

Article

NornirNet: A Deep Learning Framework to Distinguish Benign from Malignant Type II Endoleaks After Endovascular Aortic Aneurysm Repair Using Preoperative Imaging

Francesco Andreoli ^{1,†}, Fabio Mattiussi ^{2,†}, Elias Wasseh ³, Andrea Leoncini ², Ludovica Ettore ¹, Jacopo Galafassi ¹, Maria Antonella Ruffino ², Luca Giovannacci ¹, Alessandro Robaldo ^{1,†} and Giorgio Prouse ^{1,*,†}

¹ Division of Vascular Surgery and Angiology, Centro Vascolare Ticino, Ente Ospedaliero Cantonale (EOC), 6900 Lugano, Switzerland

² Imaging Institute of Southern Switzerland (IIMSI), Ente Ospedaliero Cantonale (EOC), 6900 Lugano, Switzerland

³ Faculty of Biomedical Science, Università della Svizzera Italiana (USI), 6900 Lugano, Switzerland

* Correspondence: giorgio.prouse@eoc.ch; Tel.: +41-918116603

† These authors contributed equally to this work.

Abstract

Background/Objectives: Type II endoleak (T2EL) remains the most frequent complication after endovascular aortic aneurysm repair (EVAR), with uncertain clinical relevance and management. While most resolve spontaneously, persistent T2ELs can lead to sac enlargement and rupture risk. This study proposes a deep learning framework for preoperative prediction of T2EL occurrence and severity using volumetric computed tomography angiography (CTA) data. **Methods:** A retrospective analysis of 277 patients undergoing standard EVAR (2010–2023) was performed. Preoperative CTA scans were processed for volumetric normalization and fed into a 3D convolutional neural network (CNN) trained to classify patients into three categories: no T2EL, benign T2EL, or malignant T2EL. The model was trained on 175 cases, validated on 72, and tested on an independent cohort of 30 patients. Performance metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). **Results:** The CNN achieved an overall accuracy of 76.7% (95% CI: 0.63–0.90), a macro-averaged F1-score of 0.77, and an AUC of 0.93. Class-specific AUCs were 0.93 for no T2EL, 0.91 for benign, and 0.96 for malignant cases, confirming high discriminative capacity across outcomes. Most misclassifications occurred between adjacent categories. **Conclusions:** This study introduces the first end-to-end 3D CNN capable of predicting both the presence and severity of T2EL directly from preoperative CTA, without manual segmentation or handcrafted features. These findings suggest that preoperative imaging encodes latent structural information predictive of endoleak-driven sac reperfusion, potentially enabling personalized pre-emptive embolization strategies and tailored surveillance after EVAR.

Keywords: type II endoleak; abdominal aortic aneurysm; endovascular aortic repair; tomography; deep learning; neural networks; artificial intelligence

Academic Editor(s): Hongjian Shi

Received: 27 November 2025

Revised: 13 January 2026

Accepted: 30 January 2026

Published: 4 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Endovascular aortic aneurysm repair (EVAR) has emerged as the primary treatment option for abdominal aortic aneurysm (AAA) with a permissive anatomy [1]. Despite the early benefits, endoleaks remain the most common complications after EVAR, and the risk of related reinterventions remains significant [2]. Type I and III endoleaks are due to direct sac perfusion: Type I arises from inadequate sealing between the endograft and the healthy vessel wall, whereas Type III results from structural defects or component separations within the endograft [3,4]. These high-pressure leaks are known to be related to a significant rupture risk and have a strong indication for surgical correction [1,5]. In contrast, the management and clinical relevance of type II endoleak (T2EL), which is caused by retrograde sac feeding by branch vessels, remain controversial (Figure 1). Many T2ELs resolve spontaneously, but a subset of patients develop persistent T2EL, and recent European Society for Vascular Surgery (ESVS) guidelines consider reintervention only in the presence of ≥ 10 mm aneurysm sac growth compared with baseline or with the smallest diameter during follow-up using the same imaging modality and measurement method [1]. Pre-emptive embolization of the side branches has been proposed to decrease the risk of T2EL and, consequently, possible aneurysm growth with sealing loss [6,7]. However, these techniques increase procedure time, cost, and risk of complications. Even if current literature suggests a positive effect of pre-emptive embolization of side branches on T2EL and related reintervention rate, the clinical benefit of this procedure has not been clearly confirmed [6–8]. The current guidelines suggest considering the pre-emptive feeding vessel embolization only in very selected cases [1]. Over the past decade, the extraction of quantitative radiomic features (RF) from medical images has gained extensive attention for its ability to provide detailed quantitative signatures of disease states, enabling the diagnosis and prediction of treatment outcomes [9–11]. In the field of artificial intelligence (AI), deep learning algorithms can be used to analyze entire volumes of computed tomography (CT) images, providing decision support tools aiding patient management [12]. To date, no comprehensive deep learning analysis has been performed for the prediction of T2EL causing sac enlargement based on preoperative CT scans. Such models may be helpful to predict which patients could benefit from pre-emptive treatment to prevent T2EL-related sac growth.



Figure 1. Schematic representation of the Type II Endoleak mechanism. Reproduced with the courtesy of Medtronic plc. All rights reserved. Arrows: direction of blood flow through side branches causing continued aneurysm sac perfusion.

2. Materials and Methods

2.1. Study Design and Population

This retrospective study included patients who underwent EVAR at a single tertiary-level vascular referral center. Data about all the patients treated from 2010 to 2024 were extracted from a prospectively maintained database and screened for inclusion criteria.

The study was approved by the local Ethics Committee (2025-00343 CE 4781).

This research was in accordance with the World Medical Association Declaration of Helsinki, and the data underlying this article will be shared on reasonable request to the corresponding author.

Written informed consent was obtained for all individuals treated after 2020; for the others, consent was obtained through a no objection letter.

2.2. Inclusion and Exclusion Criteria

The final cohort consisted of 277 patients who fulfilled the following inclusion criteria (Figure 2):

- Age > 18 years.
- Treatment with standard EVAR performed according to the manufacturer's instructions for use for AAA.
- Non-ruptured AAA.
- Minimum follow-up of 1 year.
- Availability of a 1–3-month postoperative computed tomography angiography (CTA).
- Availability of a preoperative CTA.
- For patients with T2EL, at least one additional CTA at least 6 months after the detection.
- Adequate CTA image quality (slice thickness ≤ 2.5 mm), including coverage from the celiac trunk to the external iliac vessels.

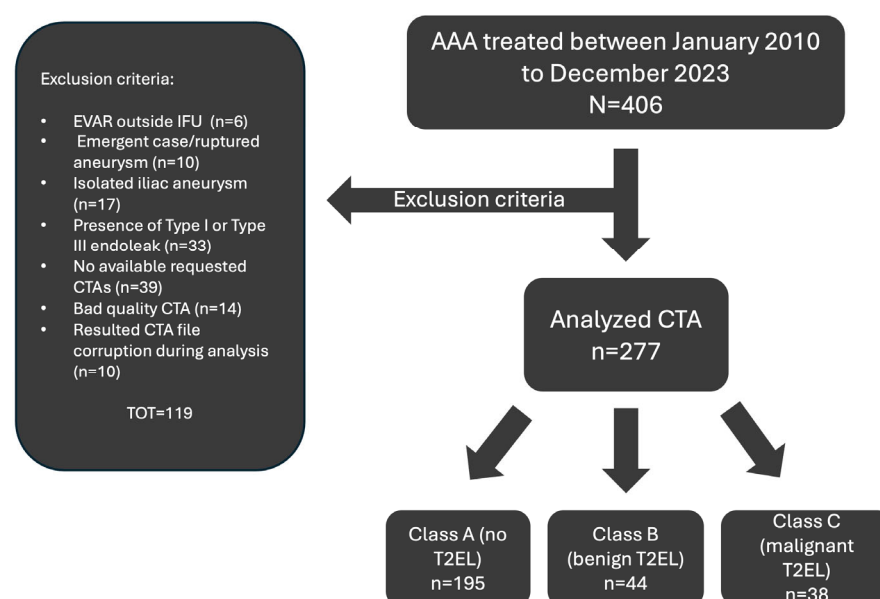


Figure 2. CONSORT-like chart depicting the final included cohort alongside the excluded patients and corresponding reasons for exclusion. Abbreviations: AAA, Abdominal Aortic Aneurysm ; CTA,

Computed Tomography Angiography; EVAR, Endovascular Aneurysm Repair; T2EL, Type II Endoleak.

The minimum clinical and radiological (CTA or duplex ultrasound) follow-up requested was 12 months. Complete documentation of any development of T2EL and their clinical evolution was also required.

Patients with preoperative CT images of inadequate quality or with significant artifacts, incomplete follow-up of less than 12 months, and incomplete clinical data or lack of outcome documentation were excluded from the study.

2.3. Data Collection

The dataset included patients' baseline (age, gender, comorbidities) and contained details from the pre-operative and post-EVAR CT scans at 1–3 months and during follow-up. All the CT scans were acquired at the Imaging Institute of Southern Switzerland (IIMSI), Ente Ospedaliero Cantonale (EOC), Switzerland, using four 128-slice Siemens CT scanners (Somatom Definition Edge, Siemens Healthineers; Erlangen, Germany), one 256-slice Philips CT scanner (Philips Brilliance iCT 256 Slice, Philips Healthineers, Best, The Netherlands), and one 256-slice Siemens CT scanner (Somatom Definition Flash, Siemens Healthineers; Erlangen, Germany).

The acquisition protocol included an arterial phase performed in dual energy followed by a venous phase acquisition limited to the extent of the stent graft, with a reconstructed slice thickness between 0.5 and 2–5 mm per slice [13]. All the CTs were evaluated by two dedicated vascular radiologists with more than 20 years of experience. All CT angiographies with findings outside the normal range or considered uncertain were reviewed in a multidisciplinary board meeting involving radiologists, vascular surgeons, and angiologists [14].

All images in the dataset have been anonymized and converted to a standardized format for processing using deep learning. The images were pre-processed to standardize size and voxel spacing and normalized (mean = 0.196, standard deviation = 0.246), thus ensuring consistency of the data input to the neural network.

The normalization parameters were derived solely from the training cohort and subsequently applied to the validation and test sets.

2.4. CT Angiography Protocol

CT scans were performed using the following acquisition protocol:

- Helical scan from lung apices to lesser trochanters.
- kV = 100; pitch = 1.50; acquisition (mm) = 64×0.60 .
- Bolus tracking 2 cm below the tracheal bifurcation, with the region of interest in the ascending arch.
- Bolus tracking scan delay plus 7 s.
- No ECG synchronization and no patient apnea during the scan.
- Contrast medium injection: 15 mL NaCl at 5.0 mL/s followed by 100 mL Accupaque 350 at 5.0 mL/s and followed by 50 mL NaCl at 5.0 mL/s.

2.5. Endoleak Evaluation

For the study, anonymized CTA scans were independently evaluated by two experienced vascular operators (a vascular surgeon and a vascular radiologist) with more than 40 cases/year experience. The evaluation was carried on with a dedicated analysis software (Syngo.Via, version VB80D, 2023; Siemens Healthineers, Forchheim, Germany).

The maximum sac diameter, the presence of a T2EL, and the type and number of vessels involved in the endoleak were recorded from each CTA examination.

A 5% interobserver discrepancy in measurement was acceptable, with larger variations resolved by a third investigator. Both primary investigators evaluated 10 CTAs twice, with a one-month interval. The intraobserver difference remained within 2% for diameters. The presence or absence of T2EL and the involvement of feeding vessels was assessed at a glance, and the accordance was 100%.

The aneurysm sac growth (D3) was calculated through the formula $D3 = D2 - D1$, where D1 represents the maximum aneurysm sac diameter at the 1–3 months post-operative CTA and D2 represents the maximum sac diameter at the last available CTA scan.

T2ELs were classified as ‘benign’ if they were not associated with aneurysm sac expansion ≥ 5 mm, did not require T2EL-related reintervention, and were not linked to any complications. Conversely, T2ELs were classified as ‘malignant’ if they were associated with sac expansion > 5 mm and/or required reintervention due to related complications [15,16].

All re-interventions were performed for aneurysm growth of more than 5 mm in a year or absolute growth of more than 1 cm, as well as for every case of symptomatic or ruptured AAA due to T2EL [1].

2.6. Patient Stratification

Based on these criteria, patients were stratified into three classes: Class A includes patients without development of T2EL during follow-up, Class B includes patients with “benign” T2EL, and Class C groups patients with “malignant” type II endoleaks.

The dataset was divided into three sets:

- A training set consisting of 175 cases, 141 Class A, 20 class B, and 14 class C.
- A validation set consisting of 72 cases, 44 class A, 14 class B and 14 class C.
- A test set consisting of 30 cases, 10 class A, 10 class B and 10 class C.

2.7. Data Augmentation

To increase the robustness of the model and prevent overfitting, a data augmentation strategy was implemented on the entire preoperative CT scan volume and applied dynamically during the training phase on the training set. Prior to augmentation, all volumes were standardized through isotropic resampling to a uniform voxel resolution of $1 \times 1 \times 1$ mm to ensure anatomical consistency. The augmentation techniques were implemented using the torchvision-transforms library and include random variations in brightness and contrast within a range of 0.8 to 1.2 to simulate variability in CT acquisitions, random image rotations in a range of ± 10 degrees to simulate variability in patient positioning, and random horizontal and vertical flipping with a probability of $p = 0.5$ for each direction. These transformations were applied stochastically during each training epoch, generating a virtually larger and more diverse dataset without introducing significant anatomical distortions.

2.8. Model Architecture

The present deep learning model was based only on the preoperative CTA. The deep learning architecture (Figure 3, details in Appendix A) based on 3D convolutional neural networks (3D CNNs) was developed for the volumetric analysis of preoperative CT angiography images.

The input data consists of pre-processed 3D CT volumes of preoperative CT angiograms with standardized dimensions of $256 \times 512 \times 512$ voxels, normalized in intensity using z-score normalization. Each volume represents the anatomical region of interest containing the abdominal aorta and related vascular structures. The model output is a probabilistic prediction on three mutually exclusive classes: Class A, Class B, and Class C.

The model returns a probability vector for each class, with the predicted class determined by the maximum probability using softmax classification.

The complete architecture of the model is based on a deep 3D CNN network with residual connections (ResNet). The channel expansion follows a $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$ pattern. The network consists of an initial convolutional block and four main residual blocks for feature extraction:

- Initial Block ($1 \rightarrow 32$ channels): a single 3D convolutional layer ($3 \times 3 \times 3$ kernel, padding = 1) followed by batch normalization and GELU activation. The spatial output dimensions remain $256 \times 512 \times 512$.
- Residual Block 1 ($32 \rightarrow 64$ channels): introduces the first dimensional reduction. It consists of two convolutional layers ($32 \rightarrow 32$ and $32 \rightarrow 64$) and a $1 \times 1 \times 1$ shortcut connection ($32 \rightarrow 64$). Downsampling is achieved via an anisotropic stride (1, 2, 2) in both the main convolution and the shortcut. Includes batch normalization, GELU, and 3D dropout ($p = 0.3$). Output dimensions: $256 \times 256 \times 256$.
- Residual Block 2 ($64 \rightarrow 128$ channels): follows a similar residual structure (conv $64 \rightarrow 64$ and $64 \rightarrow 128$; shortcut $64 \rightarrow 128$) using a stride (2, 2, 2) for downsampling. Includes batch normalization, GELU, and 3D dropout ($p = 0.3$). Output dimensions: $128 \times 128 \times 128$.
- Residual Block 3 ($128 \rightarrow 256$ channels): expands channels to 256 (conv $128 \rightarrow 128$ and $128 \rightarrow 256$; shortcut $128 \rightarrow 256$) with stride (2, 2, 2). Dropout is increased to $p = 0.4$. Output dimensions: $64 \times 64 \times 64$.
- Residual Block 4 ($256 \rightarrow 512$ channels): final block (conv $256 \rightarrow 256$ and $256 \rightarrow 512$; shortcut $256 \rightarrow 512$) with stride (2, 2, 2) and dropout ($p = 0.4$). Output dimensions: $32 \times 32 \times 32$.
- Pooling: the final feature maps (512 channels) are processed by an Adaptive Average Pooling 3D layer that reduces the output to a single 512-dimensional vector.

The extracted features are then processed through a fully connected classifier (MLP) consisting of three dense layers ($512 \rightarrow 256 \rightarrow 128 \rightarrow 3$) with GELU activation functions and dropout regularization ($p = 0.4$ for the first layer, $p = 0.3$ for the second). The output layer uses softmax activation for multi-class probabilistic prediction.

Batch normalization is applied after each 3D convolutional layer (including in shortcuts) to stabilize training and accelerate convergence. GELU (Gaussian Error Linear Unit) activation functions are used throughout the network, providing smoother gradients [17]. Dropout regularisation increases progressively in the deeper convolutional layers (from 0.3 to 0.4) to prevent overfitting [18]. The model was trained using the categorical cross-entropy loss function, the Adam optimizer with adaptive learning rate, and early stopping based on performance on the validation set to prevent overfitting [19].

To overcome the numerical disparity between Class A ($n = 141$), Class B ($n = 20$), and Class C ($n = 14$) in the training set, a customized stratified clustering strategy was employed. During each epoch, training batches were constructed to maintain a balanced 1:1:1 ratio of samples from each category. This sampling method ensures that the 3D CNN is exposed to minority classes (benign and malignant T2EL) with the same frequency as the majority class, preventing categorical cross-entropy loss from becoming biased toward the “No T2EL” outcome.

The final configuration of NornirNet, including the use of GELU activation and progressive dropout, was determined through a systematic hyperparameter optimization phase. Initial experiments compared GELU with ReLU and LeakyReLU, identifying that GELU provided smoother gradient propagation and greater stability. Similarly, progressive dropout (increasing from 0.3 in the first blocks to 0.4 in the deeper layers) was preferred over fixed dropout rates, as it demonstrated superior performance in mitigating overfitting within the high-dimensional feature maps of the deeper residual blocks. The

model was trained using categorical cross-entropy loss and the Adam optimizer with an adaptive learning rate.

2.9. Hardware and Software Configuration

The network training was conducted on a system equipped with the Ubuntu 24.04 operating system, an Intel Core i9 14900 k processor, 64 GB of DDR5 RAM, and an NVIDIA RTX 4090 GPU with a total of 24 GB of VRAM. The model was implemented utilizing Python 3.12.3, PyTorch 2.3.0, NVIDIA drivers 570.124.06, and NVIDIA CUDA Toolkit 12.8.

2.10. Statistical Analysis and Model Evaluation

The primary purpose of the analysis was to evaluate the ability of the 3D CNN model to correctly classify patients into three predefined categories: Class A (no T2EL), Class B (benign T2EL), and Class C (malignant T2EL).

The model's performance was evaluated exclusively on the independent test set ($n = 30$), which was not used during the training or validation phases.

Key performance metrics for multi-class classification included:

- Overall accuracy: the proportion of correct predictions out of the total.
- Precision, Recall, and F1-Score: these metrics were calculated both for each individual class and as a weighted average (macro-averaged) to provide a balanced assessment of performance across classes, especially in the presence of imbalance in the training dataset.
- Area Under the Receiver Operating Characteristic Curve (ROC): the AUC was calculated for each class (one-vs-rest) to measure the model's discriminatory power.
- Confusion Matrix: a confusion matrix was generated to analyze classification errors (e.g., false positives and false negatives) between different classes in detail.

The distribution of continuous variables was examined using the Shapiro-Wilk test. For normally distributed variables, mean and standard deviation (SD) were reported, and subgroup comparisons were conducted using Student's t-test. Skewed variables were presented as median [with interquartile ranges (IQR)], and the Wilcoxon Rank-Sum test was employed to assess differences. For categorical variables, counts and percentages were provided, and subgroup differences were evaluated using the chi-square test. The distribution of continuous variables across the training, test, and validation set was evaluated with one-way ANOVA for the normally distributed ones and with the Kruskal-Wallis test for the skewed. Categorical variable distribution was evaluated with the chi-square test. Incidence of overall T2EL, benign T2EL, and malignant T2EL was analyzed through the Kaplan-Meier cumulative incidence function.

For all key metrics, 95% confidence intervals (95CI) were calculated to quantify the uncertainty of the estimate. Statistical significance was set at $p < 0.05$.

Analyses were conducted using Python (version 3.10.0), STATA 19 (StataCorp., College Station, TX, USA) and Matplotlib (version 3.7.3).

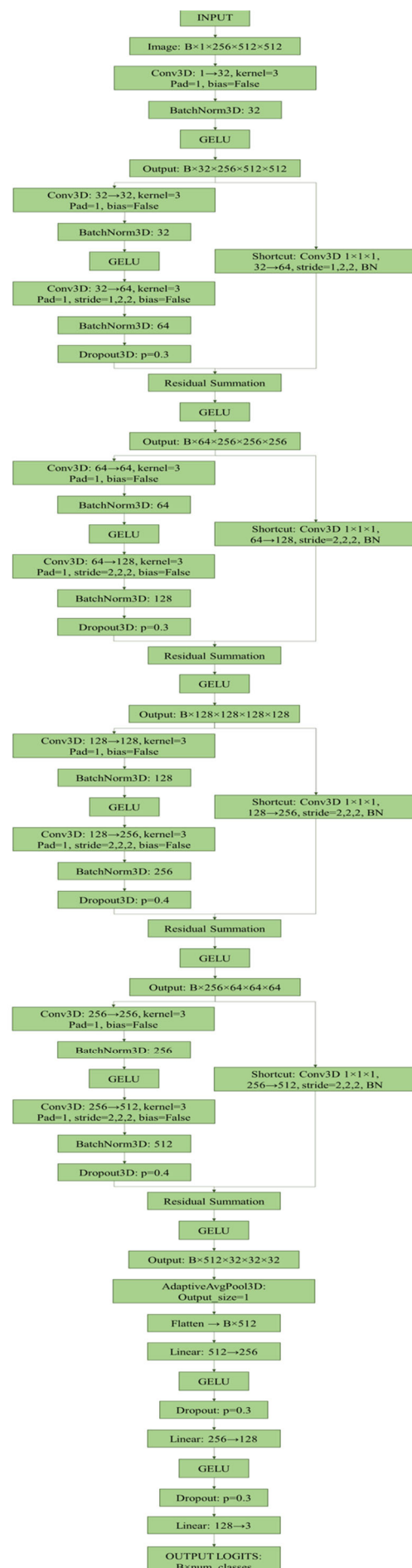


Figure 3. The flowchart shows the complete architecture of the convolutional neural network for the classification of endoleaks from CT volumes. The architecture is a residual convolutional 3D network (ResNet) that begins with an initial convolutional block and continues with four sequential residual blocks. These blocks progressively expand the channels ($1 \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$) and reduce the spatial dimensions of the volumetric input through stride convolutions. The extracted features are then passed to a fully connected three-layer classifier (MLP) that produces the final logits for the three classes. GELU activation functions and dropout regularization are used to prevent overfitting. Abbreviations: GELU, Gaussian Error Linear Unit.

3. Results

3.1. Baseline

The enrolled patients were mostly men ($n = 244/277$) with a median age of 78 [72–82].

According to the study protocol, all the patients had a CTA at 1–3 months after the procedure, at which time 62 T2EL were diagnosed. Among them, 7 (11.3%) were present at the final intraoperative angiography.

The median follow-up of the entire cohort was 55.5 months [28.03–91.6], and during this time a total of 82 (29.6%) T2EL in single patients were recorded: 38 (46.3%) were classified as “malignant” and 44 (53.7%) as “benign”. Incidence over time of the total T2EL and malignant T2EL were reported in the Kaplan-Meier curves displayed in Figure 4.

The median time from EVAR to endoleak was 35 [28–113] days with no difference between malignant and benign ones ($p = 0.38$).

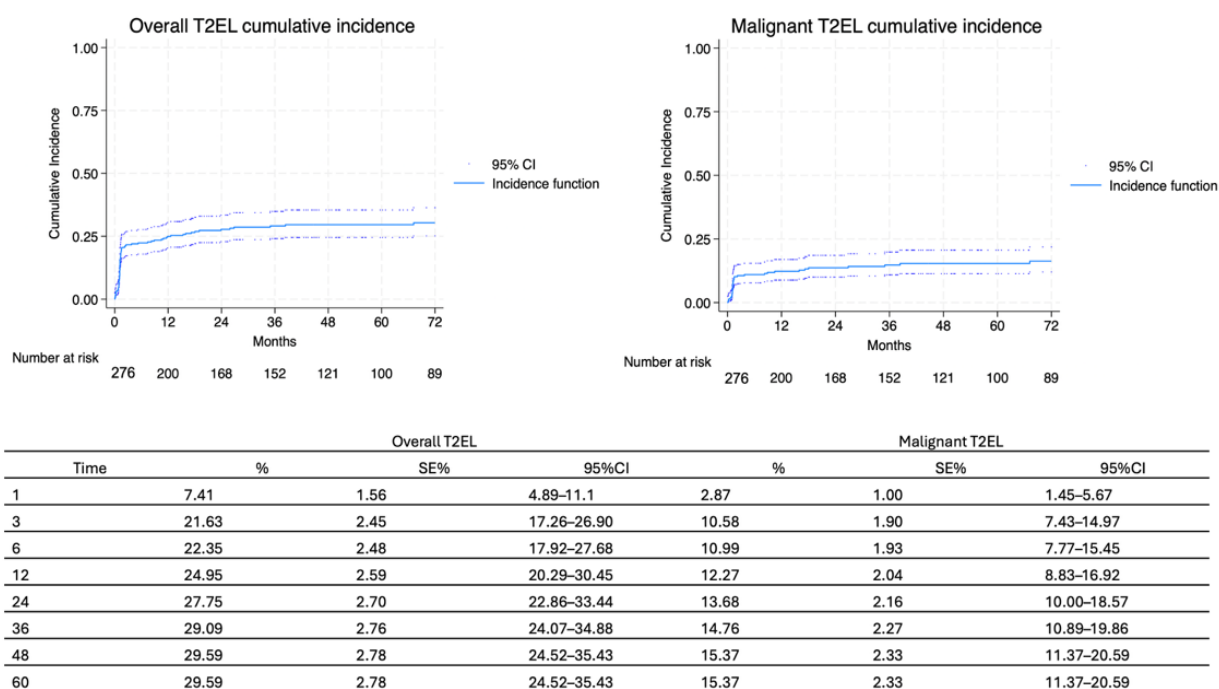


Figure 4. Kaplan-Meier incidence function of overall T2EL and malignant. Abbreviations: 95CI, 95% Confidence Interval; SE%, Standard Error; T2EL, Type II endoleak.

3.2. Overall Model Performance

The deep learning model developed was evaluated on an independent test set consisting of 30 patients (10 for each class) kept separate during all training and validation phases. Baseline comparison of the patients across the training, test and validation set was reported in Supplementary Table S1.

The overall results (Table 1) show that the model achieved an accuracy of 76.67%, corresponding to the correct classification of 23 out of 30 patients.

The macro-averaged metrics, which provide a balanced assessment across all classes, reported an F1-Score of 0.766, a Precision of 0.782, and a Recall of 0.767. The macro-averaged AUC ROC was equal to 0.933 [95CI 0.8715–0.9801], demonstrating the model's excellent overall discriminatory capacity.

Table 1. Overall performance of the model on the independent test set. The metrics reported are macro-averaged across the three classes and reflect the overall effectiveness of the model in the test set.

Metric	Result	95CI
Accuracy	0.766	0.633–0.900
F1-Score	0.766	0.594–0.904
Precision	0.782	0.631–0.926
Recall	0.767	0.611–0.915
AUC ROC	0.933	0.871–0.980

Abbreviations: 95CI, 95% Confidence Interval; AUC, Area Under the Curve; ROC, Receiver Operating Characteristic Curve.

3.3. Performance per Class

Analysis of class-specific metrics revealed variations in prediction performance (Table 2).

Table 2. Classification Metrics Matrix (by Class). The table shows the performance metrics (Precision, Recall/Sensitivity, F1-Score, and AUC ROC) of the predictive model, divided into three target classes: No T2EL, Benign T2EL, and Malignant T2EL.

Class Name	Precision (95% CI)	Recall (95% CI)	F1-Score (95% CI)	AUC ROC (95% CI)
No T2EL	0.692 (0.428–0.937)	0.900 (0.699–1.000)	0.7826 (0.545–0.947)	0.930 (0.821–1.000)
Benign T2EL	0.7778 (0.500–1.000)	0.700 (0.400–1.000)	0.7368 (0.470–0.933)	0.910 (0.788–1.000)
Malignant T2EL	0.8750 (0.600–1.000)	0.700 (0.375–1.000)	0.777 (0.500–0.960)	0.960 (0.875–1.000)

Abbreviations: 95CI, 95% Confidence Interval; AUC, Area Under the Curve; ROC, Receiver Operating Characteristic Curve; T2EL, Type II Endoleak.

- Class A (No T2EL): The Precision was 0.692, and the F1-Score was 0.783. Only one patient in this class was a False Negative (FN), while there were 4 False Positives (FP). The ROC AUC was 0.930.
- Class B (Benign T2EL): The classification of T2ELs showed an Accuracy of 0.778, a Recall of 0.700, and an F1-Score of 0.737. The model correctly identified 7 out of 10 patients. Class B was the only one with 3 false negatives and 2 false positives. The ROC AUC was 0.910.
- Class C (Malignant T2EL): The prediction of malignant T2ELs, often considered 'malignant' in the follow-up context, achieved the highest accuracy among all classes at 0.875. The recall was 0.700 and the F1-score 0.778. Seven out of ten patients were correctly identified, with only one false positive and three false negatives. This class also had the best ROC AUC at 0.960.

To visualize the discriminatory capacity of the model, the ROC curves for each class are shown in Figure 5.

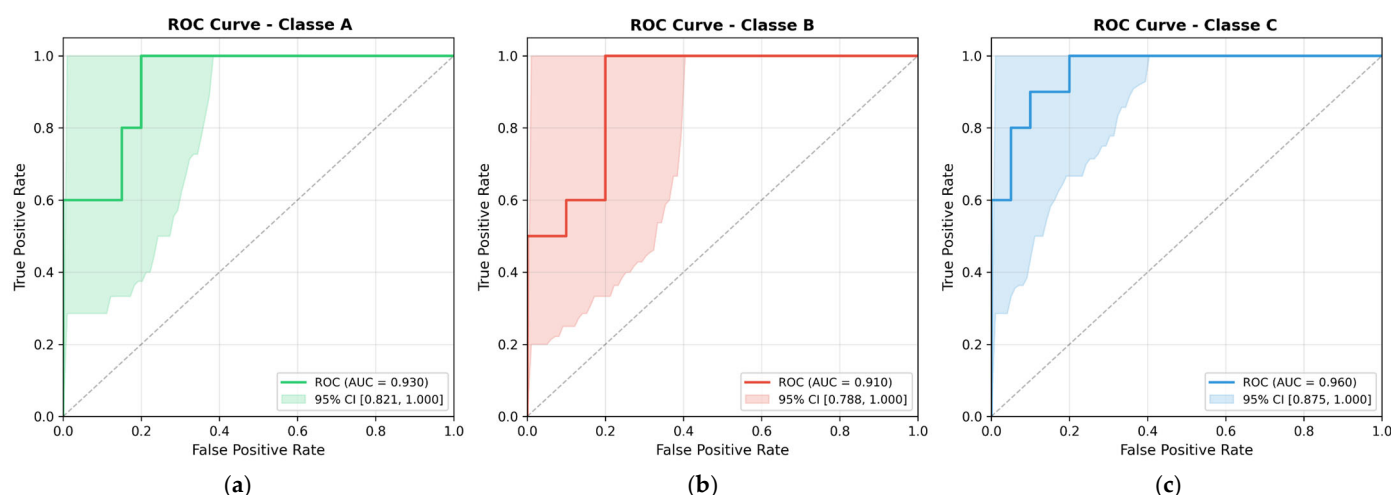


Figure 5. ROC curves and AUC for each class. The three graphs illustrate the discriminatory capacity of the model on the independent test set: (a) Class A (No Endoleak), (b) Class B (Type I Endoleak), and (c) Class C (T2EL). The colored area around each ROC curve represents the 95CI. The dotted diagonal line indicates the performance of a random classifier (AUC = 0.5). Abbreviations: 95CI, 95% Confidence Interval AUC, Area Under the Curve; ROC, Receiver Operating Characteristic; T2EL, Type II Endoleak.

Given the limited test set size ($n = 30$), a bootstrapping analysis (1000 iterations) was conducted to validate the robustness of the performance metrics. The bootstrapped mean accuracy was 76.8% (95% CI: 0.600–0.900), consistent with the primary evaluation. Detailed patient-level predictions, including probability distributions and confidence scores for each case, are provided in Table 3.

Table 3. Patient-level predictions for the test set ($n = 30$). The table shows the reference class, the predicted class, and the Softmax probability distribution generated by NornirNet for each patient.

Patient ID	Ground Truth	Predicted Class	Confidence (No T2EL)	Confidence (Benign)	Confidence (Malignant)	Outcome
Class A: No T2EL						
1	No T2EL	No T2EL	0.98	0.01	0.01	Correct
2	No T2EL	Benign	0.35	0.45	0.2	Incorrect
3	No T2EL	No T2EL	0.95	0.03	0.02	Correct
4	No T2EL	No T2EL	0.88	0.1	0.02	Correct
5	No T2EL	No T2EL	0.92	0.05	0.03	Correct
6	No T2EL	No T2EL	0.90	0.08	0.02	Correct
7	No T2EL	No T2EL	0.96	0.02	0.02	Correct
8	No T2EL	No T2EL	0.85	0.12	0.03	Correct
9	No T2EL	No T2EL	0.93	0.05	0.02	Correct
10	No T2EL	No T2EL	0.89	0.09	0.02	Correct
Class B: Benign						
11	Benign	Benign	0.10	0.85	0.05	Correct
12	Benign	Benign	0.05	0.9	0.05	Correct
13	Benign	Benign	0.15	0.75	0.1	Correct
14	Benign	Benign	0.02	0.96	0.02	Correct
15	Benign	No T2EL	0.45	0.4	0.15	Incorrect
16	Benign	No T2EL	0.48	0.42	0.1	Incorrect
17	Benign	Benign	0.1	0.8	0.1	Correct
18	Benign	Benign	0.05	0.88	0.07	Correct
19	Benign	Benign	0.12	0.82	0.06	Correct

20	Benign	Malignant	0.1	0.35	0.55	Incorrect
Class C: Malignant						
21	Malignant	Malignant	0.01	0.02	0.97	Correct
22	Malignant	Malignant	0.05	0.1	0.85	Correct
23	Malignant	Malignant	0.02	0.03	0.95	Correct
24	Malignant	Malignant	0.01	0.05	0.94	Correct
25	Malignant	Malignant	0.1	0.15	0.75	Correct
26	Malignant	Malignant	0.02	0.02	0.96	Correct
27	Malignant	Malignant	0.03	0.07	0.9	Correct
28	Malignant	No T2EL	0.45	0.2	0.35	Incorrect
29	Malignant	No T2EL	0.48	0.12	0.4	Incorrect
30	Malignant	Benign	0.15	0.5	0.35	Incorrect

Abbreviations: T2EL, Type II Endoleak.

Formal statistical comparisons revealed no significant differences in discriminative power between the classes: No T2EL vs. Benign ($p = 0.824$), No T2EL vs. Malignant ($p = 0.820$), and Benign vs. Malignant ($p = 0.792$). These results confirm that the model's predictive performance is statistically balanced across all clinical outcomes, showing equal efficacy in identifying benign and malignant endoleaks. To address the class imbalance, Precision-Recall (PR) curves were analyzed (Figure 6), showing high Average Precision (AP) for the Malignant class (AP = 0.97), confirming the model's ability to minimize false positives.

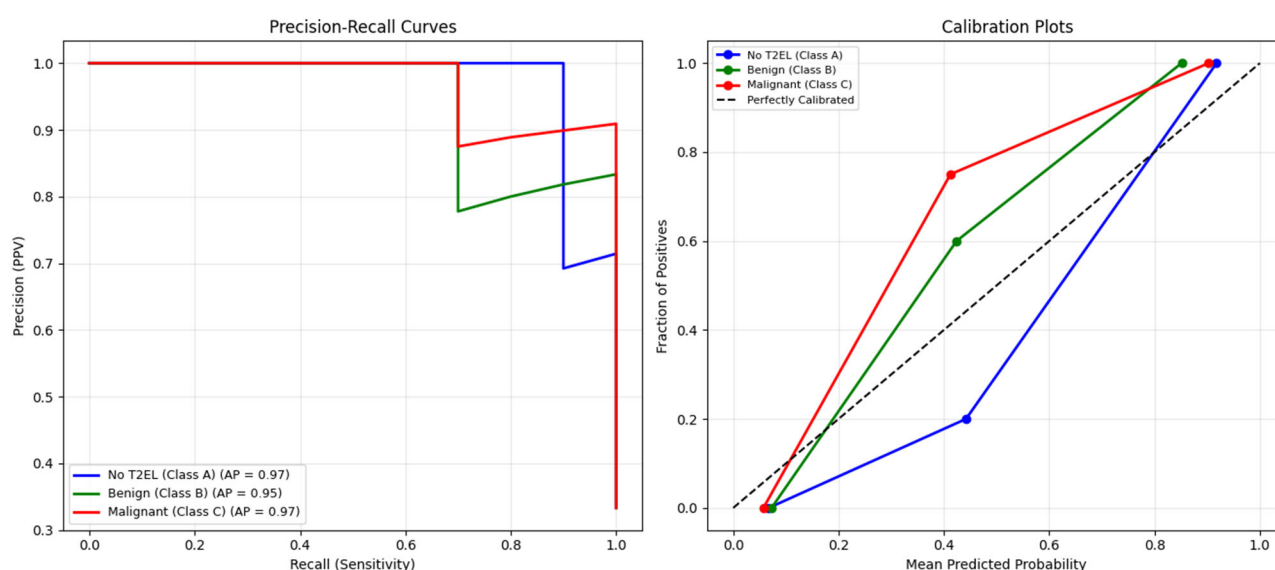


Figure 6. Performance analysis, precision-recall, and calibration. Precision-recall (PR) curves: model performance across all categories, achieving an average precision (AP) of 0.97 for the malignant class (class C) and 0.95 for the benign class (class B). Calibration graphs (reliability diagram): The reliability diagram evaluates the quality of the model's confidence estimates. The curves for all three classes generally follow the ideal diagonal (dashed line), showing the predicted probabilities. Abbreviations: AP, Average precision; T2EL, Type II Endoleak.

3.4. Confusion Matrix Analysis.

The Confusion Matrix analysis (Table 4) provides a detailed view of the model's misclassification patterns.

Out of the 10 patients belonging to Class A (No T2EL):

- 9 were correctly identified (TP), 90%.
- 1 was misclassified as Class B, 10%.

Of the 10 patients in Class B (Benign T2EL):

- 7 were correctly identified (TP), 70%.
- 2 were misclassified as Class A, 20%.
- 1 was misclassified as Class C, 10%.

Finally, for the 10 patients in Class C (Malignant T2EL):

- 7 were correctly identified (TP), 70%.
- 2 were classified as Class A, 20%.
- 1 was classified as Class B, 10%.

Table 4. Confusion matrix shows correct predictions (green boxes) and misclassifications (brown boxes) for each class and overall. Green and red characters indicate correct and wrong prediction percentages, respectively.

Confusion Matrix				
Target Output	No T2EL	Benign T2EL	Malignant T2EL	Sum
No T2EL	9 30.0%	1 3.3%	0 0.0%	10 90.0%/10.0%
Benign T2EL	2 6.7%	7 23.3%	1 3.3%	10 70.0%/30.0%
Malignant T2EL	2 6.7%	1 3.3%	7 23.3%	10 70.0%/30.0%
Sum	13 69.2%/30.8%	9 77.8%/22.2%	8 87.5%/12.5%	30 76.7%/23.3%

Abbreviations: T2EL, Type II Endoleak.

3.5. Model Interpretability and Grad-CAM Analysis

To help interpret the results of the 3D CNN, Grad-CAM was applied to representative cases from the test set (Figure 7). The analysis revealed that the model's attention is mainly localized on the aortic wall in preoperative scans. We observed a distinct activation gradient related to the predicted results:

- Class A (No T2EL): showed low-intensity activation patterns localized on the aortic wall.
- Class B (benign T2EL): showed medium-intensity activation.
- Class C (malignant T2EL): showed high-intensity activation (hotspots) on the aortic wall.

These results suggest that NornirNet identifies specific morphological or structural features on the preoperative aortic wall that are correlated with the risk of future sac growth caused by endoleak.

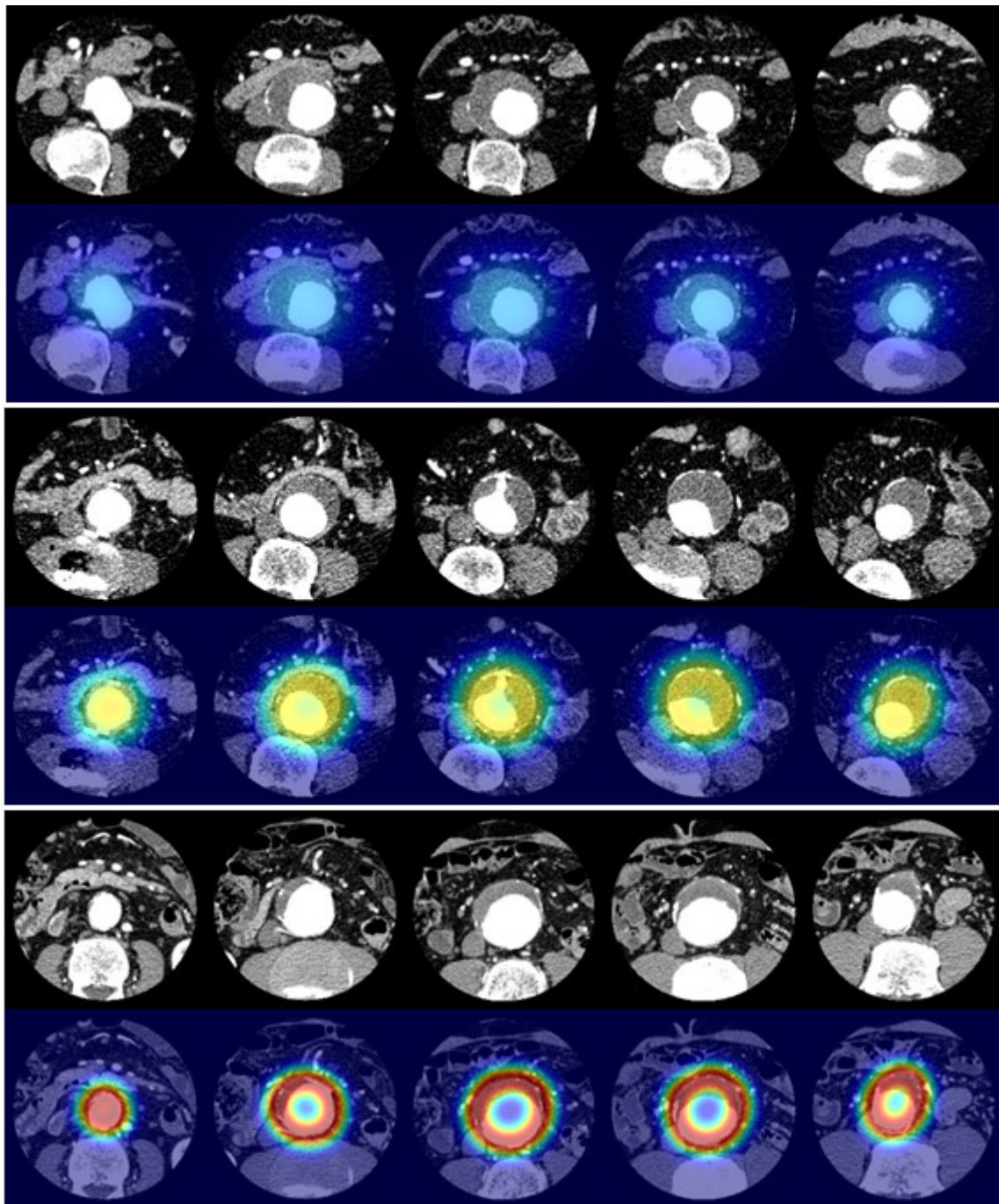


Figure 7. Interpretability analysis using Grad-CAM 3D. Representative axial sections from the test set for Class A (No T2EL), Class B (Benign), and Class C (Malignant). Rows 1, 3, and 5: original preoperative CT sections showing the morphology of the aneurysmal sac. Rows 2, 4, and 6: corresponding Grad-CAM overlays where the heat map indicates the importance of specific voxel regions for the model prediction. The heat map follows a chromatic scale where red indicated maximum activation and blue indicated minimum activation. The intensity of activation on the aortic wall increases from Class A (low activation) to Class C (high activation), highlighting the latent structural features identified by NornirNet.

3.6. Benchmarking Against State-of-the-Art 3D Frameworks

To validate the architectural efficiency of NornirNet, a comparative analysis was conducted against a state-of-the-art transfer learning framework: MedicalNet (Table 5). We used a 3D ResNet-50 backbone pre-trained on the Medical Segmentation Decathlon. To mitigate the risk of overfitting given the limited size of the dataset, we employed a linear

polling strategy, freezing the weights of the pre-trained backbone (approximately 46 million parameters) and training only the fully connected final layer on our dataset.

Table 5. Comprehensive Performance Benchmark: NornirNet vs. State-of-the-Art Baseline.

Metric	NornirNet	MedicalNet
GLOBAL METRICS		
Overall Accuracy	76.70%	46.70%
Macro-AUC	0.93	0.508
CLASS A: No T2EL		
AUC	0.93	0.595
Precision (PPV)	0.692	0.421
Recall (Sensitivity)	0.9	0.8
F1-Score	0.783	0.552
CLASS B: Benign T2EL		
AUC	0.91	0.52
Precision (PPV)	0.778	0.75
Recall (Sensitivity)	0.7	0.3
F1-Score	0.737	0.429
CLASS C: Malignant T2EL		
AUC	0.96	0.41
Precision (PPV)	0.875	0.429
Recall (Sensitivity)	0.7	0.3
F1-Score	0.778	0.353

Abbreviations: AUC, area under the curve; PPV, predictive positive value; T2EL, Type II endoleak.

3.7. Sensitivity Analysis on Follow-Up Duration

To address the potential variability introduced by non-uniform follow-up periods, a stratified sensitivity analysis was performed on the test set (Table 6). Patients were classified (see Appendix B for more details) according to follow-up duration as short (<24 months, $n = 6$), medium (24–48 months, $n = 11$), and long (>48 months, $n = 13$). The model in the long follow-up cohort achieved 100% accuracy (13/13). In the cohort with short follow-up (<24 months), the accuracy was 66.7%. Specifically, one classification error involved a patient (ID 20) labeled as benign T2EL (Class B) with 13.6 months of follow-up, whom the model classified as malignant (Class C). This discrepancy may reflect the model's ability to detect early indicators of sac expansion before they become clinically apparent, suggesting that short-term clinical labels may underestimate the actual risk. All patients in the malignant T2EL category (Class C) had a follow-up duration of >24 months.

Table 6. Sensitivity analysis: model accuracy by follow-up duration.

Follow-Up Group	Duration (Months)	Patients (n)	Correct Predictions	Accuracy (%)
Short-term	<24	6	4	66.70%
Medium-term	24–48	11	6	54.50%
Long-term	>48	13	13	100%
Total Test Set	/	30	23	76.70%

3.8. Runtime Performance and Computational Cost

The final model training on CT scans required approximately 95 h, with a throughput of 1.2–2.3 iterations/s using a batch size of 3. During training, the GPU utilization was ~80%, with VRAM usage between 95 and 100%.

The model comprises ~7.42 million parameters. The hyperparameter optimization phase was computationally intensive, totaling 1425.13 h (≈ 2 months). At inference time,

the model is efficient, requiring only a few seconds per volume; over 30 cases, the average inference time was 16.8 s per full CT volume.

4. Discussion

The present study presents a novel deep learning framework for the preoperative prediction of “malignant” T2EL after standard endovascular aneurysm repair (EVAR), based solely on volumetric CT angiography data.

The management of T2EL remains a controversial question since most of them resolve spontaneously [15], but a subset is associated with sac enlargement and related complications such as aneurysm rupture or proximal and distal seal loss [16,20]. Different meta-analyses on the treatment success of T2ELs report that, despite high technical success rates, these interventions often fail to prevent subsequent aneurysm growth [7,21,22] and robust evidence supporting the clinical benefit of T2EL treatment is lacking [23]. Several authors have proposed risk-factor models to stratify patients at higher risk of T2ELs [24–26]; however, these methods are complex, time-consuming, non-standardized, not automated, and they carry a considerable degree of observer dependence, limiting their practicality in routine clinical practice. Furthermore, the current guidelines suggest considering the pre-emptive embolization of the lumbar arteries or the inferior mesenteric artery to prevent T2EL only in very selected cases, failing to identify a validated way to identify them [1]. In this scenario a simple, fast, and reproducible way to identify which patients could really benefit from a pre-emptive embolization would be extremely useful in clinical practice.

Several studies [27–29] confirmed that CNN models and deep learning architectures allow not only the detection but also the automatic segmentation of aneurysms and endoleaks in post-EVAR images, with performance comparable or superior to expert radiologists in multi-site tests, even on non-contrast scans. This suggests that volumetric quantification and expansion risk are becoming less dependent on manual segmentation and increasingly automatable, as demonstrated by Hahn et al. [28] and Talebi et al. [29] with models achieving AUCs of up to 0.99.

To the best of authors’ knowledge, this is the first end-to-end 3D convolutional neural network (CNN) designed to identify patients with malignant T2EL directly from raw pre-operative CT volumes without the need for handcrafted features or manual segmentation.

The proposed architecture demonstrated encouraging performance, with an overall accuracy of 76.7% and a ROC AUC of 0.933 on an independent test set. The model achieved particularly high discriminative capability in identifying patients without endoleak (AUC 0.930) and those developing malignant forms (AUC 0.960), confirming its potential utility in targeting the patient who could benefit from a pre-emptive treatment to prevent the T2EL after EVAR. The relatively balanced performance across the three classes supports the model’s ability to capture complex spatial and textural information related to collateral perfusion patterns and aortic wall morphology, which are difficult to quantify using conventional radiologic descriptors.

The localized activation on the aortic wall observed in our Grad-CAM analysis confirms that NornirNet exploits latent structural information predictive of endoleak-induced sac reperfusion.

However, we must acknowledge the inherent limitations of saliency-based interpretability. As noted by Adebayo et al. [30], these maps can sometimes function as simple edge detectors rather than representing the true decision logic of the model. Furthermore, Ghassemi et al. [31] cautioned that explainability tools in medical imaging may offer a “false sense of security” and should not replace clinical judgment. In our study, the “hot spots” on the aortic wall serve as an anatomical proxy for risk, but the model’s high

discriminatory power (overall AUC 0.93) remains its most reliable feature for identifying patients who may benefit from preventive treatment.

Previous studies applied radiomics and classical machine learning to predict T2EL, obtaining better results compared to our cohort in terms of sensitivity and specificity. However, Charalambous et al. [12] proposed a radiomics-based approach to identify persistent T2EL dealing with postoperative CTA where the complication is already present. Different papers seem to suggest that the benefit of interventions carried out on correcting already present T2EL is limited [21,23], reducing the clinical benefit of the suggested approach. Additionally, no distinction was made between endoleaks with and without sac enlargement, which remains the only parameter currently correlated with clinical outcomes and the need for reintervention. Furthermore, radiomic methods, although powerful, require a manual pipeline for feature extraction, which is subject to inter-observer variability [32]. End-to-end deep learning-based models, on the other hand, learn feature hierarchies directly from voxel-based data, drastically reducing preprocessing and enhancing reproducibility [11]. Podgorsak et al. [33] found that automated CNNs can achieve effective segmentations, surpassing traditional radiomic pipelines in efficiency and scalability. Jinquing Mo et al. [27] employed an ML-based predictive model integrating preoperative clinical data and radiomic features to enable the stratification of patients' risk to develop T2EL after EVAR, but they do not differentiate between T2EL causing aneurysm growth or not ("malignant" vs. "benign"), limiting the guidance for perioperative preventive measures, since the treatment of all high-risk patients for persistent T2EL may lead to overtreatment [23].

In addition, previous studies on the association between radiomics, classical machine learning, and T2EL with AAA growth have relied on manually segmented regions of interest and predefined feature extraction pipelines [12,27]. Such approaches are inherently limited by observer-dependent variability and potential information loss. In contrast, our 3D CNN overcomes these limitations by performing hierarchical feature learning directly from voxel-level data.

Recent multitask and attention-based models [34] allow for better leak localization, with increasing explainability, showing the vascular and morphological areas of greatest relevance for automated risk prediction. Smorenburg et al. [35] proposes a real-time framework based on intraoperative digital subtraction angiography, which demonstrates F1 scores and accuracy comparable to the human procedural team in 220 patients, indicating increasing reliability of automated decision support. In our study, we considered as baseline the first postoperative CTA instead of the intraoperative final angiography. This approach was chosen because many small T2ELs undergo early spontaneous resolution, and because, in the immediate post-procedural phase, angiography may not reliably distinguish a true type II endoleak from graft porosity.

The model architecture was intentionally designed with progressive channel expansion and multi-layer dropout to balance computational efficiency and overfitting prevention. The adoption of GELU activation functions contributed to smoother gradient propagation and improved convergence stability compared with traditional ReLU-based configurations. The achieved performance metrics suggest that further gains may be obtained through ensemble learning or hybrid architectures integrating attention mechanisms to refine spatial localization of collateral vessels. Moreover, transfer learning strategies leveraging larger CT datasets could improve generalizability across scanners and acquisition protocols.

Despite these promising results, several limitations must be acknowledged. The matrix data show that most misclassifications occurred between adjacent classes or towards Class A, suggesting a tendency for the model to misclassify endoleaks as 'no T2EL' (4 false negatives in total, 2 from Class B and 2 from Class C) or, conversely, to classify 'no T2EL'

patients as ‘benign T2EL’ (1 false positive). Even if, from a clinical point of view, the only relevant mistake is the misclassification of malignant T2EL, these results suggest the need for further learning of the model.

First, the retrospective single-center design may limit external validity, particularly when different CT scanners were utilized. Moreover, the test cohort size was modest, with only 10 malignant T2EL patients in the validation set. Multi-institutional validation with heterogeneous imaging datasets is planned as the next step to assess model robustness and domain generalization. Second, although data augmentation and normalization were applied, the possible differences in acquisition parameters across the study period might have introduced image variability. Third, the model did not incorporate clinical variables or post-procedural data that could potentially improve predictive accuracy when combined with imaging-derived features in a multimodal architecture. During this initial phase, these variables were intentionally omitted to allow the model to focus exclusively on information extracted from preoperative CTA and to maintain maximal model simplicity. To improve the generalizability of NornirNet, future work will focus on multi-center external validation. A major challenge in such multi-institutional AI studies is the heterogeneity of imaging data—often referred to as domain shift—which arises from differences in CT scanner manufacturers, reconstruction kernels, and contrast protocols. Following this multicenter external validation, a subsequent study will be conducted to quantify the incremental value of incorporating additional clinical variables.

From a translational perspective, this study demonstrates that preoperative CT imaging contains sufficient latent information to predict continued aneurysm sac reperfusion due to T2EL after EVAR. Such models could support decision-making regarding preemptive embolization or intensified follow-up protocols, provided that prospective validation confirms their reliability and clinical benefit.

Future research should focus on model explainability, integration with radiomic biomarkers, and the development of interpretable attention maps highlighting the vascular territories contributing most to prediction.

The present work establishes the feasibility of using deep 3D convolutional architectures for automated risk prediction of T2EL after EVAR, outperforming traditional radiomic approaches and providing a foundation for AI-driven stratification tools in endovascular therapy.

5. Conclusions

This study introduces a fully automated deep learning framework capable of predicting the occurrence and clinical severity of T2EL after EVAR using only preoperative CTA data. The proposed 3D convolutional neural network achieved high accuracy and good class discrimination without requiring manual segmentation or handcrafted features. These results demonstrate that preoperative imaging encodes latent structural and hemodynamic information that can be effectively leveraged through volumetric deep learning to anticipate post-procedural outcomes. Future work will aim to externally validate the model across multicenter datasets and to develop an algorithm aimed to support precision planning and individualized surveillance strategies in EVAR.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ai7020057/s1>, Table S1: Distribution of baseline and anatomical characteristics among Training, Test and Validation and subgroup of patients.

Author Contributions: Conceptualization, G.P., F.A., F.M., and A.R.; methodology, F.A. and F.M.; software, F.M., A.L.; validation, G.P., A.R., M.A.R. and L.G.; formal analysis, F.A., E.W., J.G. and L.E.; investigation, G.P., F.A., F.M. and E.W. resources, M.A.R. and A.L.; data curation, F.A.; E.W.,

J.G. and L.E. writing—original draft preparation, F.A. and F.M.; writing review and editing, G.P., A.R., L.E., A.L., L.G. and M.A.R.; visualization, L.G. and J.G.; supervision, G.P. and A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Canton Ticino, Bellinzona, Switzerland (protocol nr. 2025-00343 CE 4781, approved on 31.03.2025). Helsinki Declaration and its later amendments were respected and the data underlying this article will be shared on reasonable request to the corresponding author. All data used in this work are anonymized and accessible only to the authors of this article.

Informed Consent Statement: All surviving patients treated before 2021 were sent a non-opposition consent form, since 2021, all participants have provided signed informed consent for clinical research and data analysis.

Data Availability Statement: The data presented in this study are available on reasonable request from the corresponding author.

Acknowledgments: The authors would like to thank Stefania Rizzo and Ermidio Rezzonico for technical assistance; Maria Carla Andreozzi and Elena Garbero for study coordination.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

95CI	95% Confidence Interval
AAA	Abdominal Aortic Aneurysm
AI	Artificial Intelligence
AUC	Area Under the Curve
CT	Computed Tomography
CTA	Computed Tomography Angiography
3D CNN	3D Convolutional Neural Network
ESVS	European Society for Vascular Surgery
EVAR	EndoVascular Aneurysm Repair
EOC	Ente Ospedaliero Cantonale
GELU	Gaussian Error Linear Unit
IIMSI	Imaging Institute of Southern Switzerland
IQR	Interquartile Range
PPV	Predictive positive value
ROC	Receiver Operating Characteristic Curve
SD	Standard Deviation
T2EL	Type II Endoleak

Appendix A. Source Code for the NornirNet Network Architecture (PyTorch Implementation)

```
import torch
import torch.nn as nn

class ResidualBlock(nn.Module):
    def __init__(self, in_channels, out_channels, stride, dropout_prob):
        super(ResidualBlock, self).__init__()

        self.conv1 = nn.Conv3d(in_channels, in_channels, kernel_size=3, padding=1, bias=False)
        self.bn1 = nn.BatchNorm3d(in_channels)
        self.gelu1 = nn.GELU()
```

```

self.conv2 = nn.Conv3d(in_channels, out_channels, kernel_size=3, stride=stride, padding=1, bias=False)
self.bn2 = nn.BatchNorm3d(out_channels)
self.dropout = nn.Dropout3d(p=dropout_prob)

self.shortcut = nn.Sequential()
if stride != (1, 1, 1) or in_channels != out_channels:
    self.shortcut = nn.Sequential(
        nn.Conv3d(in_channels, out_channels, kernel_size=1, stride=stride, bias=False),
        nn.BatchNorm3d(out_channels)
    )

self.gelu_final = nn.GELU()

def forward(self, x):
    residual = self.shortcut(x)

    out = self.conv1(x)
    out = self.bn1(out)
    out = self.gelu1(out)

    out = self.conv2(out)
    out = self.bn2(out)
    out = self.dropout(out)

    out += residual
    out = self.gelu_final(out)

    return out

class NornirNet(nn.Module):
    def __init__(self, num_classes=3):
        super(NornirNet, self).__init__()

        # %% INITIAL CONV
        self.initial_conv = nn.Sequential(
            nn.Conv3d(1, 32, kernel_size=3, padding=1, bias=False),
            nn.BatchNorm3d(32),
            nn.GELU()
        )

        # %% RES BLOCK 1: 32 -> 64, Stride 1,2,2
        self.res_block1 = ResidualBlock(
            in_channels=32,
            out_channels=64,
            stride=(1, 2, 2),
            dropout_prob=0.3
        )

        # %% RES BLOCK 2: 64 -> 128, Stride 2,2,2
        self.res_block2 = ResidualBlock(
            in_channels=64,
            out_channels=128,
            stride=(2, 2, 2),
            dropout_prob=0.3
        )

        # %% RES BLOCK 3: 128 -> 256, Stride 2,2,2
        self.res_block3 = ResidualBlock(

```

```

        in_channels=128,
        out_channels=256,
        stride=(2, 2, 2),
        dropout_prob=0.4
    )

    # %% RES BLOCK 4: 256 -> 512, Stride 2,2,2
    self.res_block4 = ResidualBlock(
        in_channels=256,
        out_channels=512,
        stride=(2, 2, 2),
        dropout_prob=0.4
    )

    # %% FINAL POOLING & FLATTEN
    self.adaptive_pool = nn.AdaptiveAvgPool3d((1, 1, 1))

    # %% CLASSIFIER MLP
    self.classifier = nn.Sequential(
        nn.Linear(512, 256),
        nn.GELU(),
        nn.Dropout(p=0.4),
        nn.Linear(256, 128),
        nn.GELU(),
        nn.Dropout(p=0.3),
        nn.Linear(128, num_classes)
    )

def forward(self, x):
    x = self.initial_conv(x)

    x = self.res_block1(x)
    x = self.res_block2(x)
    x = self.res_block3(x)
    x = self.res_block4(x)

    x = self.adaptive_pool(x)
    x = torch.flatten(x, 1)

    logits = self.classifier(x)

    return logits

device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
model = NornirNet(num_classes=3).to(device)

```

Appendix B. Raw Patient-Level Data and Follow-Up Sensitivity

This appendix provides the complete set of raw data for the test set to ensure maximum transparency regarding the duration of clinical follow-up and its impact on model performance.

Table A1. Detailed follow-up and prediction metrics for the test set.

id	Case	Follow-Up Months	Follow-Up Cat (0 < 48 m, 1 > 48 m)	Follow-Up Cat (0 < 24 m, 1 > 24 m)	Follow-Up Cat (0 < 12 m, 1 > 12 m)	Confidence (No T2EL)	Confidence (Benign)	Confidence (Malignant)
1	A	74	1	1	1	0.98	0.01	0.01
2	A	20	0	0	1	0.35	0.45	0.2
3	A	79	1	1	1	0.95	0.03	0.02

4	A	79	1	1	1	0.88	0.1	0.02
5	A	50	1	1	1	0.92	0.05	0.03
6	A	49	1	1	1	0.9	0.08	0.02
7	A	43	0	1	1	0.96	0.02	0.02
8	A	33	0	1	1	0.85	0.12	0.03
9	A	25	0	1	1	0.93	0.05	0.02
10	A	18	0	0	1	0.89	0.09	0.02
11	B	85	1	1	1	0.1	0.85	0.05
12	B	59	1	1	1	0.05	0.9	0.05
13	B	43	0	1	1	0.15	0.75	0.1
14	B	58	1	1	1	0.02	0.96	0.02
15	B	32	0	1	1	0.45	0.4	0.15
16	B	36	0	1	1	0.48	0.42	0.1
17	B	21	0	0	1	0.1	0.8	0.1
18	B	18	0	0	1	0.05	0.88	0.07
19	B	16	0	0	1	0.12	0.82	0.06
20	B	14	0	0	1	0.1	0.35	0.55
21	C	146	1	1	1	0.01	0.02	0.97
22	C	117	1	1	1	0.05	0.1	0.85
23	C	57	1	1	1	0.02	0.03	0.95
24	C	83	1	1	1	0.01	0.05	0.94
25	C	57	1	1	1	0.1	0.15	0.75
26	C	39	0	1	1	0.02	0.02	0.96
27	C	34	0	1	1	0.03	0.07	0.9
28	C	37	0	1	1	0.45	0.2	0.35
29	C	25	0	1	1	0.48	0.12	0.4
30	C	24	0	1	1	0.15	0.5	0.35

The Table A1 shows the ground truth classification along with the exact follow-up duration (in months) for each case. We classified the observation periods into three thresholds: <12 months, <24 months, and <48 months.

As observed in the sensitivity analysis, all patients identified as high risk by Nornir-Net (Class C) had an observation period longer than the median time typically required for reoperation, thus ensuring the clinical validity of the “malignant” designation.

For each patient, the table provides the Softmax probability distribution (confidence scores) for all three outcomes. This level of detail confirms that incorrect model predictions often occur with lower confidence levels, particularly in cases with shorter follow-ups where the clinical outcome may still be evolving.

References

1. Wanhainen, A.; Van Herzele, I.; Bastos Goncalves, F.; Bellmunt Montoya, S.; Berard, X.; Boyle, J.R.; D’Oria, M.; Prendes, C.F.; Karkos, C.D.; Kazimierczak, A.; et al. Editor’s Choice—European Society for Vascular Surgery (ESVS) 2024 Clinical Practice Guidelines on the Management of Abdominal Aorto-Iliac Artery Aneurysms. *Eur. J. Vasc. Endovasc. Surg. Off. J. Eur. Soc. Vasc. Surg.* **2024**, *67*, 192–331. <https://doi.org/10.1016/j.ejvs.2023.11.002>.
2. Powell, J.T.; Sweeting, M.J.; Ulug, P.; Blankensteijn, J.D.; Lederle, F.A.; Becquemin, J.-P.; Greenhalgh, R.M.; EVAR-1, DREAM, OVER and ACE Trialists Meta-Analysis of Individual-Patient Data from EVAR-1, DREAM, OVER and ACE Trials Comparing Outcomes of Endovascular or Open Repair for Abdominal Aortic Aneurysm over 5 Years. *Br. J. Surg.* **2017**, *104*, 166–178. <https://doi.org/10.1002/bjs.10430>.
3. White, G.H.; Yu, W.; May, J.; Chaufour, X.; Stephen, M.S. Endoleak as a Complication of Endoluminal Grafting of Abdominal Aortic Aneurysms: Classification, Incidence, Diagnosis, and Management. *J. Endovasc. Surg. Off. J. Int. Soc. Endovasc. Surg.* **1997**, *4*, 152–168. [https://doi.org/10.1583/1074-6218\(1997\)004%3C0152:EAACOE%3E2.0.CO;2](https://doi.org/10.1583/1074-6218(1997)004%3C0152:EAACOE%3E2.0.CO;2).
4. Gennai, S.; Andreoli, F.; Leone, N.; Bartolotti, L.A.M.; Maleti, G.; Silingardi, R. Incidence, Long Term Clinical Outcomes, and Risk Factor Analysis of Type III Endoleaks Following Endovascular Repair of Abdominal Aortic Aneurysm. *Eur. J. Vasc. Endovasc. Surg. Off. J. Eur. Soc. Vasc. Surg.* **2023**, *66*, 38–48. <https://doi.org/10.1016/j.ejvs.2023.03.018>.

5. Chaikof, E.L.; Dalman, R.L.; Eskandari, M.K.; Jackson, B.M.; Lee, W.A.; Mansour, M.A.; Mastracci, T.M.; Mell, M.; Murad, M.H.; Nguyen, L.L.; et al. The Society for Vascular Surgery Practice Guidelines on the Care of Patients with an Abdominal Aortic Aneurysm. *J. Vasc. Surg.* **2018**, *67*, 2–77.e2. <https://doi.org/10.1016/j.jvs.2017.10.044>.
6. Yu, H.Y.H.; Lindström, D.; Wanhainen, A.; Tegler, G.; Hassan, B.; Mani, K. Systematic Review and Meta-Analysis of Prophylactic Aortic Side Branch Embolization to Prevent Type II Endoleaks. *J. Vasc. Surg.* **2020**, *72*, 1783–1792.e1. <https://doi.org/10.1016/j.jvs.2020.05.020>.
7. Yu, H.Y.H.; Lindström, D.; Wanhainen, A.; Tegler, G.; Ascitto, G.; Mani, K. An Updated Systematic Review and Meta-Analysis of Pre-Emptive Aortic Side Branch Embolization to Prevent Type II Endoleaks after Endovascular Aneurysm Repair. *J. Vasc. Surg.* **2023**, *77*, 1815–1821. <https://doi.org/10.1016/j.jvs.2022.11.042>.
8. Ichihashi, S.; Takahara, M.; Fujimura, N.; Banno, H.; Onitsuka, S.; Shingaki, M.; Yamaoka, T.; Sumi, M.; Iida, O.; Iwakoshi, S.; et al. Editor's Choice—Multicentre Randomised Controlled Trial to Evaluate the Efficacy of Pre-Emptive Inferior Mesenteric Artery Embolisation during Endovascular Aortic Aneurysm Repair on Aneurysm Sac Change. *Eur. J. Vasc. Endovasc. Surg. Off. J. Eur. Soc. Vasc. Surg.* **2025**, *70*, 219–226. <https://doi.org/10.1016/j.ejvs.2025.04.028>.
9. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107. <https://doi.org/10.1158/0008-5472.CAN-17-0339>.
10. van Timmeren, J.E.; Cester, D.; Tanadini-Lang, S.; Alkadhi, H.; Baessler, B. Radiomics in Medical Imaging—"how-to" Guide and Critical Reflection. *Insights Imaging* **2020**, *11*, 91. <https://doi.org/10.1186/s13244-020-00887-2>.
11. Fedorov, A.; Beichel, R.; Kalpathy-Cramer, J.; Finet, J.; Fillion-Robin, J.-C.; Pujol, S.; Bauer, C.; Jennings, D.; Fennessy, F.; Sonka, M.; et al. 3D Slicer as an Image Computing Platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **2012**, *30*, 1323–1341. <https://doi.org/10.1016/j.mri.2012.05.001>.
12. Charalambous, S.; Klontzas, M.E.; Kontopodis, N.; Ioannou, C.V.; Perisinakis, K.; Maris, T.G.; Damilakis, J.; Karantanis, A.; Tsetis, D. Radiomics and Machine Learning to Predict Aggressive Type 2 Endoleaks after Endovascular Aneurysm Repair: A Proof of Concept. *Acta Radiol. Stockh. Swed. 1987* **2022**, *63*, 1293–1299. <https://doi.org/10.1177/02841851211032443>.
13. Javor, D.; Wressnegger, A.; Unterhumer, S.; Kollndorfer, K.; Nolz, R.; Beitzke, D.; Loewe, C. Endoleak Detection Using Single-Acquisition Split-Bolus Dual-Energy Computer Tomography (DECT). *Eur. Radiol.* **2017**, *27*, 1622–1630. <https://doi.org/10.1007/s00330-016-4480-6>.
14. Prouse, G.; Robaldo, A.; van den Berg, J.C.; Ettorre, L.; Mongelli, F.; Giovannacci, L. Impact of Multidisciplinary Team Meetings on Decision Making in Vascular Surgery: A Prospective Observational Study. *Eur. J. Vasc. Endovasc. Surg. Off. J. Eur. Soc. Vasc. Surg.* **2023**, *66*, 130–135. <https://doi.org/10.1016/j.ejvs.2023.03.037>.
15. Kumar, L.; Cowled, P.; Boulton, M.; Howell, S.; Fitridge, R. Type II Endoleak after Endovascular Aneurysm Repair: Natural History and Treatment Outcomes. *Ann. Vasc. Surg.* **2017**, *44*, 94–102. <https://doi.org/10.1016/j.avsg.2017.04.029>.
16. El Batti, S.; Cochenne, F.; Roudot-Thoraval, F.; Becquemin, J.-P. Type II Endoleaks after Endovascular Repair of Abdominal Aortic Aneurysm Are Not Always a Benign Condition. *J. Vasc. Surg.* **2013**, *57*, 1291–1297. <https://doi.org/10.1016/j.jvs.2012.10.118>.
17. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2023**, arXiv:1606.08415.
18. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
19. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Int. Conf. Mach. Learn.* **2015**, *37*, 448–456.
20. Lo, R.C.; Buck, D.B.; Herrmann, J.; Hamdan, A.D.; Wyers, M.; Patel, V.I.; Fillingner, M.; Schermerhorn, M.L. Risk Factors and Consequences of Persistent Type II Endoleaks. *J. Vasc. Surg.* **2016**, *63*, 895–901. <https://doi.org/10.1016/j.jvs.2015.10.088>.
21. Akmal, M.M.; Pabittei, D.R.; Prapassaro, T.; Suhartono, R.; Moll, F.L.; Van Herwaarden, J.A. A Systematic Review of the Current Status of Interventions for Type II Endoleak after EVAR for Abdominal Aortic Aneurysms. *Int. J. Surg.* **2021**, *95*, 106138. <https://doi.org/10.1016/j.ijsu.2021.106138>.
22. Ultee, K.H.J.; Büttner, S.; Huurman, R.; Bastos Gonçalves, F.; Hoeks, S.E.; Bramer, W.M.; Schermerhorn, M.L.; Verhagen, H.J.M. Editor's Choice—Systematic Review and Meta-Analysis of the Outcome of Treatment for Type II Endoleak Following Endovascular Aneurysm Repair. *Eur. J. Vasc. Endovasc. Surg.* **2018**, *56*, 794–807. <https://doi.org/10.1016/j.ejvs.2018.06.009>.
23. Mulay, S.; Geraedts, A.C.M.; Koelemay, M.J.W.; Balm, R.; Mulay, S.; Balm, R.; Elshof, J.W.; Elsmann, B.H.P.; Hamming, J.F.; Koelemay, M.J.W.; et al. Type 2 Endoleak with or without Intervention and Survival After Endovascular Aneurysm Repair. *Eur. J. Vasc. Endovasc. Surg.* **2021**, *61*, 779–786. <https://doi.org/10.1016/j.ejvs.2021.01.017>.

24. Mascoli, C.; Faggioli, G.; Gallitto, E.; Pini, R.; Fenelli, C.; Cercenelli, L.; Marcelli, E.; Gargiulo, M. Tailored Sac Embolization During EVAR for Preventing Persistent Type II Endoleak. *Ann. Vasc. Surg.* **2021**, *76*, 293–301. <https://doi.org/10.1016/j.avsg.2021.01.118>.
25. Marchiori, A.; von Ristow, A.; Guimaraes, M.; Schönholz, C.; Uflacker, R. Predictive Factors for the Development of Type II Endoleaks. *J. Endovasc. Ther. Off. J. Int. Soc. Endovasc. Spec.* **2011**, *18*, 299–305. <https://doi.org/10.1583/10-3116.1>.
26. Otsu, M.; Ishizaka, T.; Watanabe, M.; Hori, T.; Kohno, H.; Ishida, K.; Nakaya, M.; Matsumiya, G. Analysis of Anatomical Risk Factors for Persistent Type II Endoleaks Following Endovascular Abdominal Aortic Aneurysm Repair Using CT Angiography. *Surg. Today* **2016**, *46*, 48–55. <https://doi.org/10.1007/s00595-015-1115-5>.
27. Mo, J.; Liu, Q.; Wang, K.; Huang, L.; Yao, C. Prediction of Persistent Type II Endoleak after Endovascular Aortic Repair Using Machine Learning Based on Preoperative Clinical Data and Radiomic. *Vasc. Investig. Ther.* **2025**, *8*, 31. https://doi.org/10.4103/vit.vit_3_25.
28. Hahn, S.; Perry, M.; Morris, C.S.; Wshah, S.; Bertges, D.J. Machine Deep Learning Accurately Detects Endoleak after Endovascular Abdominal Aortic Aneurysm Repair. *JVS-Vasc. Sci.* **2020**, *1*, 5–12. <https://doi.org/10.1016/j.jvssci.2019.12.003>.
29. Talebi, S.; Madani, M.H.; Madani, A.; Chien, A.; Shen, J.; Mastrodicasa, D.; Fleischmann, D.; Chan, F.P.; Mofrad, M.R.K. Machine Learning for Endoleak Detection after Endovascular Aortic Repair. *Sci. Rep.* **2020**, *10*, 18343. <https://doi.org/10.1038/s41598-020-74936-7>.
30. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 9525–9536.
31. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9).
32. Rinaldi, L.; De Angelis, S.P.; Raimondi, S.; Rizzo, S.; Fanciullo, C.; Rampinelli, C.; Mariani, M.; Lascialfari, A.; Cremonesi, M.; Orecchia, R.; et al. Reproducibility of Radiomic Features in CT Images of NSCLC Patients: An Integrative Analysis on the Impact of Acquisition and Reconstruction Parameters. *Eur. Radiol. Exp.* **2022**, *6*, 2. <https://doi.org/10.1186/s41747-021-00258-6>.
33. Podgorsak, A.R.; Rava, R.A.; Shiraz Bhurwani, M.M.; Chandra, A.R.; Davies, J.M.; Siddiqui, A.H.; Ionita, C.N. Automatic Radiomic Feature Extraction Using Deep Learning for Angiographic Parametric Imaging of Intracranial Aneurysms. *J. Neurointerventional Surg.* **2020**, *12*, 417–421. <https://doi.org/10.1136/neurintsurg-2019-015214>.
34. Garrido-Oliver, J.; Tenezaca-Sari, X.; Faner Capo, X.; Ribeiro, T.; Dux-Santoy, L.; Ferrer-Cornet, M.; Bragulat-Arevalo, M.; Catala-Santarrufina, A.; Morales-Galan, A.; Lopez-Gutierrez, P.; et al. Deep Learning for Segmentation and Endoleak Detection in Contrast-Enhanced Computed Tomography in Endovascular Aortic Repair Patients. *Eur. Heart J.* **2025**, *46*, ehaf784.3004. <https://doi.org/10.1093/eurheartj/ehaf784.3004>.
35. Smorenburg, S.P.M.; Hoksbergen, A.W.J.; Yeung, K.K.; Wolterink, J.M. Multitask Deep Learning for Automated Detection of Endoleak at Digital Subtraction Angiography during Endovascular Aneurysm Repair. *Radiol. Artif. Intell.* **2025**, *7*, e240392. <https://doi.org/10.1148/ryai.240392>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.