coarse skulls (dimension: 30^3, 60^3, and 120^3) with simple holes. Li et al. [6], [22] extended the concept of skull shape completion to high-resolution data, i.e., $512^2 \times Z$ with much irregular synthetic defects. Their approach showed potential for clinical use according to the evaluation results on real defective skulls from craniotomy. Their dataset is not yet public. Kodym et al. [23] trained a cascade of convolutional neural networks to predict the implants directly from synthetically defective skulls, using a publicly available dataset. The trained model can also be generalized well to real head trauma-related defects, as the synthetic defects the authors created closely mimic real ones. The real defects used in this study are not publicly available. Matzkin et al. [24] focused on cranial implant design for decompressive craniectomy. The authors explored the possibility of both skull shape completion and predicting the implant directly from defective skulls using a U-Net style network. Similar to the studies described above, only synthetic defects are used for training, while, for evaluation, real cases are included. However, the dataset is not yet publicly accessible. These prior studies reveal that the algorithms, if carefully designed, can be generalized to real clinical defects, even if only synthetic defects are used during training. These prior algorithms also accept as input the 3D binary skull images (voxel grids) and produce the implants in the same format. However, the implants need to be converted to meshes in order to be 3D printed. The deep learning method by Zhang et al. [25] is focused on the maxilla area of the skull.

C. Shape Completion

As earlier studies discussed in Section B. cast automatic cranial implant design as a volumetric shape completion problem, this section reviews the general shape completion algorithms used for various data modalities (points, meshes, and voxel grids).

Voxel Grid Completion Classical shape completion algorithms [26], [27] deal with volumetric data, which are voxelized from a point representation using a signed distance function. Voxel grid methods have been prevalent in recent studies using convolutional neural networks (CNN) on volumetric images. Both works employ an encoder-decoder style network, which is, however, restricted to accept as input coarse voxel grids (e.g., 32^3). Meshes can be extracted from the final completed grids.

Point/Mesh Completion Recent development in deep learning enables a CNN to learn from unstructured point clouds efficiently. An encoder-decoder can be used to perform shape completion directly on the raw point data [28], [29] derived from Shapenet [30], which is often used as a benchmark dataset for both voxel grid and point-based shape completion studies. Liu et al. [31] propose a two-step approach to complete dense point clouds. The first step predicts a completed but coarse point cloud using an encoder-decoder style network. In the second step, a residual network is used to produce a dense (high-fidelity) version of the completed point cloud, given as input a combination of the coarse output from the previous step and the partial point cloud. Early studies from Liepa [32] and Kraevoy et al. [33] perform shape completion directly on triangular meshes using classical geometry processing and mesh editing techniques. Shape completion on triangular meshes tends to be much more complicated than on binary voxel grids, as the former data structure can carry much richer information (e.g., texture, color) of an object compared to the latter.

Medical Images Shape completion has also been applied to medical images. Prutsch et al. [34] used a GAN to complete 2D aortic dissection (AD) images (CT), in order to generate the healthy aorta images prior to dissection. Armanious et al. [35] trained a GAN-style network to complete arbitrarily shaped regions on 2D brain images. Gapon et al. [36] adopted a patch similarity matching method to remove metal artifacts on 2D CT and MRI images. A multi-layer perceptron (MLP) was trained to search the best matching patches to the missing region across an image. Manjón et al. [37] used a 3D U-Net to remove lesions on brain MRI images. The trained network can complete the missing region without requiring manual delineation of the lesions, while other studies all require explicit definition of the region of interest (usually done manually) before completion. These medical shape completion applications require the restoration of not only the shape but also the voxel/pixel intensities of the missing region, as medical images are usually grayscale. However, for the skull shape completion task in our challenge, we consider primarily the restoration of the missing shapes as the skull data are binary, containing only 0 and 1.

III The AutoImplant Challenge

A. Organization, Evaluation, and Ranking

The challenge was organized as a satellite event in MICCAI 2020, held virtually due to the COVID-19 pandemic. To our knowledge, this is also the first public challenge targeting the automatic design of cranial implants. Ten teams submitted their prediction results valid for evaluation, along with ten full papers. For ease of reference, the algorithms in the ten papers are denoted as A1 [38], A2 [39], A3 [40], A3 (s) [40], A4 [41], A5 [42], A6 [43], A7 [44], A8 [45], A8 (re) [45], A9 (r) [46], A9 (p) [46], A10 (r) [47], and A10 (bbox) [47], respectively.

Among the algorithms, some [40, 45] have reported an enhanced version of their algorithm, denoted as A3 (s) and A8 (re), besides the base implementations A3 and A8. Two papers [46, 47] reported approaches for comparison, denoted as A9 (r), A9 (p), and A10 (r), A10 (bbox).

Two metrics, Dice Similarity Coefficient (DSC) and symmetric Hausdorff distance (HD, measured in millimeters) are used for the quantitative evaluation of the results. DSC and HD are first ranked separately in descending and ascending order, respectively, and the final ranking is obtained by taking the average of the two rankings, as shown in Figure 1. For [40, 45–47], A3 (s), A8 (re), A9 (p), and A10 (bbox) are used for ranking.

Table I shows the quantitative results (mean DSC and HD) of each algorithm. To get the results, participants needed to submit the predicted implants to the organizers, and a .csv file containing the DSC and HD of each test case was returned to them. It was required that the predicted implants have the same dimensions as the corresponding defective skulls, i.e., $512^2 \times Z$, to be considered valid submissions.

Metrics\Alg	A1	A2	A3	A3 (s)	A4	A5	A6	A7	A8	A8 (re)	A9 (r)	A9 (p)	A10 (r)	A10 (bbox)
DSC (100)	0.917	0.931	0.913	0.845	0.944	0.920	0.907	0.896	0.887	0.891	0.735	0.889	0.810	0.856
DSC (10)	0.919	0.924	0.769	0.816	0.932	0.910	0.870	-	0.351	0.473	-	-	-	-
HD (100)	4.336	3.660	4.067	6.414	3.564	4.137	4.180	4.602	7.017	6.909	7.243	5.534	5.440	5.183
HD (10)	3.987	4.090	8.585	5.952	3.934	4.707	4.760	-	29.476	21.049	-	-	-	-

Table I. Quantitative Results (Mean DSC and HD) of the Participating Algorithms on $D_{test100}$ and D_{test10}

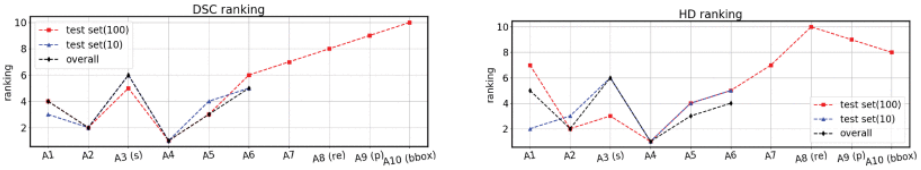


Fig. 1. DSC and HD Rankings of the Algorithms $A1$ -$A10$, on $D_{test100}$, D_{test10} and the overall test set.

Table II shows the t-test for DSC and HD on the entire test set ($D_{test100}$ and D_{test10} combined together) among the leading methods (A_4, A_2, A_1, A_5, A_6, A_3 (s)). We can see that most of the p values are far smaller than 0.05, indicating that the differences among these leading methods are statistically significant. In particular, the winning method (A_4) can beat its followers by a large margin, statistically speaking. In contrast, the difference between A_1 and A_5 is not significant for both DSC and HD. We also show the t-test between some network variants, i.e., A_3 (s) $\leftrightarrow$ A_3, A_8 (re) $\leftrightarrow$ A_8, A_9 (p) $\leftrightarrow$ A_9 (r) and A_{10} (bbox) $\leftrightarrow$ A_{10} (r) in Table II. Except A_9 (p) $\leftrightarrow$ A_9 (r) and A_{10} (bbox) $\leftrightarrow$ A_{10} (r), the t-test is run on the entire test set.

	DSC	HD
$A4 \leftrightarrow A2$	$3.1e^{-4}$	$\mathbf{5.8e^{-1}}$
$A4 \leftrightarrow A1$	$2.4e^{-10}$	$3.0e^{-3}$
$A4 \leftrightarrow A5$	$6.7e^{-8}$	$1.2e^{-2}$
$A4 \leftrightarrow A6$	$9.6e^{-13}$	$2.3e^{-3}$
$A4 \leftrightarrow A3(s)$	$1.4e^{-16}$	$1.3e^{-3}$
$A2 \leftrightarrow A1$	$2.3e^{-3}$	$9.5e^{-3}$
$A2 \leftrightarrow A5$	$1.5e^{-2}$	$3.4e^{-2}$
$A2 \leftrightarrow A6$	$1.2e^{-6}$	$8.5e^{-3}$
$A2 \leftrightarrow A3(s)$	$1.3e^{-13}$	$1.8e^{-3}$
$A1 \leftrightarrow A5$	$\mathbf{7.4e^{-1}}$	$\mathbf{6.7e^{-1}}$
$A1 \leftrightarrow A6$	$1.2e^{-2}$	$\mathbf{7.8e^{-1}}$
$A1 \leftrightarrow A3(s)$	$9.7e^{-11}$	$1.7e^{-2}$
$A5 \leftrightarrow A6$	$8.0e^{-3}$	$\mathbf{8.6e^{-1}}$
$A5 \leftrightarrow A3(s)$	$6.4e^{-11}$	$1.1e^{-2}$
$A6 \leftrightarrow A3(s)$	$1.8e^{-7}$	$1.3e^{-2}$
$A3(s) \leftrightarrow A3$	$3.4e^{-6}$	$3.0e^{-2}$
$A8(re) \leftrightarrow A8$	$\mathbf{5.3e^{-1}}$	$\mathbf{7.3e^{-1}}$
$A9(p) \leftrightarrow A9(r)$	$1.7e^{-47}$	$1.2e^{-3}$
$A10(bbox) \leftrightarrow A10(r)$	$5.9e^{-8}$	$\mathbf{4.5e^{-1}}$

Table II. t-Test Between the Top Ranking Methods (A_4, A_2, A_1, A_5, A_6, A_3 (s)) and Some Network Variants for DSC and HD. p Values Larger Than 0.05 (5×10^{-2}) Are Highlighted

B. Challenge Dataset

We included a data descriptor [48] of the challenge dataset in the challenge proceedings, which detailed the origin, creation and statistics of the challenge dataset. However, a brief description of the training set and test set is provided in this section to make the contribution self-contained. We use the term *complete skull* to refer to the undamaged skull and *defective skull* to refer to a skull with a defect. The complete skulls in the challenge dataset are segmented from a public head CT collection CQ500 (http://headctstudy.qure.ai/dataset), using a thresholding technique (150 Hounsfield Units). The defective skulls are created automatically by removing part of the skull bone from the complete skulls.

Training Set The training set contains 100 complete skulls from different head CT scans and their corresponding synthetic defective skulls and implants. An implant is simply the logical XOR of the corresponding complete skull and defective skull. The defects in the training set follow a similar pattern as illustrated in Figure 2 (A), regarding the size, shape and position. Participants were free to create and use additional defects on the complete skulls provided for training.

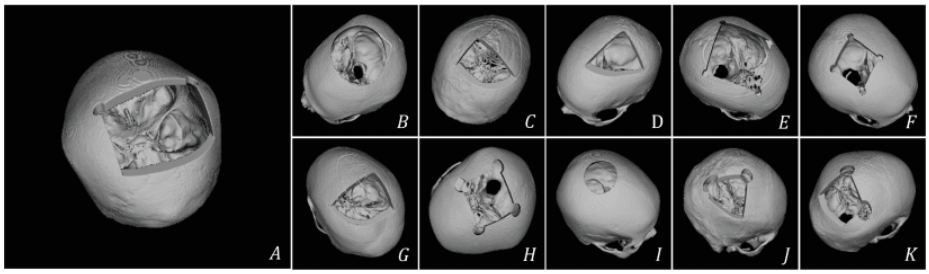


Fig. 2. Illustration of the skull defects in $D_{test100}$ and D_{test10}. The defect in (A) is representative of the defects in the training set and $D_{test100}$, where the defects are similar in terms of shape, size and position. (B)-(K) show the defects in D_{test10}, where there are three types of defects: spherical (B, I), cubic (C, D, G) and cubic with cylinders on the corners (E, F, H, J, K).

Test Set Two independent test sets, denoted as $D_{test100}$ and D_{test10}, were created for evaluation of the submissions. $D_{test100}$ contains 100 defective skulls (created out of 100 skulls different from those in the training set), with defects in $D_{test100}$ similar to those in the training set (Figure 2(A)). D_{test10} contains 10 defective skulls with varying defects, as shown in Figure 2(B)–(K). Figures 1, 5 and 7 denote the two test sets as (100) and (10).

We created the two test sets to evaluate the generalization performance of participants' algorithms. Considering that the skull shape is patient-specific and the shape of the defects from craniotomy also depends on the specific pathological

conditions e.g., the size and position of the brain tumor, of the patients, we expect the participants' algorithms to generalize well across different skulls and defects, which are desired in cranioplasty. According to [48], the defect variation in D_{test10} is much greater than that in $D_{test100}$, which is primarily used to evaluate how well the algorithms can generalize across varied skull shapes, while D_{test10} evaluates whether the algorithms can generalize to randomly shaped, sized and positioned defects, especially when trained only on the training set with a fixed defect pattern shown in Figure 2(A).

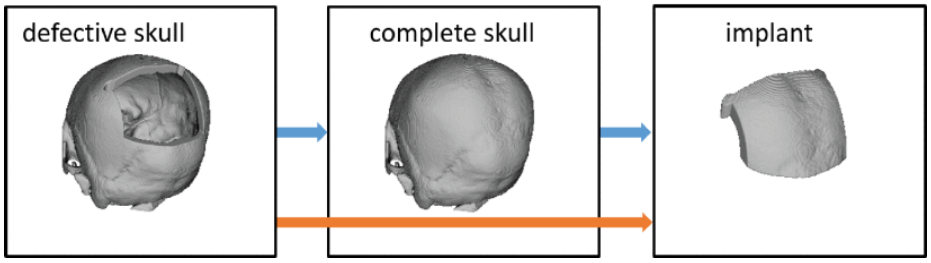


Fig. 3. Two types of problem formulation used among the submitted algorithms. The blue arrowindicates that the algorithms reconstruct a complete skull first, and then the implant is obtainedthrough the subtraction of the defective skull from the reconstructed skull, which defines a *shape completion* problem. The orange arrow indicates that the algorithms reconstruct theimplant directly from a defective skull.

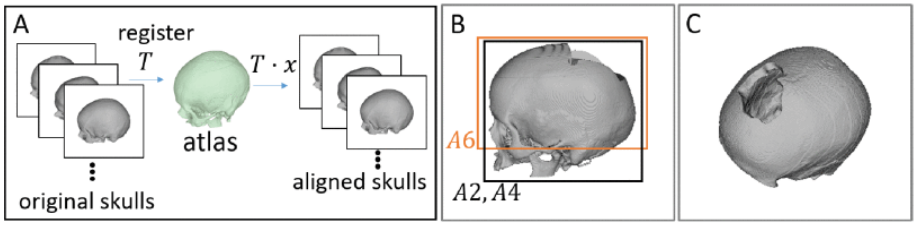


Fig. 4. Illustration of the skull preprocessing methods. (A) Aligning the skulls to a common skull atlas ($A3$). (B) Image background cropping ($A2$, $A4$, $A6$) and (C) Aligning the four anatomical landmarks on each skull onto a common axial plane ($A5$).

Algorithms A_1 – A_6 succeeded on both test sets, while A_7 – A_{10} failed on D_{test10}, and thus A_7 – A_{10} are not included in the ranking on D_{test10}, as shown in Figure 1. We also calculated the ranking of A_1 – A_6 on the entire test set ($D_{test100}$ and D_{test10} combined together).

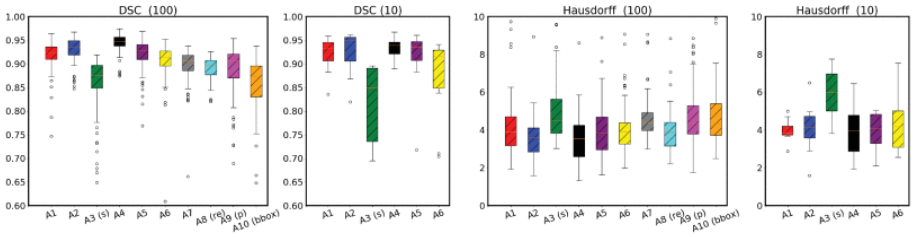
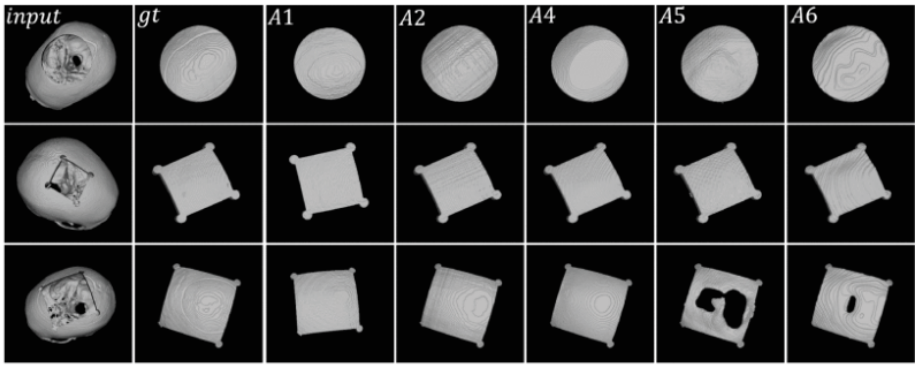
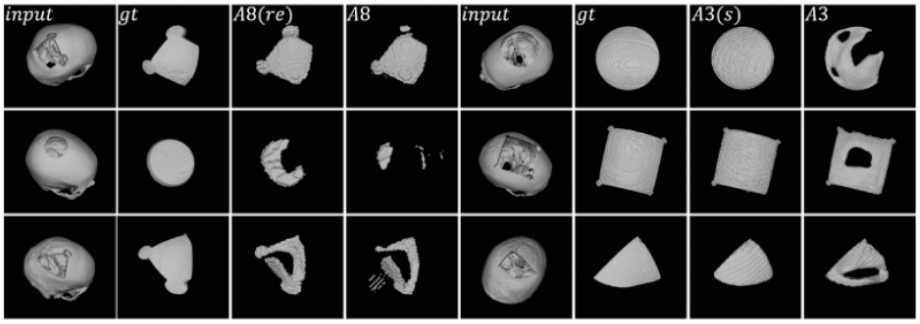


Fig. 5. DSC and HD of algorithms $A1$-$A10$ on $D_{test100}$ (100) and D_{test10} (10). Among the algorithms, $A7$-$A10$ failed on D_{test10} (10).



(a) Implant predictions from $A1$, $A2$, $A4$, $A5$ and $A6$ on D_{test10}.



(b) Comparison between the implant produced by $A8$ and $A8$ (re), $A3$ and $A3$ (s) on D_{test10}.

Fig. 6. Implant predictions on D_{test10}. (a) First to last column: the input, ground truth, predictions from $A1$, $A2$, $A4$, $A5$ and $A6$. (b) The first and fifth column show the input. The second and sixth column show the ground truth. The third and seventh column show the predictions from $A8$ (re) and A3 (s). The fourth and eighth column show the predictions from $A8$ and $A3$

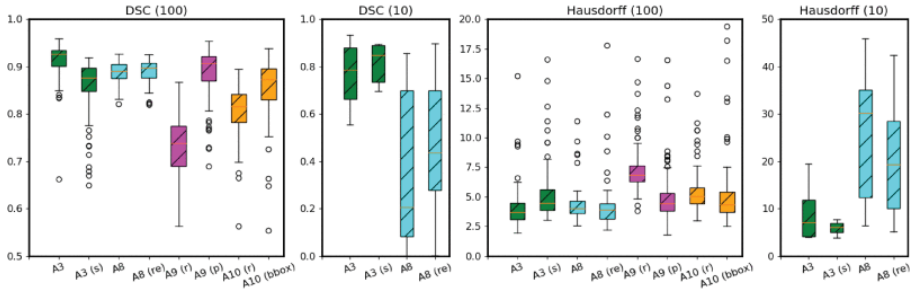


Fig. 7. Comparison between $A3$ and $A3$ (s), $A8$ and $A8$ (re), $A9$ (r) and $A9$ (p), $A10$ (r) and $A10$ (bbox) on $D_{test100}$ and D_{test10} .

All the images in the challenge dataset have dimension $512^2 \times Z$. The ground truth of the test set, i.e., the corresponding complete skulls and implants, were kept secret by the organizers.

Rationale As introduced above, synthetic defective skulls are used in both the training and evaluation phase of our challenge. However, the synthetic defects, as shown in Figure 2(A), are created to resemble the real craniotomy defects by including the drilling holes on the defect borders. In craniotomy, a cranial drill is used by neurosurgeons for drilling small roundish holes on human skulls in order to create an opening in the skull. A craniotome is further used to remove a bone flap to access the brain underneath. This course of action can result in a skull defect with small roundish corners, similar to the artificial ones used in our challenge. The drilling holes are also important for the insertion and fixation of a cranial implant in cranioplasty. However, real craniotomy defects tend to have rough boundaries, as the skull is cut manually using the craniotome, in contrast to the synthetic defects which have smooth and straight boundaries. The rationale for using synthetic defects in our challenge is twofold: First, as discussed in Section B., the algorithms can generalize to real craniotomy defects even if only synthetic defects are involved in the training phase. Second, using real defects in our current challenge is neither practical (not enough data and privacy restrictions) nor efficient (expert evaluations are needed due to a lack of ground truth for the real defects) especially when dozens of submissions are to be expected.

IV Summary and Comparison of the Algorithms

This section summarizes the algorithms from the perspective of problem formulation, skull pre-processing, defect augmentation, network architecture, post-processing and skull dimension. Emphasis will be placed on how the submitted algorithms deal with high-dimensional skull data and on the generalization per-

formance of these algorithms to highly varied skull defects. Table III provides a summary. Specific details of the algorithms can be found in the proceedings [49].

Algorithm	Architecture	Input Dim	Hardware	Skull Preproc.	Defect Augment.	Use of Shape Prior	Output	D_{test10}	# Param
*A*1 [38]	SSM + 2D GAN	256 × 256	4× RTX 6000	yes	no	yes	skull	yes	229.18M
*A*2 [39]	ED + SE block	512 × 512	GTX 1080+GTX 960	yes	yes	no	implant	yes	3.17M
*A*3 [40]	U-Net	304 × 304 × 224	TITAN Xp	yes	yes	no	implant	yes	6.77M
*A*3 (s) [40]	U-Net + shape prior	304 × 304 × 224	TITAN Xp	yes	yes	yes	implant	yes	5.10M
*A*4 [41]	U-Net with residual block	176 × 224 × 144	2× V100	yes	yes	no	skull	yes	68.56M
*A*5 [42]	Cascade U-Net	128 × 128 × 128	Titan Xp	yes	yes	no	implant	yes	5.96M
*A*6 [43]	U-Net	192 × 256 × 128	RTX Titan	yes	yes	no	skull	yes	6.49M
*A*7 [44]	ED with residual block	180 × 180 × 180	Quadro P6000	no	no	no	skull	no	1.49M
*A*8 [45]	RDU-Net	128 × 128 × 64	3× RTX 2080 Ti	no	no	no	skull	no	2.51M
*A*8 (re) [45]	RDU-Net + VAE	128 × 128 × 64	3× RTX 2080 Ti	no	no	yes	skull	partially	25.46M
*A*9 (r) [46]	V-Net + resizing	256 × 256 × 64	RTX Titan	no	no	no	implant	no	45.60M
*A*9 (p) [46]	V-Net + patch	256 × 256 × 64	RTX Titan	no	no	no	implant	no	45.60M
*A*10 (r) [47]	ED + resizing	128 × 128 × 64	GTX 1070 Ti	no	no	no	implant	no	82.00M
*A*10 (bbox) [47]	ED + boundingbox	256 × 256 × 128	GTX 1070 Ti	no	no	no	implant	no	82.00M

Table III. Summary and Comparison of the Algorithms

A. Problem Formulation

As illustrated in Figure 3, two types of problem formulation are used among the algorithms:

1. Some participants formulated the problem of cranial implant design as a volumetric shape completion task. In this formulation, a complete skull is first reconstructed from a defective skull, and the implant is viewed as the difference between the complete skull and the defective skull.
2. Others view the problem as a shape learning task, and the shape of an implant is learned directly from the shape of a defective skull.

The Output column in Table III shows the formulation adopted by each algorithm.

B. Preprocessing of the Skull

Preprocessing the skulls by removing the rotation, translation and other image-related variations helps the deep neural networks to focus on learning the shape variations of the skulls and defects, which is the primary concern of the challenge. This course of action also reduces the difference between the training and test set and thus can potentially help improve the final results. The commonly used techniques for this purpose include image background cropping [39, 41, 43] and skull alignment via registration [40, 42]. Table IV shows a description of the skull preprocessing techniques (if used) per algorithm.

Algorithms	**Preprocessing Methods**
A2 [39]	Image background cropping and resizing the skull region to 512^3, so that the axial, sagittal and coronal slices have the same dimension 512^2.
A3 [40]	Align the training/test set to a common skull atlas via rigid registration.
A4 [41]	Image background cropping and image re-orientation
A5 [42]	Align the skull (on the x/y plane) based on four skull landmarks using rigid registration.
A6 [43]	Crop the image background and the area below the skull base and re-scale all images to 192×256×128 non-isotropically (different scaling factors in the $x-$, $y-$ and $z-$ direction).

Table IV. Description of the Skull Preprocessing Methods Used by Participants

Background Cropping Zero-valued background voxels outside the skull's bounding box provide no useful information for shape learning. Cropping the background reduces image size and thus the memory consumption. In practice,

instead of cropping the entire background, some margins are usually kept (Figure 4, B). After cropping, A_2 further resized the cropped images to 512^3 so that the 2D slices of axial, sagittal and coronal planes have the same size (512^2). A_6 cropped bone structures below the skull base irrelevant to the task, e.g., mandibles (Figure 4, B, A_6). The cropped images were then downsampled to $192 \times 256 \times 128$ to get a fixed input size for the shape completion network. A_4 downsampled the cropped images to $176 \times 224 \times 144$. Note that downsampling the original image volume ($512^2 \times Z$) directly to such low size tends to lead to considerable degradation of image quality and compromises the algorithmic performance. Cropping before downsampling can mitigate such adverse effects.

Skull Alignment To reduce the rotational and translation variations, A_3 and A_3 (s) aligned all the skulls in the training and test set to a common skull atlas using rigid registration, as shown in Figure 4(A). The skull atlas is constructed by averaging the shapes of several complete skulls. Such transformation also resamples the images to an intermediate size of $304^2 \times 224$. In A_5, instead of using a pre-defined skull atlas, the alignment is based on four anatomical landmarks on the skulls, i.e., the left and right auditory meatus and left and right supraorbital notch. These landmarks are aligned onto the same axial plane using a rigid transformation. A U-Net style network is trained to detect the four landmarks automatically on the test set. Another benefit of such a transformation is that the unwanted bone structures below the alignment plane (e.g., midface, mandibles) can be discarded automatically (Figure 4, C). The Skull Preprocessing column in Table III shows whether the algorithms used a preprocessing step. As shown in Figure 5, algorithms that used skull preprocessing generally outperform those that did not. An ablation study performed for [42] demonstrated that using skull alignment improved the quantitative results on a validation set. However, this course of action adds another dimension of complexity. For example, in order to obtain the final prediction y given a test case x, A_3 needs to consider the transformation matrix from registration T and its inverse T^{-1},

$$y = T^{-1} \cdot f(T \cdot x), \tag{2}$$

where f represents the deep neural network. Furthermore, such transformation can usually resample the images to a smaller size [40], which reduces the memory needed to process the skull data. Note that A_4 re-oriented all the images to Right, Anterior, Superior (RAS), which has a similar effect to aligning the skulls in that the data are submitted to the U-Net in the same orientation. As a bonus, re-orientation can be done at runtime and does not require a registration.

C. Defect Augmentation

Among the submitted algorithms, augmentation of skull defects was a dominant factor contributing to the generalization of the algorithms to highly varied defects in D_{test10}. The defects in the training set have limited variations regarding the shape, size, and position, while the defects in the D_{test10} are of much greater

irregularity. It is therefore challenging for the algorithms to generalize well to D_{test10} without the creation and use of additional defects during training, if a standard network configuration is used. In Table III, the Defect Augmentation column shows whether the algorithms have created and used additional defects for training besides those provided in the original challenge dataset. The D_{test10} column shows whether the algorithms can generalize to D_{test10}. We can see that algorithms that generalize to D_{test10} generally also used defect augmentation, with the exception of A_1, which used a strong shape prior to guide the skull reconstruction process. Table V summarizes the defect augmentation techniques (if any) of the algorithms, which can be classified into three groups:

1. Create defects similar to those in $D_{test100}$ and D_{test10} shown in Figure 2 (A_2, A_3, and A_6)
2. Create random defects without resemblance of the defects in the test sets (A_5)
3. Create additional defective skulls using pair-wise registration and warping (A_4)

Algorithms	Augmentation Methods
A2 [39]	Creating defects on 2D slices in axial, sagittal and coronal planes using a rectangular mask.
A3 [40]	Creating 3D defects using random sized masks similar to the defects in D_{test10} (spherical, cubic, cube-cylinder).
A4 [41]	Permutation, scaling, translation and pair-wise non-linear registration and warping.
A5 [42]	Random lateral flipping and creating five random defects per skull.
A6 [43]	Creating defects using spherical and cubic masks.

Table V. Description of the Data Augmentation Methods Used by Participants

Creating Defects Resembling the Test Set A_2 generated additional defects using a rectangular mask in axial, sagittal and coronal slices. To generate defects similar to $D_{test100}$ and D_{test10}, the rectangular mask was tailored according to the defect distributions in the respective test set. A_3 and A_6 generate defects directly on 3D volumes using 3D spherical and cubic masks. A_3 created a mask combining cubes with cylinders to generate defects similar to the defects illustrated in Figure 2(E, F, H, J, K). These augmentation strategies seek to create similar defect distributions on the training set to those of D_{test10}, which further allows the algorithms to generalize to D_{test10}. As introduced in Section B., for evaluation, only synthetic defects were used, which were similar but simplified compared to real craniotomy defects. The success of these augmentation strategies implies that creating synthetic defects that are closely resembling the real

defects (craniotomy, traumatic brain injury or TBI, etc.) for training might help to increase the success rates of the algorithms in clinical scenarios.

Creating Random Defects As shown in Figure 4(C), instead of trying to generate defects similar to those of the test set, A_5 created five defects with random shape, size and position on each skull, resulting in a total of $90 \times 5 = 450$ training pairs (10 skulls in the training set were reserved as a validation set). An ablation study revealed that the algorithm A_5 can generalize to these random defects only if these augmented defects are also involved in the training phase.

Augmentation via Registration and Space Warping D. G. Ellis et al. [41] augment the dataset by registering each skull in the training set with the remaining 99 skulls. For each registration, each skull can be warped into the space of the remaining 99 skulls using the corresponding transformation, yielding 99 uniquely warped skulls. This course of action substantially increases the number of training pairs to $99 \times 100 = 9900$, excluding the defective skulls in the challenge dataset. A_4 used a total of 9803 pairs for training (197 registrations failed).

The three defect augmentation strategies all have proven to be effective in improving the generalization performance of the algorithms on D_{test10}, as illustrated in Figure 6. For training, two algorithms [39, 40] intentionally created defects similar to the test sets, so that the algorithms can naturally generalize well to both of the test sets. However, even if A_6 only augmented spherical and cubic defects, it can still generalize well to the defect pattern shown in the second row of Figure 6(a). Similarly, A_5 augmented random defects, but can generalize to other defect patterns (e.g., spherical defects), according to the first two rows in Figure 6(a). The third row shows that A_5 and A_6 tend to perform worse on large defects. For A_4, the good generalization performance can be largely attributed to the massive augmentation enabled by warping each training skull into the space of the remaining training cases. Unlike other augmentation techniques, which only try to increase the variations of the defects, while the shape variations of the skull are limited to the original training set, this course of action essentially created new skulls.

D. Architecture and Network Configurations

The Architecture column in Table III lists the deep learning models upon which the algorithms are built. We can see that most algorithms, A_2, A_7, A_{10} (r) and A_{10} (bbox), are based on an encoder-decoder (ED) architecture or ED with skip connections, i.e., U-Net for A_3, A_3 (s), A_4, A_5, A_6 and A_8. For A_1, the primary part of the algorithm is based on a statistical shape model (SSM), which reconstructs a complete skull given a defective skull. A generative adversarial network (GAN) is further used to refine the output of the SSM. The GAN is trained using 2D slices from complete skulls from the training set. The generator component of the GAN is an auto-encoder network, which is trained to generate refined 2D skull slices. During the inference stage, the generator takes as input

a combination of 2D slices from the test case (defective skull) and the complete skull reconstructed by the previous SSM, and produces the completed 2D slices, which are aggregated to form the final complete skull in 3D. The corresponding implant is obtained by subtracting the test case from the final reconstructed 3D complete skull. A_3, A_6, A_9 and Li et al. [47] used standard U-Net, V-Net or ED configurations, while variants of ED and U-Net were explored by the other algorithms.

A_2 uses a Squeeze-and-Excitation (SE) block [50], which introduces channel attention mechanisms, and an auxiliary path formed by several convolutional layers, to connect the encoder and decoder part of the network. A_2 uses a standard U-Net to directly predict the implants from defective skulls. However, A_3 proposed a way to incorporate shape prior of the skull into the network, which aims to improve the generalization ability of the network. The base algorithm and the shape-prior-enhanced version are denoted as A_3 and A_3 (s), respectively. A_4 used a U-Net with residual blocks [51] in each level of the convolutional and deconvolutional layers. A_5 used two U-Nets in a cascaded fashion; the first U-Net was responsible for producing a coarse implant of low resolution (128^3) given a downsampled defective skull as input, and the second U-Net was used as a super-resolution network to upsample the low-resolution prediction to high resolution, given as input a patch (128^3) of the prediction concatenated with the corresponding patch of the original high-resolution defective skull. Similar to A_5, A_7 also followed a two-step process to generate high-resolution predictions. First, a 3D ED was used to generate a low-resolution prediction of dimension 180^3. In the ED, residual blocks were used to connect the encoder part of the network with the decoder part. Second, a 2D decoder consisting of several convolutional layers with residual blocks and SE blocks was used to super-resolve the predictions to the original high resolution in a slice-wise manner. A_8 used a Residual Dense U-Net (RDU-Net) [52] as the base implementation. The loss function of the network was enhanced using a shape regularization term derived from a pre-trained variational auto-encoder network (VAE). We denote the network trained with and without the regularization term in the loss function as A_8 and A_8 (re). A_9 (r) and A_9 (p) were based on a V-Net architecture [53]. A_9 (r) used a downsampled version of the skulls for training and produced low-resolution implants ($256^2 \times 64$), which were upsampled to the original dimension using simple image resizing. A_9 (p) used a patch-based method to train on the original skull data using a V-Net. The base implementation of Li et al. [47] is A_{10} (r), which used a standard encoder-decoder to predict implant in low resolution ($128^2 \times 64$). Simple image resizing was used to upsample the implants to the original dimension. A_{10} (bbox) was built upon the output of A_{10} (r), which was used to extract a bounding box (bbox) on the original high-dimensional defective skulls. A_{10} (bbox) used another encoder-decoder network to predict high-resolution implants directly from the bounding box, which has a much smaller size than the original skulls. Most of the networks use DSC loss or a combination of DSC loss and cross-entropy loss as the objective function for this task.

The last column (# Param) of Table III shows the number of trainable parameters of the deep learning components of the algorithms. For A_1, the number refers to the GAN. Note that the top-ranking algorithm A_4 has significantly more parameters (68.56M) than its followers A_2, which has only 3.17M parameters and A_5, which has 5.96M parameters. A_{10} (r), A_9 (r) and A_9 (p) have the second and fourth largest number of parameters, while their performances are ranked the last in our challenge.

E. Shape Priors

Besides defect augmentation, the exploitation of skull shape priors proves to be another effective measure to improve the generalization performance to varied defect patterns. The Use of Shape Prior column in Table III shows whether the skull shape priors is used by each algorithm. We see that even if algorithms A_1 and A_8 (re) used no augmented defects for training, skull shape priors let them still (partially) generalize to D_{test10}. Both quantitative and qualitative comparisons (Figure 7, Figure 6) demonstrate the advantages of shape priors, especially when it comes to defects different from the ones in the training set. The shape prior can be introduced either on-the-fly during the reconstruction process or during the learning process, using shape constraints or contextual information. Among the algorithms submitted, there are three different strategies for using the shape of a complete skull as prior knowledge: (1) Building a statistical model of the complete skulls (A_1), (2) using the shape prior as contextual information during learning (A_3), (3) using the shape prior as shape constraints in the loss function (A_8).

Statistical Shape Model A statistical shape model of the skull represents the average shape as well as principal shape variations of human skulls. The shape representation ability of a SSM is decided largely by the size and diversity of the skull dataset on which the SSM is built. A_1 created a 3D skull SSM using the complete skulls from the training set, using principal component analysis (PCA). In the test phase, a defective skull is fitted to the SSM to find the shape variations that best match the shape of the given test case, during which the SSM acts as a strong shape prior to guide the skull reconstruction. The fitted shape serves as an initial approximation of the reconstructed complete skull corresponding to the test case and is further refined using a GAN. Note that, unlike other algorithms that used both defective skulls and complete skulls or the implant for training, the construction of the skull SSM and the training of the GAN only requires the complete skulls. Such unsupervised learning enables the algorithm to be independent from the defect patterns, and thus its performance is not affected by the shape, size and position of the defects. We can see from Figure 5 that it performs almost equally well on $D_{test100}$ and D_{test10}, even without augmenting the defects. Figure 6(a) shows an illustration of the reconstruction results of A_1.

Shape Prior as Contextual Information A_3 and A_3 (s) are used to evaluate how the incorporation of shape prior affects the performance of the algorithm

on D_{test10}. Both algorithm variants follow the same network and training configuration, except that A_3 (s) uses a skull atlas as an additional input channel for the network during training. The atlas is the same as used for alignment in A_3, which represents the average shape of several complete skulls. By doing so, the skull atlas can provide the contextual information beneficial for the learning process, which distracts the model from overfitting to the defect patterns in the training set and consequently improves robustness of the model. The ablation study of A_3 shows that the algorithm performs better on D_{test10} when a shape prior is incorporated into the network, i.e., A_3 (s) performs better than A_3 regarding DSC and HD, as can be seen from the boxplot in Figure 7. According to Table II, the improvement of DSC and HD on the test set due to the introduction of the shape prior is statistically significant. Qualitatively, we can also see from Figure 6(b) that A_3 failed partially, whereas A_3 (s) can succeed on some of the test cases from D_{test10}.

Shape Constraints in the Loss Function B. Wang et al. [45] reported a comparison of a deep neural network trained with and without shape prior, denoted as A_8 (re) and A_8. In A_8 (re), the shape prior is implemented as a regularization term in the loss function, which tries to minimize the Euclidean distance between the prediction and the ground truth in a latent feature space learned using a VAE. The VAE was trained end-to-end using the complete skulls in the training set to learn a compact and latent shape representation of the complete skulls. A RDU-Net was then used for skull shape completion. During training, the output of the RDU-Net and the corresponding ground truth is encoded into the latent feature space using the encoder part of the pre-trained VAE, and their distance in the latent space is used as a constraint in the learning process, which forces the network to produces anatomically and geometrically plausible skulls. Applying the shape constraint during the training process is similar to the shape fitting stage of the SSM method, where the prior knowledge about the shape of complete skulls is exploited on-the-fly. Besides, this course of action also diverts the attention of the network from the defects to the shape of the skull and thus eases the overfitting to defect patterns. The qualitative and quantitative comparison of A_8 and A_8 (re), according to Figure 6(b) and Figure 7, shows the advantages of using the shape constraints. However, the t-test reported in Table II reveals that the improvement is not statistically significant regarding the quantitative metrics.

F. Post-Processing

Post-processing refers to the final steps taken in order to refine the output, including noise removal, hole filling, etc. These steps are closely related to the choice of problem formulation illustrated in Figure 3.

If the implant is obtained by subtracting the defective skull from a reconstructed complete skull, the resulting implant tends to contain both noise at the implant boundaries and isolated noise, which comes from the mismatch between

the two skulls outside of the defective area. Morphological opening can be used to remove the noise attached to the implant boundaries. The isolated noise can be removed by keeping only the largest component, i.e., the implant, identified via connected component analysis (CCA). A_1 and A_4 applied morphological opening and CCA to the implant sequentially. For A_6, after selecting the largest component using CCA, a spherical topological filter [54] was used to remove the attached noise non-destructively; a morphological closing and anti-aliasing filter was used to fill holes interior of the implant.

Direct implant prediction leads to isolated and attached noise as well. Unlike implants obtained from subtraction, directly predicted implants suffer mainly from attached noise. As before, isolated noise can be removed using CCA, and morphological opening can be used to remove attached noise [44]. However, morphological opening tends to remove not only noise but also fine details of the implant. Thus, A_5 used a detail-preserving strategy to suppress such over-smoothing. An additional morphological dilation operation is applied to the implant after opening, which makes the implant slightly larger than the original implant. A clean implant preserving the fine details can be obtained by masking the original implant with the dilated implant using a logical AND operation.

G. Skull Dimension

The high dimensionality of the skull data posed a major challenge, as it was required that the predicted implants should be of the same dimension as the corresponding skulls. Direct processing of high dimensional skulls is, in many situations, not feasible due to hardware limitations (see the Hardware

Column in Table III). Therefore, most of the algorithms downsampled the skulls to a smaller size before submitting them into the network, as can be seen from the Input dim column in Table III. However, downsampling can cause loss of image quality, and learning from low-quality images yields coarse output (e.g., Figure 8, $A8$ (re)). Comparing these predictions with the ground truth, we can see that the surface of the implants produced by $A8$ (re), $A9$ (r) and $A10$ (r) is severely degraded with terracing artifacts, which is undesirable for this task. These algorithms used standard image resizing (interpolation) techniques to upsample the output to the original dimension for submission and cannot restore surface details.

To produce both high-dimensional and high-quality implants, three different strategies were explored: (1) $A3$, $A4$ and $A6$ reduce the image size before downsampling, as already discussed in Section IV (B) and Table IV. (2) $A5$ and $A9$ use patch-based training. (3) $A5$, $A7$, and $A10$ (bbox) apply a coarse-to-fine framework.

1) Patch-Based Training:

Dividing an image volume into several smaller patches and using these patches to train the network is a commonly applied strategy to deal with high-dimensional data [22]. Using a patch-based training method can lead to substantial improvement of the implant quality, as demonstrated by the comparison of $A9$ (r) and $A9$ (p) in Figure 8. We can see that there are obvious terracing artifacts on

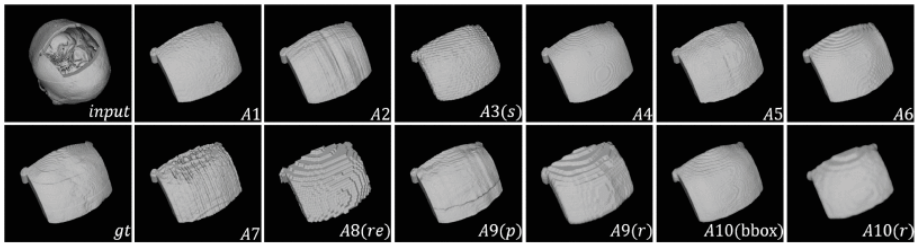


Fig. 8. Illustration of the implants (dimension 5122 × Z) predicted by the algorithms. Left: the input defective skull and the corresponding ground truth implant. Right: the predictions for the algorithms.

the surface of the implant from $A9$ (r) while the implant surface of $A9$ (p) is much smoother. Quantitatively, Figure 7 and Table II also show that $A9$ (p) outperforms $A9$ (r) in terms of DSC and HD by a large margin (statistically significant).

2) Coarse-to-Fine Framework:

$A5$, $A7$ and $A10$ (bbox) adopted a coarse-to-fine framework to produce the desirable implants in two steps; each step is based on a deep neural network. The initial network is trained on downsampled skulls and therefore produces coarse implants. The second network produces fine implants based on the initial coarse prediction. For $A5$ and $A7$, upsampling a coarse implant, while at the same time restoring the geometric details on the implant surface is cast as a volumetric super-resolution task. For $A10$ (bbox), the coarse implant from the first network is used to extract the defective region (2562 × 128) on the original high-dimensional skull, and the second network predicts the fine implant directly from the extracted region, which is much smaller than the original volume. Figure 8 shows a comparison of the implants produced by $A10$ (r), which used standard interpolation for upsampling, and $A10$ (bbox). We can see that the implant from $A10$ (r) looks coarse and blurred on the surface while the implant from $A10$ (bbox) is of much higher quality. $A10$ (bbox) also beats $A10$ (r) regarding DSC and HD according to Figure 7. For DSC, the improvement of $A10$ (bbox) over $A10$ (r) is statistically significant according to Table II. Despite $A10$ (bbox) having better performance than $A10$ (r), its model is significantly more lightweight than that of $A10$ (r) as can be seen in Table III (# Param).

V Discussions

A. Desired Algorithms Characteristics

From both a technical and application perspective, good generalization performance for various cranial defects and the ability to produce high-resolution and high-quality implants with affordable hardware (e.g., a desktop GPU) are among the most desirable characteristics of the algorithms for this challenge. For deep

learning methods, the use of shape priors and defect augmentation can effectively increase the robustness. Besides, a statistical shape model (SSM) of the skull, which represents the general shape of a skull population and is independent from the defects, theoretically has the best generalization ability in this regard. However, the disadvantage of SSM methods is that inference tends to take much longer than with deep learning methods, up to 7–12 minutes per case [38]. The robustness of both deep learning and SSM to highly deformed skulls is restricted to the training samples and can only be increased effectively by including representative deformed cases in the training phase. Processing high-dimensional 3D data, such as the skull data in this challenge, requires ample memory, often exceeding the capacity of commodity hardware. Downsampling the data as a workaround results in severe degradation of image quality. A two-step coarse-to-fine strategy, as used by $A5$, $A7$ and $A10\,(bbox)$, proves to be a solution to this problem.

For the algorithm produced implants, another desired characteristic is that the implants should be in congruency with the skulls in terms of shape and boundary for cosmetic and functional considerations. Figure 9 shows in 3D the reconstructed skulls by $A4$, $A6$, $A7$, $A8\,(re)$ and $A9\,(r)$, given as input a defective skull shown on the top left. It shows that these algorithms can successfully complete the defective skull and restore the missing skull bone, while the surface quality of the reconstructed skulls differs, similar to the implants shown in Figure 8. The reconstructed skulls are further overlaid onto the defective skull to examine how well they overlap in 2D sagittal views. On the defected region shows the difference between the reconstructed and defective skulls, i.e., the implant that can be obtained via a subtraction process illustrated in Figure 3. Ideally, a reconstructed skull should have a 100% overlap with the defective skull except on the defected region, and the implant should fit the skull in terms of shape (e.g., the surface curvature) and bone thickness on the edges. Figure 10 shows an implant created by the winning algorithm ($A4$) overlaid onto the corresponding defective skull. From the 3D view, we can see that the shape of the implant is compatible with its surrounding skull structures in terms of shape and boundary, so that the skull aesthetics can be restored. From the 2D views, we can see that the implant fits tightly against the defect edges on both the interior and exterior skull surfaces. We consider the tight edge contact a desirable characteristic for the implants produced by participants' algorithms.

Extrapolating from the findings of our MICCAI challenge, we see the following directions worthy of future study:

- Expand the training population for SSM and deep learning models, curate collections of head CT (normal, pathological, pediatric etc.) as training datasets [48].
- Explore alternative ways to create and incorporate shape priors or constraints of the skull into deep learning.
- Develop tools to generate synthetic defects from healthy skulls so that they closely resemble the defects from craniotomy, craniectomy and trauma.
- Preprocess the skulls before training to increase the learning efficiency.

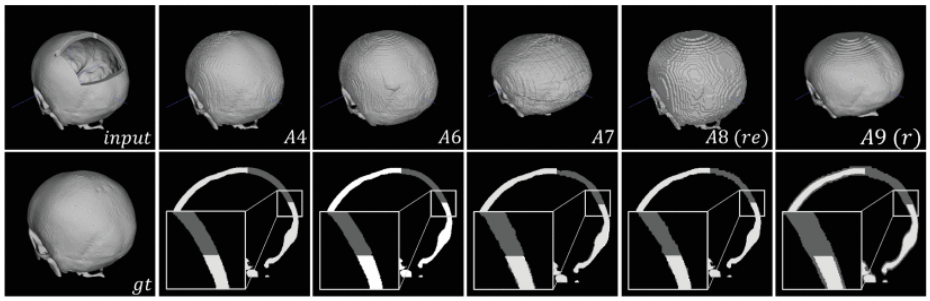


Fig. 9. An illustration of the reconstructed skulls by $A4$, $A6$, $A7$, $A8\,(re)$ and $A9\,(r)$, given as input a defective skull shown on the top left. The second row shows the ground truth skull in 3D and an overlay of the reconstructed skull (dark gray) onto the defective skull (light gray) in sagittal view.

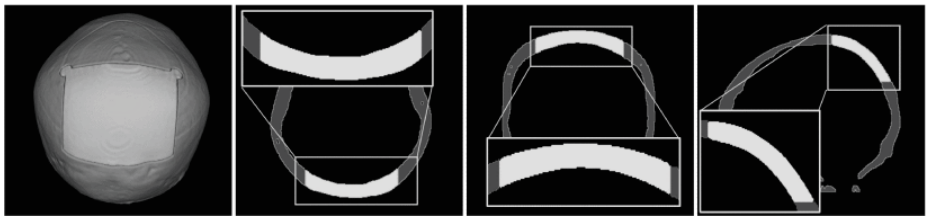


Fig. 10. An overlay of the implant from the winning algorithm ($A4$) onto a defective skull, viewed in 3D, axial, sagittal, and coronal plane. The implanted area is zoomed in for the 2D views. To differentiate, skull and implant are in different colors.

– Extend current deep learning methods to data structures other than a voxel grid representation, such as a point cloud or mesh.

B. Limitations of the 1st AutoImplant Challenge

Dataset The synthetic defects provided for training and evaluation in AutoImplant 2020 challenge are realistic, but simplified compared to real clinical defects from craniotomy and trauma. Defective skulls are hard to obtain clinically, let alone make public to participants, due to both the rarity of such operations and privacy restrictions. Even if defective skulls from clinical routine could be provided for evaluation, there often lack ground truth (implants), making quantitative evaluations impractical. It is therefore hard to judge the performance of the submitted algorithms on real defects. However, even if only synthetic defects were used, this challenge tried to encourage participants to improve the generalization performance of their algorithms to varied skull defects through either defect augmentation or the incorporation of skull shape priors, which have been proven effective to obtain top rankings in the challenge. For this reason, two test sets, D_{test100} and D_{test10}, were used during the evaluation stage. A1 [38] is one of the representative algorithms with excellent generalization performance regarding skull defects, as the algorithm is built only upon healthy skulls and thus is independent from defective pattern.

Another limitation is the small number of unique skulls provided for training. While participants could create indefinite synthetic defects per skull to enlarge the training set and increase the defect variations, shape variations of the skulls were limited to the original 100 skulls provided. No algorithms had data to generalize to pathologically deformed skulls. This limitation can only be overcome by including more skulls in the training set. Hence, we have devised and open-sourced a pipeline [48] to convert collections of head CT, which are much easier to acquire than clinical defective skulls, into trainable datasets for the purpose of cranial implant design. At the core of the pipeline lies the creation of defective skulls out of complete skulls through the injection of synthetic defects. The pipeline can be extended to inject more realistic defects, or even allow multiple defects at once. Such a pipeline can also encourage the incorporation of skull data from different scanners, protocols, or populations into training.

Evaluation Metrics Two quantitative metrics, DSC and HD, were used for the evaluation and ranking of the algorithms. The predicted implant that matches exactly with the ground truth (highest DSC and lowest HD) fits precisely with the defective area on the skull. However, instead of fitting exactly with the defect on the defect boundary, clinically usable implants should be minimally fault-tolerant in case of bone growth (ossification), the presence of scar tissues and osteolysis at the edge of the defects. Furthermore, cranial implant design is an ill-posed problem: An infinite variety of implants can serve the purpose of restoring the mechanical, protective and aesthetic functions. In other words, the ground truth used in the challenge is just one of the many possible solutions. However,

current quantitative metrics constrain the solution to the ground truth, and other implants that are equally clinically usable are penalized during scoring. More work will be needed to formalize the subjective judgement of neurosurgeons based on their professional experience. Besides, the implant boundaries are considered to be critical in cranioplasty and therefore should be given more emphasis compared to other parts of the implant during the evaluation phase. Current metrics, however, treat the implant as a whole and the boundary areas are not distinguished. Boundary-specific evaluation metrics are therefore highly desired in a future edition of the challenge.

C. Commercially Designed Versus Algorithm Produced Implants

In this section, we discuss how far the implants generated by the participants' algorithms are to the commercially designed implants, which are currently the clinical standard. According to our collaborating neurosurgeons, the actual cranial implants used in cranioplasty are usually thinner than the skull bone on the defected region, so that the interior surface of the implants will not apply pressure to the brain (more precisely, to the dura mater). Besides, a clinically usable implant does not necessarily have a tight contact with the skull on the edges. Instead, small gaps (in the order of one millimeter) around the borders of the implant and the skull defect are allowed and sometimes preferred, taking into consideration the bone regeneration over time. Therefore, when necessary, even the commercially designed and manufactured implants require some manual post-processing (e.g., rasping) by neurosurgeons before they can be used, especially when the design and manufacturing of the cranial implant takes a long time [5].

However, our challenge was designed to generate implants that can tightly fit the skull defects, as can be seen from Figure 9 and Figure 10. By doing so, the implants produced by the participants' algorithms can be further post-processed and rasped where necessary. Conversely, a too small implant is neither usable nor remediable via post-processing.

VI Conclusion

This paper is aimed at giving a comprehensive overview of the first AutoImplant challenge hosted at MICCAI 2020. Contributions, approaches, evaluation results and algorithmic trends have been presented and discussed. We also included a critical judgement of current limitations for practical usage from clinical partners. With numerous participants and contributions from academia and industry around the world, the challenge provided a strong stimulus for automatic cranial implant design. To date, the challenge website remains open for post-challenge registrations and submissions, which has been accepted by the community as demonstrated by dozens of new registrations since the official end of the first challenge deadline. Should there be a future edition of the AutoImplant challenge, real defective skulls from craniotomy should be provided and the neurosurgeons' judgement on the clinical usability of the predicted implants should

be involved in the evaluation phase. The scope could also be expanded to other medical scenarios involving computer-aided implant design, such as the lower jawbone [55] or ribs [56].

References

1. R. Stefini, G. Esposito, B. Zanotti, C. Iaccarino, M. M. Fontanella, and F. Servadei. Use of "custom made" porous hydroxyapatite implants for cranioplasty: postoperative analysis of complications in 1549 patients. *Surgical Neurology International*, 4, 2013.
2. A. Morais. Automated computer-aided design of cranial implants-a deep learning approach. Master's thesis, Universidade do Minho, 2018.
3. J. V. Rosenfeld and J. W. Tee. Complications after decompressive craniectomy and cranioplasty. In *Complications in Neurosurgery*, pages 266–273. Elsevier, 2019.
4. D. B. Kurland, A. Khaladj-Ghom, J. A. Stokum, B. Carusillo, J. K. Karimy, V. Gerzanich, J. Sahuquillo, and J. M. Simard. Complications associated with decompressive craniectomy: a systematic review. *Neurocritical Care*, 23(2):292–304, 2015.
5. G. von Campe and K. Pistracher. Patient specific implants (psi): Cranioplasty in the neurosurgical clinical routine. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 1–9. Springer, 2020.
6. J. Li, A. Pepe, C. Gsaxner, and J. Egger. An online platform for automatic skull defect restoration and cranial implant design. In *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11598, page 115981Q. International Society for Optics and Photonics, 2021.
7. X. Chen, L. Xu, X. Li, and J. Egger. Computer-aided implant design for the restoration of cranial defects. *Scientific Reports*, pages 1–10, 2017.
8. A. Marzola et al. A semi-automatic hybrid approach for defective skulls reconstruction. *Computer-Aided Design and Applications*, 17:190–204, 2019.
9. M. Gall, X. Li, X. Chen, D. Schmalstieg, and J. Egger. Computer-aided planning and reconstruction of cranial 3d implants. In *IEEE Engineering in Medicine and Biology Society*, pages 1179–1183, 2016.
10. J. Egger et al. Interactive reconstructions of cranial 3d implants under mevislab as an alternative to commercial planning software. *PLoS ONE*, 12:20, 2017.
11. L. Mei, M. Figl, A. Darzi, D. Rueckert, and P. Edwards. Sample sufficiency and pca dimension for statistical shape models. In *European Conference on Computer Vision*, pages 492–503. Springer, 2008.
12. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.
13. C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
14. O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.
15. J. Carr, W. R. Fright, and R. Beatson. Surface interpolation with radial basis functions for medical imaging. *IEEE Transactions on Medical Imaging*, 16:96–107, 1997.

16. W. Semper-Hogg et al. Virtual reconstruction of midface defects using statistical shape models. *Journal of cranio-maxillo-facial surgery*, 45(4):461–466, 2017.
17. M. A. Fuessinger et al. Virtual reconstruction of bilateral midfacial defects by using statistical shape modeling. *Journal of Craniomaxillofacial Surgery*, 47:1054–1059, 2019.
18. M. A. Fuessinger et al. Planning of skull reconstruction based on a statistical shape model combined with geometric morphometrics. *International Journal of Computer Assisted Radiology and Surgery*, 13:519–529, 2017.
19. H. Lamecker. *Variational and statistical shape modeling for 3D geometry reconstruction*. PhD thesis, Zuse-Institut Berlin, 2008.
20. Z. Kun. Dense correspondence and statistical shape reconstruction of fractured, incomplete skulls. Master's thesis, National University of Singapore, 2014.
21. A. Morais, J. Egger, and V. Alves. Automated computer-aided design of cranial implants using a deep volumetric convolutional denoising autoencoder. In *World Conference on Information Systems and Technologies*, pages 151–160. Springer, 2019.
22. J. Li. Deep learning for cranial defect reconstruction. Master's thesis, Graz University of Technology, 2020.
23. O. Kodym, M. Španěl, and A. Herout. Skull shape reconstruction using cascaded convolutional networks. *Computers in Biology and Medicine*, 123:103886, 2020.
24. F. Matzkin, V. Newcombe, S. Stevenson, A. Khetani, T. Newman, R. Digby, A. Stevens, B. Glocker, and E. Ferrante. Self-supervised skull reconstruction in brain ct images with decompressive craniectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 390–399. Springer, 2020.
25. Y. Zhang, Y. Pei, Y. Guo, S. Chen, T. Xu, and H. Zha. Cleft volume estimation and maxilla completion using cascaded deep neural networks. In *International Workshop on Machine Learning in Medical Imaging*, 2020.
26. X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *IEEE International Conference on Computer Vision*, pages 85–93, 2017.
27. A. Dai, C. Ruizhongtai Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017.
28. X. Wen, T. Li, Z. Han, and Y.-S. Liu. Point cloud completion by skip-attention network with hierarchical folding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1939–1948, 2020.
29. W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. Pcn: Point completion network. In *International Conference on 3D Vision*, pages 728–737. IEEE, 2018.
30. A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
31. M. Liu, L. Sheng, S. Yang, J. Shao, and S.-M. Hu. Morphing and sampling network for dense point cloud completion. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11596–11603, 2020.
32. P. Liepa. Filling holes in meshes. In *Eurographics Symposium on Geometry Processing*, pages 200–205, 2003.
33. V. Kraevoy and A. Sheffer. Template-based mesh completion. In *Eurographics Symposium on Geometry Processing*, volume 385, pages 13–22, 2005.

34. A. Prutsch, A. Pepe, and J. Egger. Design and development of a web-based tool for inpainting of dissected aortae in angiography images. *arXiv preprint arXiv:2005.02760*, 2020.
35. K. Armanious, V. Kumar, S. Abdulatif, T. Hepp, S. Gatidis, and B. Yang. ipa-medgan: Inpainting of arbitrary regions in medical imaging. In *IEEE International Conference on Image Processing*, pages 3005–3009. IEEE, 2020.
36. N. Gapon, V. Voronin, R. Sizyakin, D. Bakaev, and A. Skorikova. Medical image inpainting using multi-scale patches and neural networks concepts. In *IOP Conference Series: Materials Science and Engineering*, volume 680, page 012040. IOP Publishing, 2019.
37. J. V. Manjón, J. E. Romero, R. Vivo-Hernando, G. Rubio, F. Aparici, M. de la Iglesia-Vaya, T. Tourdias, and P. Coupé. Blind mri brain lesion inpainting using deep learning. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 41–49. Springer, 2020.
38. P. Pimentel et al. Automated virtual reconstruction of large skull defects using statistical shape models and generative adversarial networks. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 16–27. Springer, 2020.
39. H. Shi and X. Chen. Cranial implant design through multiaxial slice inpainting using deep learning. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 28–36. Springer, 2020.
40. F. Matzkin, V. Newcombe, B. Glocker, and E. Ferrante. Cranial implant design via virtual craniectomy with shape priors. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 37–46. Springer, 2020.
41. D. G. Ellis and M. R. Aizenberg. Deep learning using augmentation via registration: 1st place solution to the autoimplant 2020 challenge. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 47–55. Springer, 2020.
42. O. Kodym, M. Španěl, and A. Herout. Cranial defect reconstruction using cascaded cnn with alignment. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 56–64. Springer, 2020.
43. J. G. Mainprize, Z. Fishman, and M. R. Hardisty. Shape completion by u-net: An approach to the autoimplant miccai cranial implant design challenge. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 65–76. Springer, 2020.
44. A. Bayat, S. Shit, A. Kilian, J. T. Liechtenstein, J. S. Kirschke, and B. H. Menze. Cranial implant prediction using low-resolution 3d shape completion and high-resolution 2d refinement. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 77–84. Springer, 2020.
45. B. Wang et al. Cranial implant design using a deep learning method with anatomical regularization. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 85–93. Springer, 2020.
46. Y. Jin, J. Li, and J. Egger. High-resolution cranial implant prediction via patch-wise training. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 94–103. Springer, 2020.
47. J. Li, A. Pepe, C. Gsaxner, G. von Campe, and J. Egger. A baseline approach for autoimplant: the miccai 2020 cranial implant design challenge. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 75–84. Springer, 2020.

48. J. Li and J. Egger. Dataset descriptor for the autoimplant cranial implant design challenge. In *Towards the Automatization of Cranial Implant Design in Cranioplasty*, pages 10–15. Springer, 2020.
49. J. Li and J. Egger. *Towards the Automatization of Cranial Implant Design in Cranioplasty: First Challenge, AutoImplant 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings*. Lecture Notes in Computer Science. Springer International Publishing, 2020.
50. J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.
51. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
52. P. Shamsolmoali, M. Zareapoor, R. Wang, H. Zhou, and J. Yang. A novel deep structure u-net for sea-land segmentation in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(9):3219–3232, 2019.
53. F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571. IEEE, 2016.
54. M. Styner et al. Framework for the statistical shape analysis of brain structures using spharm-pdm. *Insight Journal*, (1071), 2006.
55. L. Nickels. World's first patient-specific jaw implant. *Metal Powder Report*, 67(2):12–14, 2012.
56. J. Kang, L. Wang, C. Yang, L. Wang, C. Yi, J. He, and D. Li. Custom design and biomechanical analysis of 3d-printed peek rib prostheses. *Biomechanics and Modeling in Mechanobiology*, 17(4):1083–1092, 2018.

Improving uncertainty estimates under domain shift in white matter hyperintensity segmentation via maximum-entropy regularization

Towards Reliable WMH Segmentation under Domain Shift: An Application Study using Maximum Entropy Regularization to Improve Uncertainty Estimation

Franco Matzkin[1], Agostina Larrazabal[3], Diego H Milone[1], Jose Dolz[2], and Enzo Ferrante[4]

[1] Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, Argentina
[2] Laboratory for Imagery, Vision and Artificial Intelligence, LIVIA, ETS, Montreal, Canada
[3] Tryolabs, Uruguay
[4] Institute of Computer Sciences, ICC, CONICET-Universidad de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina

Abstract. Background: Accurate segmentation of white matter hyperintensities (WMH) is crucial for clinical decision-making, particularly in the context of multiple sclerosis. However, domain shifts, such as variations in MRI machine types or acquisition parameters, pose significant challenges to model calibration and uncertainty estimation. This comparative study investigates the impact of domain shift on WMH segmentation, proposing maximum-entropy regularization techniques to enhance model calibration and uncertainty estimation. The purpose is to identify errors appearing after model deployment in clinical scenarios using predictive uncertainty as a proxy measure, since it does not require ground-truth labels to be computed.

Methods: We conducted experiments using a classic U-Net architecture and evaluated maximum entropy regularization schemes to improve model calibration under domain shift on two publicly available datasets: the WMH Segmentation Challenge and the 3D-MR-MS dataset. Performance is assessed with Dice coefficient, Hausdorff distance, expected calibration error, and entropy-based uncertainty estimates.

Results: Entropy-based uncertainty estimates can anticipate segmentation errors, both in-distribution and out-of-distribution, with maximum-entropy regularization further strengthening the correlation between uncertainty and segmentation performance, while also improving model calibration under domain shift.

Conclusions: Maximum-entropy regularization improves uncertainty estimation for WMH segmentation under domain shift. By strengthening the relationship between predictive uncertainty and segmentation errors, these methods allow models to better flag unreliable predictions without requiring ground-truth annotations. Additionally, maximum-entropy

regularization contributes to better model calibration, supporting more reliable and safer deployment of deep learning models in multi-center and heterogeneous clinical environments.

Keywords: White Matter Hyperintensity · Uncertainty Estimation · Domain Shift · Medical Image Segmentation · Maximum-Entropy Regularization

Highlights

- Entropy-based uncertainty estimates can be used as a proxy for segmentation errors.
- Maximum-entropy regularization improves model calibration and uncertainty quantification under domain shift in WMH segmentation.
- Models trained with maximum-entropy regularization achieve stronger alignment between uncertainty and segmentation errors.
- Validation performed on multicenter WMH datasets highlights robustness to different imaging conditions.

1 Introduction

Accurate segmentation in medical imaging is crucial for a variety of clinical applications, from computer-aided diagnostics to treatment planning [21]. In the context of Multiple Sclerosis (MS) research, the segmentation of hyperintensity areas, identifiable on head MRI scans, are indicative of pathological changes in the brain and are closely associated with MS pathology. Developing robust automated segmentation methods is crucial to improve the understanding of white matter hyperintensities (WMH) and enhancing diagnosis, monitoring, and treatment strategies for MS patients, making WMH segmentation a key predictor [15]. Accurate and reliable WMH segmentation directly impacts patient care and clinical decision-making, as it helps in estimating lesion load, an important marker for disease progression and treatment response [2].

This task is usually approached using deep learning strategies based on convolutional neural networks (CNNs) [20]. Models based on CNNs, known for their outstanding performance in segmentation tasks, heavily rely on consistent distributions between training and test datasets. When confronted with changes in distribution, such as variations in MRI machine types or acquisition parameters across different medical centers, a phenomenon known as domain shift occurs, usually leading to a decline in segmentation accuracy. This presents a significant challenge as it compromises the model's ability to generalize effectively across diverse imaging scenarios. In addition to compromising the discriminative performance of the model, domain shift can also impact its calibration [17, 11, 14]. Calibration, which refers to the alignment between predicted probabilities and observed outcomes, is essential for accurate decision-making [19]. When faced with domain shift, the model predictions could become less calibrated in

the target domain, potentially misleading the clinician's interpretation of the results. Poor calibration can lead to overconfidence in wrong decisions or unnecessary doubts about correct ones. While one would expect the probabilistic outputs of CNN segmentation models to be affected by domain shift, manifesting higher uncertainty in the predictions, this is not usually the case. Instead, models tend to remain overconfident even in situations where predictions are wrong (e.g. producing predictions close to 0 –background– or 1 –lesion– in a binary lesion segmentation scenario, instead of assigning values close to 0.5 which would better reflect uncertainty about the unknown data distribution).

Therefore, addressing domain shift is not only important to ensure accurate segmentation, but also plays a vital role in maintaining the calibration of the model across domains, ultimately enhancing its utility in clinical practice. In this work, we are interested in quantifying model uncertainty under domain shift scenarios, a concept closely related to model calibration. In cases when we go from in-distribution (ID) data samples, which are similar to the training data, to out-of-distribution (OOD) samples, which deviate from the training data distribution, uncertainty quantification (UQ) can allow us to flag segmentation cases which require intervention [9].

In the context of medical imaging, models trained with classical loss functions (such as the popular Cross Entropy –CE– or soft Dice loss [10]) may exhibit overconfidence in their predictions when faced with OOD data, leading to suboptimal outcomes. An example is shown in Figure 1 (central column), where a WMH segmentation prediction generated by a model trained using a classical pixel level CE as the loss function, shows the label likelihood close to one across the entire segmented area. However, it would be more beneficial for the model to express uncertainty in areas where less consensus between raters could be expected, such as at lesion boundaries or in small, isolated lesions distant from larger lesion areas (as in the right column). This discrepancy underscores the need for more sophisticated loss functions and training strategies that can effectively address domain shift challenges in medical imaging applications, encouraging the model to doubt in OOD scenarios, instead of producing overconfident predictions.

Previous work has proposed the use of regularization methods to discourage overconfident predictions. In the context of classification problems, [16] proposed to increase entropy in the probabilistic output (i.e. preventing peaked distributions and promoting uniformity) of classification models by incorporating an additional regularization term to the loss function, representing the negative entropy of the output probability. Since confident predictions correspond to output distributions that have low entropy, this regularization term that prevents peaked distributions was shown to help avoid overconfidence for ID data. However, it was not evaluated under distribution shift scenarios. This idea was further refined in [7] where, instead of penalizing low entropy for all predictions, only the erroneous ones were penalized, resulting in more accurate segmentations for ID data.

So far, the use of maximum entropy methods for image segmentation has mostly been limited to ID data. At the same time, previous work [13] inves-

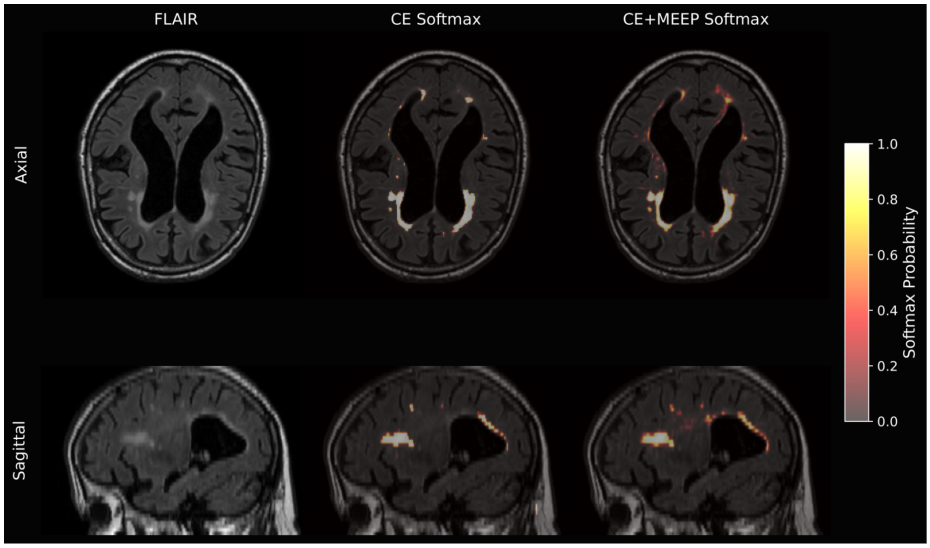


Fig. 1. Comparison of White Matter Hyperintensity (WMH) segmentation on axial and sagittal FLAIR MRI from a multiple sclerosis patient. The input FLAIR image (left) and an overconfident output from a CE Softmax model (center) are contrasted with the result from CE_{MEEP} Softmax (right). This latter approach yields more detailed probabilistic segmentations, capturing uncertainty more effectively through intermediate values—especially around lesion boundaries and in small WMH regions.

tigated the use of entropy as a measure for uncertainty quantification in the context of WMH segmentation. However, they did not explore maximum entropy methods to enhance these estimates, nor did they address the implications of distribution shifts, which are a critical issue in multi-centric scenarios. Here we study maximum entropy methods to improve UQ *under distribution shifts* in WMH segmentation. In particular, we will examine whether these models can maintain accurate uncertainty estimation when confronted with changes in data distribution, crucial for reliable decision-making in clinical settings. Additionally, we will explore model calibration under OOD scenarios, providing insights into the effectiveness of the maximum entropy methods in detecting erroneous cases. By assessing the model performance across various medical centers and imaging scenarios, our goal is to uncover its adaptation and generalization capacity in diverse clinical environments, ultimately aiming to provide valuable guidance for integrating deep learning models into clinical practice and advancing patient care outcomes.

Contributions: Our main contributions are threefold: 1) we investigate the impact of domain shift on model calibration for WMH segmentation, 2) we propose the use of maximum entropy regularization for improving uncertainty estimates in WMH segmentation under domain shift, and 3) we assess the correlation between uncertainty and segmentation errors in this scenario. By achieving these goals, we aim to enhance the reliability and clinical applicability of deep learning models in the context of WMH segmentation for MS patients. Specifically, we hypothesize that higher entropy values will correlate with lower Dice scores, particularly under domain shift conditions, enabling entropy-based uncertainty estimates to serve as reliable proxies for segmentation performance. To validate this hypothesis, we systematically evaluate existing entropy-based regularization methods on multicentric MRI datasets acquired under varying scanning protocols and patient populations. In our experiments, maximum entropy regularization methods indeed improved uncertainty estimation and calibration under domain shift.

2 Materials and methods

Let us say we have a segmentation model $S : X \rightarrow Y$ that, given an image X, returns a probabilistic voxel level segmentation map Y, as $Y = S(X)$. For every voxel i, Y will assign a probability y_i for the WMH lesion class, and $1 - y_i$ will be the probability of healthy tissue. Without loss of generality, in our case the model S is an encoder-decoder convolutional neural network which follows a U-Net architecture [18]. Note that this formulation is model-agnostic, and hence other architectures could also be considered. Given the probabilistic segmentation map, we aim to estimate voxel-level uncertainty. In this study, we focus on predictive entropy as the uncertainty metric.

2.1 Entropy-based uncertainty estimation

Various methods have been proposed for estimating uncertainty in medical image segmentation, including Monte Carlo Dropout [4], model ensembling, and Probabilistic U-Net [5]. In this work, we focus on predictive entropy, a widely adopted approach [3, 13] due to its simplicity and interpretability.

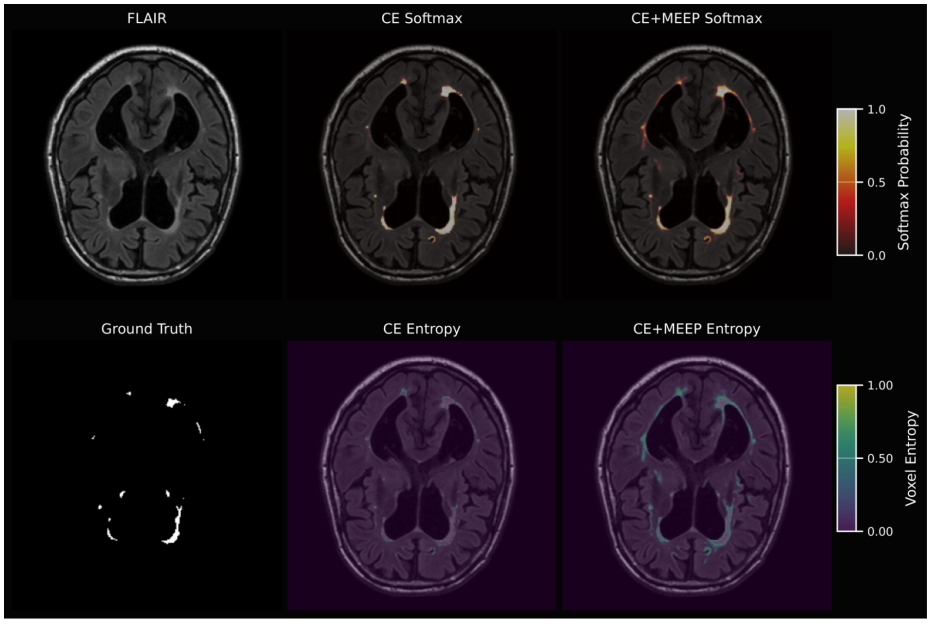


Fig. 2. Input FLAIR MRI (top left) and ground truth segmentation (bottom left) for White Matter Hyperintensities (WMH). These are shown alongside softmax probability outputs from CE Softmax (top center) and CE_{MEEP} Softmax (top right) models, and their respective voxel entropy maps: CE Entropy (bottom center) and CE_{MEEP} Entropy (bottom right). Notably, the CE_{MEEP} Entropy maps more distinctly highlights uncertainty in small WMH lesions visible in the ground truth, compared to the CE Entropy map.

Uncertainty in model predictions can be estimated using predictive entropy. For binary segmentation, the binary entropy of the segmented region has been employed to provide insights into the confidence levels associated with the predictions [3, 9, 13]. Given a Bernoulli probability distribution parameterized by p, its binary entropy is defined as

$$H_b(p) = -p \log_2(p) - (1-p) \log_2(1-p),$$

where p stands for the probability that a voxel or data point belongs to the foreground class, which, in case of WMH segmentation, is the probability associated with the lesion class. The binary entropy H_b could range from 0 to 1: when

$H_b = 0$, the outcome is entirely predictable, and when $H_b = 1$ it is completely unpredictable or random. In a binary segmentation scenario, if a model assigns a probability close to 1 for a voxel belonging to the target class, then the entropy will be very low, indicating high confidence. Alternatively, if the model assigns a probability of 0.5, the entropy will be maximum, indicating high uncertainty (see Figure 2). This allows practitioners to identify uncertain regions, potentially requiring further inspection or intervention, thus enhancing the model reliability and interpretability.

2.2 Improving entropy-based uncertainty estimation via maximum entropy methods

We propose three strategies to promote higher entropy distributions and evaluate their effectiveness in terms of uncertainty estimation under domain shift. Such strategies are implemented as an additional term in the loss function for training the neural network. In general, we will train our models using the following loss function:

$$L = L_{\text{seg}}(Y, \hat{Y}) + L_{\text{reg}}(Y),$$

where L_{seg} is the data term (either cross entropy or soft Dice loss) computed by comparing the predicted segmentation mask Y with the ground-truth label $\hat{Y}$, and L_{reg} is a regularization term defined to encourage high entropy. In what follows, we introduce three alternatives for this regularization term.

Overall confidence penalty As previously discussed, overconfident models tend to assign all probability into a single class. To avoid such behavior, we first propose to encourage high entropy for all voxel predictions. We follow the idea introduced by [16] in the context of image classification, adapting it to the context of image segmentation. Thus the entropy of *all* voxel predictions y_i in the predicted segmentation mask Y is computed, defining the regularization term as

$$\mathcal{L}_a(Y) = -H_b(Y) = -\sum_i y_i \log_2(y_i) - (1 - y_i) \log_2(1 - y_i).$$

This term is added in the the overall loss function, encouraging maximum entropy for all voxel predictions:

$$L(Y, \hat{Y}) = L_{\text{seg}}(Y, \hat{Y}) + \mathcal{L}_a(Y).$$

This approach systematically enforces higher entropy in the outputs of the model, acting as a strong regularizer and improving generalization by reducing overconfidence even in correct predictions.

Maximum entropy on erroneous predictions The term defined in the previous section penalizes high confidence for all voxel predictions. However, if a prediction is correct, in principle there is nothing wrong with the model being

confident about it. Indeed, we argue that one would like to avoid overconfident predictions especially in cases where those predictions are wrong. Thus, we resort to the maximum entropy on erroneous predictions (MEEP) regularizer, $L_m(Y_w)$, which penalizes low entropy only for erroneous predictions. We will use $\mathbf{y}_w$ to define the set of voxels whose label was incorrectly predicted, and hence we can define the regularizer as

$$L_m(Y_w) = -H_b(Y_w) = -\sum_{i \in Y_w} y_i \log_2(y_i) - (1 - y_i) \log_2(1 - y_i).$$

This regularizer will penalize low entropy (i.e. peaky) distributions only when the predictions are wrong, which intuitively encourages uniform predictions in highly uncertain situations. In particular, we hypothesize that this term will help in domain shift scenarios due to changes in intensity distributions when facing multicentric datasets. Similarly as before, we will add this term to the overall loss function, encouraging maximum entropy only for voxels which were wrongly predicted, resulting in the following loss

$$L(Y, \hat{Y}) = L_{\text{seg}}(Y, \hat{Y}) + L_m(Y_w).$$

Maximum entropy on erroneous predictions via KL divergence We evaluate a third approach where we also encourage high entropy in erroneous predictions but following a different strategy. Instead of subtracting the entropy of misclassified voxels from the overall loss function, we introduce a regularization term to encourage their predictions to be uniformly distributed, by minimizing the Kullback-Leibler (KL) divergence with respect to a uniform distribution. The KL divergence $D_{KL}(Q\|P)$ provides a notion of difference between two probability distributions P and Q. Since the uniform distribution has maximum entropy, we will minimize the difference between the predicted distribution for misclassified voxels Y_w and the uniform distribution Q, by adding a regularization term

$$L_{KL}(Y_w) = -D_{KL}(Q\|Y_w),$$

resulting in the loss function:

$$L(Y, \hat{Y}) = L_{\text{seg}}(Y, \hat{Y}) + L_{KL}(Y_w).$$

Note that although $L_{KL}(Y_w)$ and $L_m(Y_w)$ drive Y_w towards a uniform distribution, their gradient dynamics differ, resulting in different effects on the neural weight updates during training. In this study, we conduct an experimental analysis to determine which term yields better UQ under domain shift.

2.3 Metrics and evaluation protocols

Here we are interested in assessing how WMH segmentation models behave under domain shift, improve their performance both in terms of discrimination and calibration, and understand if the entropy of the predictions can be used as a

proxy to anticipate potential failures. These aspects provide a comprehensive understanding of the overall performance and its suitability for real-world applications. In what follows, we describe the metrics that are used to evaluate each of these aspects.

Discrimination metrics Discriminative ability is achieved when the model can effectively distinguish between different classes. For the segmentation tasks, the **Dice Coefficient** was used. This widely used metric measures the overlap between the predicted segmentation and the ground truth. It is calculated as

$$\text{Dice} = \frac{2 \cdot |G \cap P|}{|G| + |P|},$$

where G represents the ground truth set and P the predicted set, and $|\cdot|$ denotes the number of elements in the set.

Calibration metrics Calibration metrics are crucial for assessing how well the predicted probabilities of a model align with actual outcomes. Previous studies have shown that segmentation models trained with Dice loss tend to be overconfident [22, 12], while cross-entropy training typically leads to better calibrated models [9]. Among calibration metrics, the Expected Calibration Error (ECE) is useful for assessing the reliability of probability estimates. To calculate it, we first allocate each voxel prediction to a bin, depending on the predicted probability value. Here we consider bin separation of 0.1, resulting in $M = 10$ bins of the form $\{B_0 = [0, 0.1), B_1 = [0.1, 0.2), \ldots, B_{10} = [0.9, 1]\}$. ECE is then calculated as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|,$$

where B_m is the set of samples in bin m, n is the total number of samples, $\text{acc}(B_m)$ is the accuracy of voxels in bin B_m, and $\text{conf}(B_m)$ is the average confidence (predicted probability value) of bin B_m. This metric captures the average discrepancy between predicted probability and actual accuracy across all bins.

Another essential tool for assessing calibration is the **reliability plot**. This graphical representation plots the average predicted probability, p, against the actual fraction of positives, f_p, for each bin. Ideally, the points in a reliability plot should lie on the line $p = f_p$, indicating perfect calibration where the predicted probability matches the observed frequency of the event. This visualization helps identify areas where the model is overconfident or underconfident in its predictions. By incorporating these metrics, we can comprehensively evaluate both the discriminative power and the calibration quality of machine learning models, ensuring their reliability and effectiveness in clinical practice.

Uncertainty quantification protocols To evaluate the relationship between segmentation performance and uncertainty estimates, we computed the Pearson correlation between the average foreground entropy and the Dice coefficient

across scans. For each case, we first filtered voxels classified by the model as foreground (predicted probability > 0.5) and then computed the mean entropy over these voxels. This approach simulates a clinical scenario where ground-truth labels are unavailable, focusing the uncertainty analysis on the model's positive predictions.

Evaluation on different lesion sizes Previous work has shown that WMH segmentation methods tend to present lower quality for smaller lesions [2]. Thus, one would expect that entropy-based uncertainty estimates present higher values for patients with smaller lesion load. We thus examine in Section 3.2 how uncertainty varies with lesion size in a comparative analysis, grouping the lesions according to their volume (smaller than 5 mL, between 5 mL and 15 mL and bigger than 15 mL).

2.4 Datasets

This retrospective study analyzed two WMH segmentation datasets:

- **White Matter Hyperintensity (WMH) Segmentation Challenge:** The WMH Segmentation Challenge dataset consists of brain MR images (T1 and FLAIR) with manual annotations of WMH. The dataset includes 60 training sets of T1/FLAIR images from three different institutions, annotated by experts in WMH scoring and 110 test sets from five different scanners. The dataset was derived from patients with various degrees of aging-related degenerative and vascular pathologies to ensure generalizability of segmentation methods across scanners and patient variability. The participants had a mean age of approximately 70 years (70.1 $\pm$ 9.3 years), with an equal gender distribution (50% male). WMH burden varied widely, with mean WMH volume of 16.9 $\pm$ 21.6 ml and a mean lesion count of 62 $\pm$ 35 lesions per subject. This dataset was created as part of the WMH Segmentation Challenge, associated with MICCAI 2017, and was active from 2017 to 2022. The challenge aimed to evaluate and compare methods for the automatic segmentation of WMH of presumed vascular origin. Participants trained their models on the provided training data and submitted their methods for evaluation using the unreleased test data. Results of this challenge have been published in [6].
- **3D MR Image Database of Multiple Sclerosis Patients with White Matter Lesion Segmentations (3D-MR-MS):** The 3D-MR-MS dataset [8] comprises magnetic resonance (MR) images from 30 patients with multiple sclerosis (MS), acquired at the University Medical Center Ljubljana. The dataset includes co-registered and bias-corrected T1-weighted (T1W), contrast-enhanced T1-weighted (T1WKS), T2-weighted (T2W), and FLAIR images, as well as corresponding brain masks and intra-study transform parameters. The patients had a median age of 39 years (range: 25 to 64), with a female-to-male ratio of 23:7. The dataset is designed to support research in

automated lesion segmentation for neurodegenerative diseases like MS. Lesion burden varied significantly, with a total of 3316 lesions segmented and an overall lesion volume (total lesion load, TLL) of 567 ml. The median lesion volume per subject was 15.2 ml (range: 0.337–57.5 ml, interquartile range: 31.1 ml). Lesion sizes ranged from 2 μl to 250 μl (5th to 95th percentile).

Our analysis utilized existing MRI scans and their corresponding manual WMH segmentations to develop and evaluate the proposed methods for uncertainty estimation in WMH segmentation. To ensure consistency, we applied identical preprocessing steps to both datasets, including resampling images to match their spatial resolutions, z-score intensity standardization, and N4 bias field correction (already provided for the 3D-MR-MS dataset).

2.5 WMH segmentation model details

For all experiments in this study, we employed a 3D U-Net architecture [18] for WMH segmentation, implemented using the MONAI framework [1]. The model was designed for 3D volumetric MRI data, accepting two input channels (FLAIR and T1-weighted images) and producing two output channels representing the background and WMH classes. The network consisted of four downsampling/upsampling levels, with feature channels set to (8, 16, 32, 64). Downsampling was achieved using 2×2×2 strided convolutions. A dropout rate of 0.2 was applied within the network during training for regularization. The model was optimized with the Adam method, using an initial learning rate of 0.001 and no weight decay. Training was patch-based, with a batch size of 64 patches of size 32×32×32 voxels extracted from the input volumes, and proceeded for up to 800 epochs. Regularization weights for the different regularization terms were selected individually for each strategy through grid search, balancing segmentation performance and the quality of uncertainty estimation. Inference was also performed using a patch-based sliding window approach with the same patch size (32×32×32), aggregating predictions to reconstruct full-volume segmentations. This standardized model and preprocessing configuration (described in Section 2.4) provided a robust baseline, allowing for the evaluation of regularization strategies on model performance, calibration, and uncertainty estimation under domain shift.

3 Results

In this section, we empirically evaluate the proposed methods, investigating the relationship between model confidence, segmentation quality, and robustness of the model when exposed to OOD samples. We use the previously discussed White Matter Hyperintensity (WMH) Segmentation Challenge dataset (which is considered to be ID) and the 3D MR Image Database of Multiple Sclerosis Patients (3D-MR-MS), considered to be OOD.

3.1 Entropy as a proxy for error prediction in domain shift scenarios

We evaluated the relationship between segmentation performance and uncertainty estimates by analyzing the Pearson correlation between average foreground entropy (as described in Section 2.3.3) and Dice scores across scans. Figure 3 presents scatter plots of entropy as a function of Dice for both ID and OOD data across the four strategies: cross-entropy (CE), CE regularized with Maximum Entropy on Erroneous Predictions (CE_{MEEP}), CE regularized with Kullback-Leibler divergence (CE_{KL}), and CE with Maximum Entropy on All Predictions (CE_{MEALL}). Linear regression lines are fitted to each set of data points, revealing distinct trends for each loss function, regardless of the medical center. The Pearson correlation coefficient is provided for each loss function.

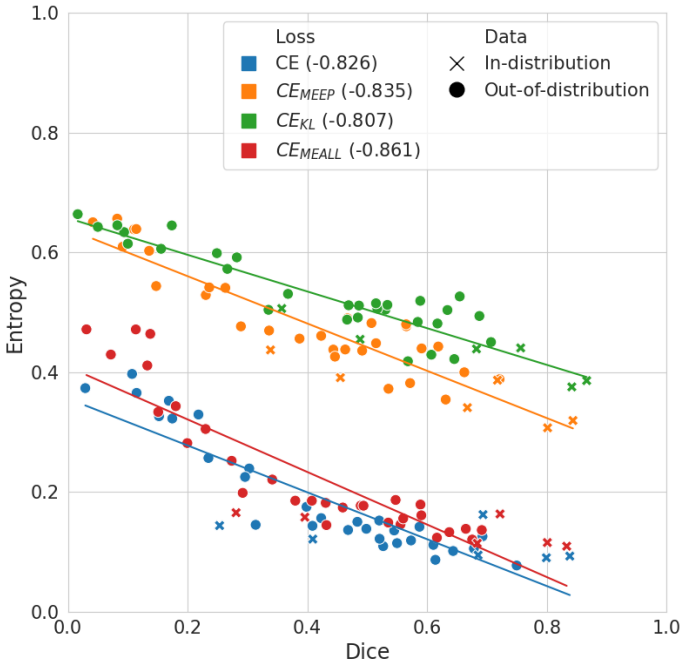


Fig. 3. Scatter plot comparing entropy of foreground predictions and Dice coefficient, per image, for ID and OOD patients. Pearson correlation coefficient between entropy and Dice is shown in parenthesis in the legend box. It can be observed that entropy estimates for MEEP and KL yield better anti-correlation, thus serving as predictors of potential failures.

Although the differences in Pearson correlation coefficients are not very large, consistent trends are observed: regularization methods targeting uncertainty improvement (CE_{MEEP} and CE_{KL}) systematically achieve stronger negative cor-

relations between entropy and Dice scores compared to standard cross-entropy (CE). This indicates that entropy-based regularization leads to uncertainty estimates that more reliably reflect segmentation performance, supporting their use as practical predictors of potential failures, particularly under domain shift.

Across all loss functions, a negative correlation is observed between Dice and entropy, indicating that higher segmentation quality is generally associated with lower uncertainty. However, the strength of this correlation varies across loss functions, with CE_{MEEP} and CE_{KL} exhibiting stronger negative correlations (-0.835 and -0.807, respectively) compared to CE (-0.826) and CE_{MEALL} (-0.861). Specifically, the Pearson correlation coefficients between average foreground entropy and Dice were -0.826 for CE, -0.835 for CE_{MEEP}, -0.807 for CE_{KL}, and -0.861 for CE_{MEALL}. This suggests that CE_{MEEP} and CE_{KL} may provide more reliable uncertainty estimates, as their entropy values more closely track the actual segmentation performance.

To further investigate the behavior of uncertainty estimates under domain shift, we examine their distribution across different types of prediction errors. Figure 4 presents a scatterplot where each point represents a voxel, color-coded based on the classification outcome: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In addition, out-of-distribution (OOD) data are indicated by blue bars, while in-distribution (ID) data are shown with orange bars.

In the case of TP, CE and CE_{MEALL} exhibit the lowest uncertainty, while CE_{MEEP} and CE_{KL} yield higher uncertainty, both for ID (blue) and OOD (orange) cases. For TN, a similar behavior is observed, although less dispersed, with uncertainty medians close to zero for all methods. As expected, FP exhibit higher uncertainties since the model is making incorrect predictions. Notably, CE_{MEEP} and CE_{KL} offer the highest uncertainty for these cases, both in and out of distribution. This heightened uncertainty for FP is desirable, as it allows for the identification of potentially erroneous segmentations, particularly in the challenging OOD setting where the model is more likely to find unfamiliar data distributions. FP often occur in regions with ambiguous image characteristics, making it difficult for the model to confidently distinguish them from TP. Finally, for FN, CE_{MEEP} and CE_{KL} again show higher uncertainty, indicating their ability to express doubt when the model is incorrect.

To gain deeper insights into how maximum entropy regularizers affect the uncertainty estimates, we first analyze entropy levels across ID and OOD data, as shown in Figure 5. As stated previously, the outcomes display two distinctive patterns: standard cross-entropy (CE) and CE_{MEALL} exhibit lower entropy values, (i.e. which translate into higher confidence in their predictions). Conversely, CE_{MEEP} and CE_{KL} demonstrate elevated entropy levels, particularly for OOD data, suggesting increased sensitivity to domain shift and a greater ability to capture uncertainty in challenging scenarios. A Mann-Whitney U test confirms this observation, revealing statistically significant differences in entropy levels between ID and OOD samples for CE_{MEEP} and CE_{KL}, further supporting their effectiveness in distinguishing between the two scenarios. This ability to

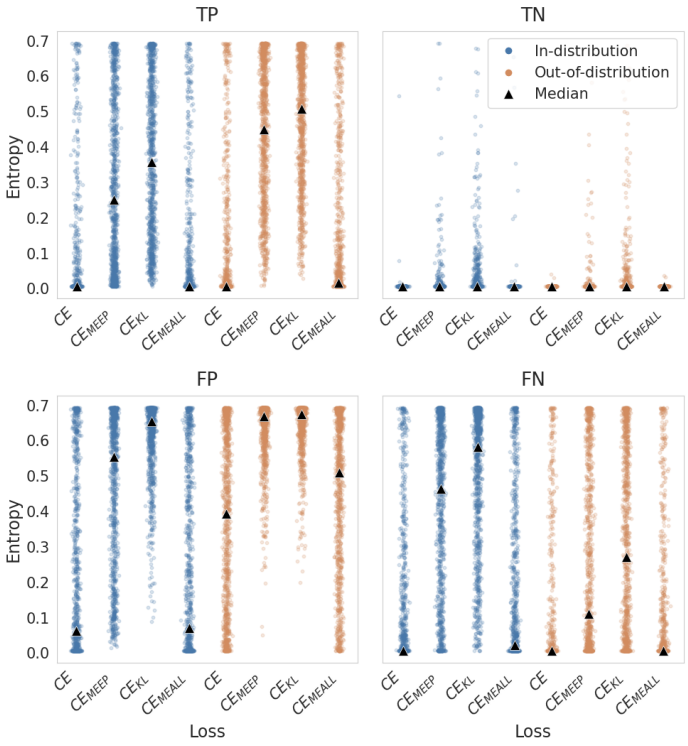


Fig. 4. Distribution of uncertainty estimates across different prediction outcomes (True Positives, True Negatives, False Positives, False Negatives) for various training strategies under ID and OOD scenarios. Each point represents a voxel, with blue indicating ID data and orange representing OOD data. The x-axis shows different training strategies, while the y-axis represents entropy values. Black triangles denote median entropy values. This visualization allows for comparison of uncertainty behaviors across different loss functions, revealing how methods like CE_{MEEP} and CE_{KL} tend to yield higher uncertainties, particularly for false positives and false negatives, in both ID and OOD settings.

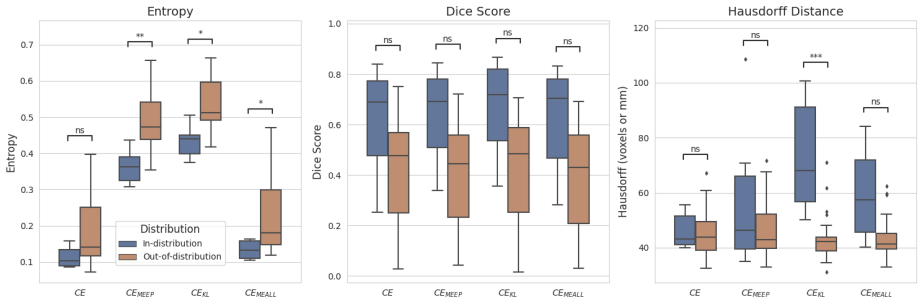


Fig. 5. Boxplots comparing metrics across in-distribution (ID) and out-of-distribution (OOD) data for different loss functions. (Left): Average entropy for voxels predicted as positive, showing a general increase in uncertainty under domain shift, especially for CE_{MEEP} and CE_{KL}. (Middle): Dice score performance across loss functions, with ID scores consistently higher than OOD scores. (Right): Hausdorff distances illustrating boundary localization performance across ID and OOD cases. Statistical significance is indicated where applicable according to the Mann–Whitney U test.

differentiate between ID and OOD data based on uncertainty estimates is crucial for identifying unreliable predictions and ensuring the model's robustness in real-world clinical settings.

3.2 Uncertainty and lesion size analysis

Figure 6 shows that smaller lesions tend to have higher entropy across all loss functions. This observation aligns with the difficulty of reaching expert consensus on ground-truth labels for smaller lesions, as their subtle appearance can make them difficult to identify and delineate. Larger lesions are generally associated with lower entropy values, indicating higher model confidence, and this tendency is consistently observed for both ID and OOD cases.

Quantitatively, with CE the median entropy for small lesions ($< 5mL$) was approximately 0.58, compared to 0.23 for large lesions ($> 15mL$), illustrating the decrease in model uncertainty with increasing lesion size. Notably, the CE_{MEEP} regularization strategy specifically targets these smaller lesions by pushing uncertainty levels toward the maximum, reflecting the inherent ambiguity and potential for disagreement in these cases. This targeted approach could be particularly valuable in clinical practice, as it allows the model to flag its own limitations and prompt further investigation or consultation for uncertain, small lesions.

3.3 Model calibration in domain-shift scenarios

Finally, to assess the impact of domain shift on model calibration, we analyze reliability diagrams and ECE for each loss function considering both ID and OOD scenarios (Figure 7). In the ID scenario, CE_{MEEP} outperforms other losses in terms of Expected Calibration Error (ECE), while in the OOD scenario, all

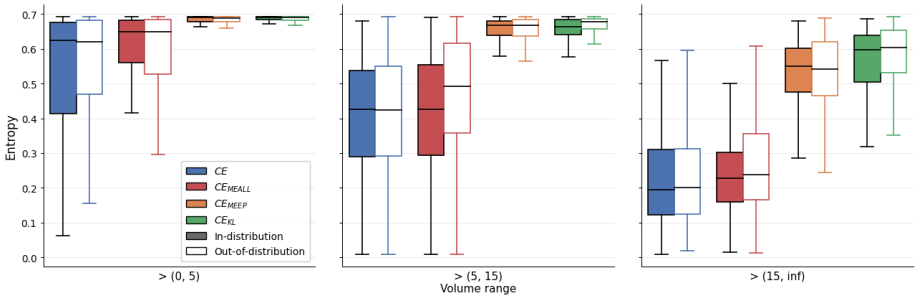


Fig. 6. Boxplots comparing average entropy for voxels predicted as positive across different strategies in three lesion volume ranges. The plot distinguishes between ID (filled boxes) and OOD (unfilled boxes) data. We observe that larger lesion volumes are generally associated with lower entropy, confirming that it can serve as an indicator of model uncertainty. Notably, this tendency is conserved for both ID and OOD cases.

loss functions exhibit poorer calibration, except for the KL-based loss, which demonstrates superior calibration and robustness to domain shift.

4 Discussion

In this study, we investigated the impact of domain shift on model calibration and uncertainty estimation in white matter hyperintensity (WMH) segmentation. Our findings demonstrate that entropy-based uncertainty estimates could be used as a proxy for anticipating segmentation errors in unseen domains. Specifically, we observed a significant correlation between increasing segmentation errors due to domain shifts and rising entropy-based uncertainty estimates. By incorporating maximum-entropy regularization techniques, such as CE_{MEEP} and CE_{KL}, we further strengthened this correlation and improved model calibration.

Our analysis also revealed that the choice of loss function significantly influences the uncertainty quantification quality. While standard cross-entropy and CE_{MEALL} loss functions tend to produce lower entropy values, CE_{MEEP} and CE_{KL} yield higher uncertainty levels, particularly for OOD data. This suggests that CE_{MEEP} and CE_{KL} are more sensitive to domain shifts and better at capturing uncertainty in challenging scenarios. Additionally, our investigation into the relationship between lesion size and uncertainty revealed that smaller lesions tend to have higher uncertainty across all loss functions. This finding highlights the importance of considering lesion size when interpreting model predictions and emphasizes the need for further research into uncertainty estimation for small lesions. Models trained with maximum-entropy regularization achieved lower ECE values compared to standard training, further confirming the effectiveness of entropy-based regularization for maintaining reliable probabilistic outputs across distributions.

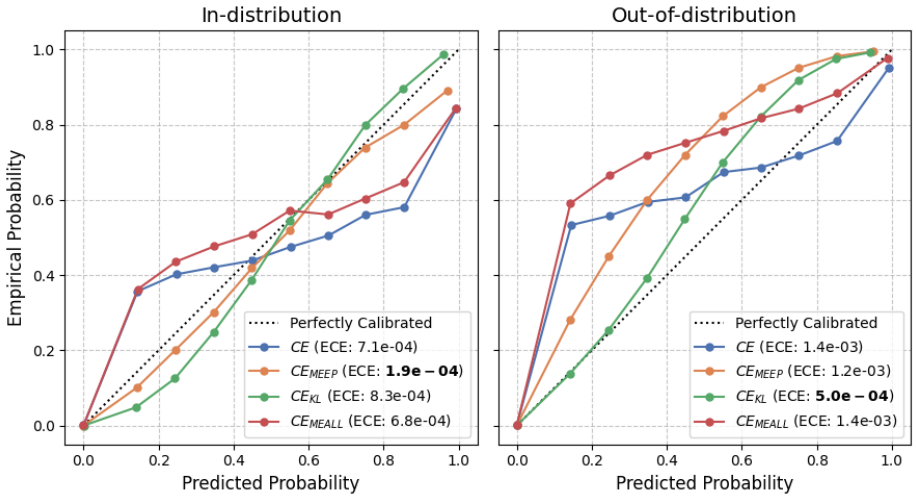


Fig. 7. Reliability plots for different loss functions on ID and OOD data. Each colored line corresponds to a different loss function, with the ECE shown in parentheses (best ones are shown in bold). Points above the diagonal indicate underconfidence, while points below indicate overconfidence. A well-calibrated model should approximate the dashed diagonal line (representing perfect calibration).

The analysis of prediction outcomes showed that uncertainty levels were higher for incorrect predictions (false positives and false negatives) in regularized models, especially under domain shift. This behavior is desirable in clinical practice, as it helps to identify unreliable segmentations and regions that may require expert review, enhancing the interpretability and safety of models. Notably, maximum-entropy regularization amplified uncertainty, particularly in smaller lesions, aligning model uncertainty with regions of greater clinical ambiguity. This could be valuable for detecting subtle or borderline lesions, which are typically harder to segment accurately.

In conclusion, our study underscores the importance of uncertainty estimation and model calibration in mitigating the challenges posed by domain shift in medical image analysis. By incorporating maximum-entropy regularization techniques and carefully considering the choice of loss function, more robust and reliable deep learning models for WMH segmentation can be developed. These strategies not only improve segmentation performance but also provide better indicators of prediction confidence, which are essential for safe clinical deployment in multi-center and heterogeneous imaging environments. Future work could extend this analysis by evaluating a-entropy regularization across different segmentation architectures, providing deeper support to the robustness and generalizability of these techniques.

Acknowledgements

The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, the support of Universidad Nacional del Litoral with the CAID program and Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación for the support with the PICT program. EF was supported by the Google Award for Inclusion Research (AIR) Program. VFM was partially supported by the Emerging Leaders in the Americas Program (ELAP) program. We also thank Calcul Quebec and Compute Canada.

Data Availability Statement

The datasets used in this study are publicly available:

1. The White Matter Hyperintensity (WMH) Segmentation Challenge dataset is available at `https://wmh.isi.uu.nl/` under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).
2. The 3D MR Image Database of Multiple Sclerosis Patients with White Matter Lesion Segmentations (3D-MR-MS) is available at `https://lit.fe.uni-lj.si/en/research/resources/3D-MR-MS/` under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

References

1. Cardoso, M.J., Li, W., Brown, J.M., Maier-Hein, K., Dawn, T., Murrey, N., ..., Glocker, B.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022), `https://arxiv.org/abs/2211.02701`
2. Chaves, H., Serra, M., Shalom, D., et al.: Assessing robustness and generalization of a deep neural network for brain ms lesion segmentation on real-world data. European Radiology **34**, 2024–2035 (2024). https://doi.org/10.1007/s00330-023-10093-5
3. Czolbe, L., Sedghi, A., Kirschke, J.S., Zimmer, C., Wiestler, B., Menze, B.H.: Evaluation of uncertainty estimation methods for brain tumor segmentation with deep learning. Frontiers in Neuroscience **15**, 680645 (2021). https://doi.org/10.3389/fnins.2021.680645
4. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. ICML'16, vol. 48, pp. 1050–1059. JMLR.org (2016)
5. Kohl, S.A.A., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.M.A., Rezende, D.J., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. In: Advances in Neural Information Processing Systems. vol. 31, pp. 6965–6975. Curran Associates, Inc. (2018), `https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf`

6. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R.J., Andermatt, S., Bento, M., ..., Biessels, G.J.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE Transactions on Medical Imaging **38**(11), 2556–2568 (2019). https://doi.org/10.1109/TMI.2019.2905770
7. Larrazabal, A., Martínez, C., Dolz, J., Ferrante, E.: Maximum entropy on erroneous predictions: Improving model calibration for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol. 14222, pp. 273–283. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43898-1_27
8. Lesjak, Z., Galimzianova, A., Spiclin, Z., Pernus, F., Likar, B., Sersa, I.: A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics **16**, 51–63 (2018). https://doi.org/10.1007/s12021-017-9353-3
9. Mehrtash, A., Wells, W., Tempany, C., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging **39**(12), 3868–3878 (2020). https://doi.org/10.1109/TMI.2020.3006437
10. Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016). https://doi.org/10.1109/3DV.2016.79
11. Mosquera, C., Ferrer, L., Milone, D., Luna, D., Ferrante, E.: Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance. European Radiology pp. 1–9 (2024). https://doi.org/10.1007/s00330-024-10834-0
12. Murugesan, B., Liu, B., Galdran, A., Ayed, I., Dolz, J.: Calibrating segmentation networks with margin-based label smoothing. Medical Image Analysis **87**, 102826 (2023). https://doi.org/10.1016/j.media.2023.102826
13. Nair, T., Precup, D., Arnold, D., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical Image Analysis **59**, 101557 (2020). https://doi.org/10.1016/j.media.2019.101557
14. Ovadia, Y., Fertig, E., Ren, J., et al.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf
15. Palladino, J., Slezak, D., Ferrante, E.: Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter mr images. In: 16th International Symposium on Medical Information Processing and Analysis. vol. 11583, p. 1158302. SPIE (2020). https://doi.org/10.1117/12.2579548
16. Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., Hinton, G.: Regularizing neural networks by penalizing confident output distributions. In: ICLR Workshop (2017), https://openreview.net/forum?id=HyhbYrGYe
17. Ricci Lara, M., Mosquera, C., Ferrante, E., Echeveste, R.: Towards unraveling calibration biases in medical image analysis. In: Wesarg, S., et al. (eds.) Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging, Lecture Notes in Computer Science, vol. 14242. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-45249-9_13

18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Sambyal, A., Niyaz, U., Krishnan, N., Bathula, D.: Understanding calibration of deep neural networks for medical image classification. Computers in Biology and Medicine **242**, 107816 (2023). https://doi.org/10.1016/j.cmpb.2023.107816
20. Tran, P., Thoprakarn, U., Gourieux, E., et al.: Automatic segmentation of white matter hyperintensities: validation and comparison with state-of-the-art methods on both multiple sclerosis and elderly subjects. NeuroImage: Clinical **33**, 102940 (2022). https://doi.org/10.1016/j.nicl.2022.102940
21. Yang, R., Yu, Y.: Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Frontiers in Oncology **11**, 638182 (2021). https://doi.org/10.3389/fonc.2021.638182
22. Yeung, M., Rundo, L., Nan, Y., Sala, E., Schönlieb, C., Yang, G.: Calibrating the dice loss to handle neural network overconfidence for biomedical image segmentation. Journal of Digital Imaging **36**(3), 739–752 (2023). https://doi.org/10.1007/s10278-022-00735-3