ENSTA IP PARIS



Conception d'une architecture neuronale hybride à fusion multimodale pour la classification automatique d'événements sismiques volcaniques

Application au volcan Llaima (Chili)

KU KASETSART UNIVERSITY

FISE 2027

Robotique Autonome – ISE

2026

Juliette LOURDAULT
juliette.lourdault@ensta.fr

Remerciements

Je tiens à remercier Piya Limcharoen, mon tuteur de stage, pour son encadrement, la qualité de ses conseils architecturaux tout au long du projet, et la confiance qu'il m'a accordée dans la conduite de ce travail jusqu'à sa valorisation sous la forme d'un article scientifique co-signé.

Je remercie également l'ensemble de l'équipe du Department of Computer Engineering de la Faculty of Engineering de Kasetsart University pour leur accueil et leur disponibilité durant ces quatre mois de stage.

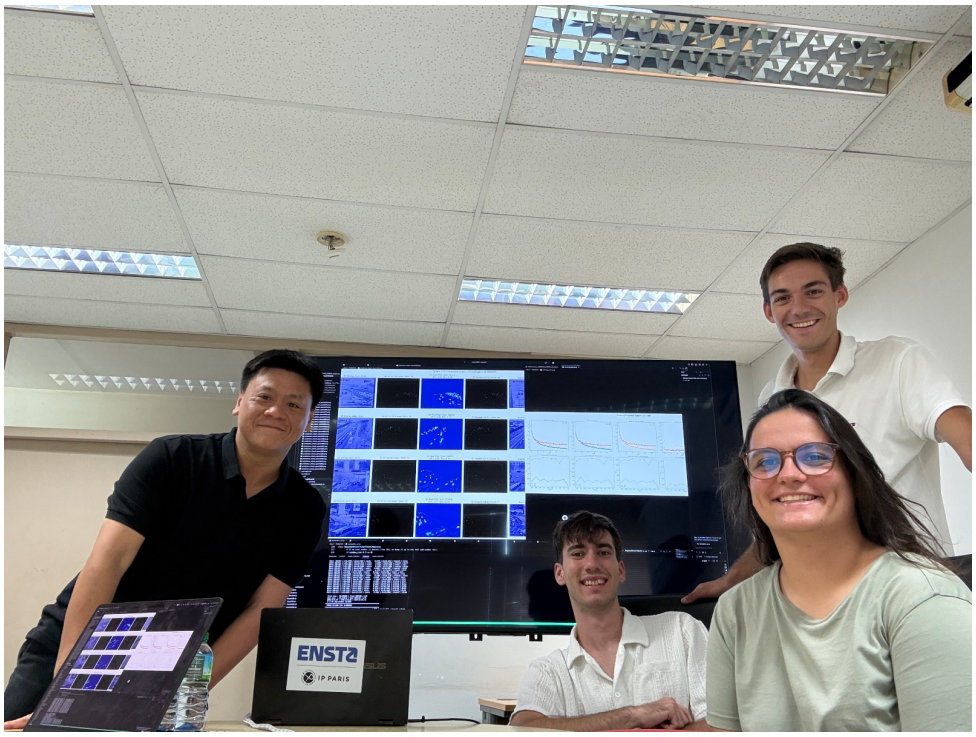


Figure 1: Photo de Clément, Florian et moi avec notre maître de stage Piya Limcharoen

Résumé

La surveillance sismique automatisée des volcans actifs repose de plus en plus sur des méthodes d'apprentissage profond capables de classer, en continu et sans intervention humaine, les différents types d'événements enregistrés par les réseaux de stations. Ce stage, effectué au sein du Department of Computer Engineering de la Faculty of Engineering de Kasetsart University, portait sur la classification automatique de quatre types d'événements sismiques du volcan Llaima (Chili) : longue période (LP), tremor (TR), volcano-tectonique (VT) et tectonique (TC), à partir de signaux bruts échantillonnés à 100 Hz sur des fenêtres de 60 secondes.

Le travail a consisté, dans un premier temps, à ré-implémenter trois méthodes de référence de la littérature : un classifieur convolutif appliqué à des spectrogrammes (Curilem et al., 2018), une comparaison d'architectures profondes convolutives et récurrentes (Canário et al., 2020), et un classifieur de type perceptron multicouche appliqué à des descripteurs d'enveloppe temporels et spectraux, optimisé par algorithme génétique (Ait Laasri et al., 2025). Ces travaux ont servi de socle à une démarche de conception itérative, documentée dans ce rapport en cinq étapes : d'un ensemble CNN + BiLSTM à vote moyen, vers une pondération optimisée par recherche sur grille, puis vers des architectures hybrides bout-en-bout (CNN→BiLSTM et BiLSTM→CNN), avant l'intégration d'une seconde entrée fondée sur le signal brut, et enfin l'ajout d'un bloc de fusion non linéaire de type perceptron multicouche entre les deux branches. Cette dernière architecture bimodale, entraînée et évaluée par validation croisée stratifiée à 5 plis sur les 3 592 signaux du jeu de données, constitue la contribution originale du stage et a fait l'objet d'un article scientifique. Avec seulement 808 036 paramètres, elle atteint une exactitude de 97,77 % ± 0,41, un F1-score macro de 96,77 % ± 0,41 et un kappa de Cohen de 0,967, surpassant les trois méthodes de référence ré-implémentées dans les mêmes conditions (CWT + CNN : 95,85 % ; CWT + BiLSTM : 97,44 % ; enveloppes + MLP : 95,60 %). L'écart est plus marqué en F1-macro qu'en exactitude, ce qui indique que le gain apporté par la fusion bimodale se concentre sur les classes minoritaires (VT, TR).

Mots-clés : apprentissage profond, classification d'événements sismiques, volcan Llaima, transformée en ondelettes continue, réseaux de neurones hybrides CNN-BiLSTM, fusion multimodale.

Abstract

Automated seismic monitoring of active volcanoes increasingly relies on deep learning methods able to classify, continuously and without human intervention, the different types

of events recorded by monitoring networks. This internship, carried out at the Department of Computer Engineering of the Faculty of Engineering of Kasetsart University, focused on the automatic classification of four types of seismic events from the Llaima volcano (Chile) : long-period (LP), tremor (TR), volcano-tectonic (VT) and tectonic (TC), from raw signals sampled at 100 Hz over 60-second windows.

The work first consisted in re-implementing three reference methods from the literature : a convolutional classifier applied to spectrograms (Curilem et al., 2018), a comparison of deep convolutional and recurrent architectures (Canário et al., 2020), and a multilayer-perceptron classifier applied to temporal and spectral envelope descriptors optimised with a genetic algorithm (Ait Laasri et al., 2025). These works served as the foundation for an iterative design process, documented in this report in five stages : from a simple-average CNN + BiLSTM ensemble, to a grid-search-optimised weighted ensemble, to end-to-end hybrid architectures (CNN→BiLSTM and BiLSTM→CNN), the integration of a second raw-signal input, and finally the addition of a non-linear multilayer-perceptron fusion block between the two branches. This final architecture, trained and evaluated through 5-fold stratified cross-validation on the 3,592 signals of the dataset, constitutes the original contribution of the internship and was the subject of a scientific article. With only 808,036 parameters, it reaches an accuracy of 97.77 % ± 0.41, a macro F1-score of 96.77 % ± 0.41 and a Cohen's kappa of 0.967, outperforming the three re-implemented baseline methods (CWT + CNN: 95.85 %; CWT + BiLSTM: 97.44 %; envelopes + MLP: 95.60 %). The margin is larger in macro F1 than in accuracy, indicating that the gain from bimodal fusion is concentrated on the minority classes (VT, TR).

Keywords: deep learning, seismic event classification, Llaima volcano, continuous wavelet transform, hybrid CNN-BiLSTM networks, multimodal fusion.

Table des matières

Remerciements ii

Résumé iii

Abstract iii

Liste des abréviations vii

1 Introduction générale 1

2 Contexte du stage 1

2.1 L'organisme d'accueil . . . 1
2.2 Le suivi sismo-volcanique : enjeux généraux . . . 2
2.3 Intégration dans l'équipe et périmètre du stage . . . 2
2.4 Objectifs initiaux du stage . . . 3

3 Problématique et enjeux 3

3.1 Enjeux technologiques . . . 3
3.2 Enjeux humains et sociétaux . . . 3
3.3 De la définition initiale du stage à sa réalité . . . 4

4 État de l'art : trois méthodes de référence 4

4.1 Classification par CNN de spectrogrammes (Curilem et al., 2018) . . . 5
4.2 Comparaison d'architectures profondes (Canário et al., 2020) . . . 5
4.3 Classifieur MLP sur descripteurs d'enveloppe optimisé par algorithme génétique (Ait Laasri et al., 2025) . . . 5
4.4 Synthèse et positionnement du stage . . . 6

5 Données et prétraitement 6

5.1 Le jeu de données du volcan Llaima . . . 6
5.2 Représentation temps-fréquence par transformée en ondelettes continue . . . 7
5.3 Le signal brut : seconde modalité . . . 9
5.4 Augmentation de données et régularisation . . . 9

6 Démarche méthodologique : conception itérative de l'architecture 10

6.1 Vue d'ensemble . . . 11
6.2 Étape 1 — Ensemble CNN + BiLSTM à vote moyen . . . 11
6.3 Étape 2 — Pondération optimisée de l'ensemble . . . 12
6.4 Étape 3 — Architectures hybrides bout-en-bout . . . 12

6.5 Étape 4 — Intégration du signal brut 13
6.6 Étape 5 — Fusion par bloc MLP explicite (architecture finale) 14
6.7 Protocole expérimental commun 15
6.8 Ré-implémentation rigoureuse des trois méthodes de référence 16

7 Résultats **17**
7.1 Résultats par pli 17
7.2 Résultats agrégés 18
7.3 Matrice de confusion 18
7.4 Performance par classe 20
7.5 Comparaison avec les méthodes de référence 21
7.6 Discussion 21

8 Conclusion **22**
8.1 Bilan au regard des enjeux du stage 23
8.2 Apports personnels et professionnels 23

A Rapport d'évaluation du tuteur de stage **26**

B Article scientifique **29**

Liste des abréviations

BN	Batch Normalization (normalisation par lot)
BiLSTM	Bidirectional Long Short-Term Memory (LSTM bidirectionnel)
CNN	Convolutional Neural Network (réseau de neurones convolutif)
CReLU	Concatenated ReLU
CV	Validation croisée (*cross-validation*)
CWT	Continuous Wavelet Transform (transformée en ondelettes continue)
GA	Algorithme génétique (*Genetic Algorithm*)
LP	Longue période (*Long-Period event*)
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron (perceptron multicouche)
OVDAS	Observatorio Vulcanológico de los Andes del Sur
ReLU	Rectified Linear Unit
TC	Tectonique (*Tectonic event*)
TR	Tremor
TTA	Test Time Augmentation
VT	Volcano-tectonique (*Volcano-Tectonic event*)

1 Introduction générale

Les volcans actifs font l'objet d'une surveillance sismique continue destinée à détecter, le plus tôt possible, les signaux précurseurs d'une reprise d'activité. Cette surveillance repose sur des réseaux de stations qui enregistrent en permanence le bruit sismique du volcan, dans lequel plusieurs types d'événements se superposent : des séismes liés au mouvement de fluides (magma, gaz) dans les conduits volcaniques, des séismes de rupture cassante de la roche encaissante, et des séismes tectoniques classiques sans lien direct avec le système volcanique. Distinguer automatiquement ces types d'événements, en continu et à grande échelle, est une condition nécessaire à toute analyse fine de l'état d'un volcan et à l'émission d'alertes fiables pour les populations riveraines.

C'est dans ce contexte que s'inscrit le stage dont ce rapport rend compte, centré sur la classification automatique de quatre types d'événements sismiques (longue période, tremor, volcano-tectonique et tectonique) enregistrés au volcan Llaima, dans le sud du Chili, à l'aide de méthodes d'apprentissage profond. La mission confiée s'est déroulée en deux temps : une phase de ré-implémentation et d'amélioration de méthodes existantes de la littérature, puis une phase de recherche méthodologique plus personnelle, conduite avec l'appui architectural de Piya Limcharoen, ayant abouti à la conception d'une architecture originale de fusion multimodale et à la rédaction d'un article scientifique.

Ce rapport est organisé en six parties. La Section 2 présente le contexte du stage et l'organisme d'accueil. La Section 3 en expose la problématique et les enjeux, tels qu'ils avaient été définis au départ et tels qu'ils se sont précisés au cours du stage. La Section 4 synthétise les trois méthodes de référence ré-implémentées. La Section 5 décrit le jeu de données et la chaîne de prétraitement commune à l'ensemble des expériences. La Section 6 constitue le cœur technique du rapport : elle retrace, en cinq étapes successives, la démarche de conception qui a mené de l'ensemble initial à l'architecture finale de fusion. La Section 7 présente les résultats expérimentaux obtenus, et la Section 8 conclut sur les apports du stage, tant pour l'organisme d'accueil que pour ma propre formation.

2 Contexte du stage

2.1 L'organisme d'accueil

Le stage s'est déroulé au **Department of Computer Engineering** de la **Faculty of Engineering** de **Kasetsart University**, sur le campus principal de Bangkok (50 Ngamwongwan Road, Lat Yao, Chatuchak, Bangkok 10900, Thaïlande). Fondée en 1943, Kasetsart University est l'une des plus grandes et plus anciennes universités publiques de Thaïlande ; sa Faculty of Engineering regroupe une dizaine de départements (génie informatique, aérospatial, chimique, mécanique, électrique, civil, industriel, etc.). Le De-

partment of Computer Engineering y forme aux différentes branches du génie informatique (informatique, logiciel, réseaux) et mène des activités de recherche appliquée, notamment en apprentissage automatique.

2.2 Le suivi sismo-volcanique : enjeux généraux

La classification automatique des événements sismiques volcaniques répond à un besoin opérationnel concret. Les observatoires volcanologiques, tels que l'OVDAS (Observatorio Vulcanológico de los Andes del Sur) au Chili, assurent une surveillance continue de plusieurs dizaines de volcans à partir de réseaux de stations sismiques disséminées sur chaque édifice. Le volume de données ainsi généré rend le dépouillement manuel, historiquement réalisé par des analystes experts, difficilement compatible avec une surveillance réellement continue : un système de classification automatique fiable permet de désengorger ce travail d'expertise et de le recentrer sur les cas ambigus ou les évolutions de fond, plutôt que sur le tri événement par événement.

Sur le plan physique, les quatre classes d'événements considérées dans ce stage correspondent à des origines distinctes. Les événements volcano-tectoniques (VT) traduisent une rupture cassante de la roche, sous l'effet de contraintes proches de celles observées en contexte purement tectonique, et présentent un spectre large bande pouvant atteindre une dizaine de hertz. Les événements longue période (LP) résultent de la résonance de fluides (magma, gaz, eau) circulant dans les conduits volcaniques et se caractérisent par un contenu spectral plus étroit et de plus basse fréquence. Le tremor volcanique (TR) correspond à une activité de résonance soutenue dans le temps, souvent associée à un écoulement continu de fluides, tandis que les événements tectoniques (TC) sont des séismes classiques, sans rapport direct avec la dynamique du volcan, mais enregistrés par le même réseau et qu'il convient de savoir écarter. Distinguer correctement ces quatre familles est donc un prérequis pour toute interprétation de l'état d'activité du volcan.

2.3 Intégration dans l'équipe et périmètre du stage

Nous étions trois stagiaires français de l'ENSTA Campus Brest, accueillis simultanément dans le même laboratoire, tous encadrés par le même maître de stage. Deux doctorants étaient également présents au quotidien dans le laboratoire, et le chef du département, francophone, venait régulièrement échanger avec nous, ce qui a facilité notre intégration.

Les recherches de notre tuteur portent sur l'intelligence artificielle appliquée aux capteurs, en particulier sur la manière d'exploiter les données issues de différents types de capteurs à l'aide de cette dernière. C'est dans ce cadre général qu'il a demandé à chacun des trois stagiaires de choisir un sujet spécifique. Mon projet de classification des événements sismiques du volcan Llaima s'inscrit ainsi dans cette thématique plus large portée par le laboratoire.

2.4 Objectifs initiaux du stage

Le sujet du stage, tel que défini initialement, consistait à ré-implémenter des méthodes existantes de classification d'événements sismiques du volcan Llaima, en s'appuyant sur trois travaux de référence de la littérature (Curilem et al., 2018; Canário et al., 2020; Ait Laasri et al., 2025), avant d'évoluer, une fois cette base maîtrisée, vers la proposition d'une méthode originale pouvant faire l'objet d'une valorisation scientifique. La Section 3 détaille la manière dont cet objectif s'est précisé et transformé au fil du stage.

3 Problématique et enjeux

3.1 Enjeux technologiques

Sur le plan technologique, la classification d'événements sismiques volcaniques se heurte à plusieurs difficultés qui structurent l'ensemble des choix méthodologiques présentés dans ce rapport. D'une part, les classes d'événements sont fortement déséquilibrées et présentent une grande variabilité intra-classe : deux événements de type LP peuvent différer sensiblement en amplitude, en durée effective ou en contenu fréquentiel selon l'état du système hydrothermal au moment de leur occurrence. D'autre part, la représentation du signal choisie en amont conditionne fortement les performances obtenues : un simple signal temporel brut ne rend pas explicite le contenu fréquentiel pourtant discriminant entre classes, tandis qu'une représentation temps-fréquence bien choisie peut faire apparaître des motifs directement exploitables par un réseau convolutif.

L'enjeu technologique central du stage a donc été double : choisir et calibrer une représentation du signal adaptée (ici la transformée en ondelettes continue, détaillée en Section 5), puis concevoir une architecture neuronale capable d'exploiter conjointement l'information locale en temps-fréquence, mieux captée par des couches convolutives, et les dépendances temporelles à plus longue portée, mieux captées par des couches récurrentes. La suite de ce rapport montre que cette articulation entre représentation et architecture ne se résume pas à un choix unique : elle a fait l'objet d'une démarche exploratoire progressive, chaque étape révélant de nouvelles limites (déséquilibre de paramètres entre branches, perte d'information lors d'une simple moyenne de probabilités, absence de mélange non linéaire explicite entre modalités) que l'étape suivante cherchait à corriger.

3.2 Enjeux humains et sociétaux

Au-delà de la performance brute d'un classifieur, la fiabilité d'un système de classification automatique d'événements sismiques volcaniques a une incidence directe sur la sécurité des populations vivant à proximité d'un édifice actif. Une sous-détection d'événements précurseurs (tremor soutenu, multiplication d'événements LP) peut retarder l'émission

d'une alerte, tandis qu'un système trop sensible, générant de nombreux faux positifs, tend à éroder la confiance des autorités et des populations dans le dispositif de surveillance. Cela peut créer un phénomène de lassitude face aux fausses alertes largement documenté en gestion de crise. Le volcan Llaima, en particulier, surplombe une zone rurale habitée où se pratiquent l'agriculture et l'élevage, et constitue également un site touristique : la qualité du système de classification a donc, indirectement, une incidence sur les plans de sécurité civile et sur l'activité économique locale.

Ces enjeux humains ne se traduisent pas directement en spécifications techniques mesurables, mais ils justifient certains choix méthodologiques retenus dans ce stage : la préférence donnée à des métriques équilibrées entre classes (F1-score macro, kappa de Cohen) plutôt qu'à la seule exactitude globale, qui masquerait une éventuelle faiblesse sur les classes les plus rares mais les plus critiques pour la surveillance (typiquement VT et TR).

3.3 De la définition initiale du stage à sa réalité

Le sujet du stage, tel que formulé au départ, se limitait à la ré-implémentation de méthodes existantes de classification d'événements sismiques du volcan Llaima. Cette phase initiale, décrite en Section 4, a permis de construire une compréhension fine des forces et des limites des approches déjà publiées et de disposer d'un pipeline de données et d'évaluation robuste (chargement des fichiers HDF5, transformation en ondelettes continue, validation croisée stratifiée, métriques). Cette base a naturellement ouvert la voie à une seconde phase, plus exploratoire : sur proposition et avec l'appui architectural de mon tuteur, le sujet s'est élargi à la conception d'une architecture originale. C'est cette seconde phase, retracée en cinq étapes en Section 6, qui constitue la contribution la plus personnelle du stage et qui a débouché sur la rédaction d'un article scientifique.

4 État de l'art : trois méthodes de référence

La première partie du stage a consisté à ré-implémenter trois méthodes publiées de classification d'événements sismiques du volcan Llaima, choisies pour leur complémentarité : une approche convolutive appliquée à une représentation image du signal, une étude comparative de plusieurs familles d'architectures profondes, et une approche fondée sur des descripteurs d'enveloppe combinés à un perceptron multicouche optimisé par algorithme génétique.

4.1 Classification par CNN de spectrogrammes (Curilem et al., 2018)

Curilem et al. (2018) proposent un réseau de neurones convolutif appliqué à des spectrogrammes convertis en images RGB de taille 20×20 pixels, avec une architecture volontairement simple à deux couches de convolution, pour distinguer les événements volcano-tectoniques, longue période, tremor et tectoniques du volcan Llaima. Ce travail établit qu'une représentation temps-fréquence, même à faible résolution, permet à un CNN de capturer les différences de structure locale (bande passante, amplitude) qui caractérisent chaque type d'événement. C'est cette approche qui constitue le point de départ de la branche convolutive développée durant le stage : la résolution des images a été portée à 32×32 pixels et la représentation spectrale remplacée par une transformée en ondelettes continue (Section 5), dans l'objectif d'obtenir une résolution temps-fréquence supérieure à celle d'un spectrogramme classique.

4.2 Comparaison d'architectures profondes (Canário et al., 2020)

Canário et al. (2020) mènent une comparaison systématique de huit architectures de réseaux profonds sur le même jeu de données du volcan Llaima : trois réseaux convolutifs (DCNN), trois réseaux hybrides convolutifs-récurrents (DCNN-RNN) et deux auto-encodeurs. Ce travail met en évidence que les architectures purement récurrentes testées obtiennent les résultats les plus faibles, pénalisées par un phénomène de surapprentissage lié au nombre limité d'exemples d'entraînement, alors que les architectures combinant convolution et récurrence figurent parmi les plus performantes. C'est ce constat qui a motivé, dès la première étape du stage (Section 6.2), le choix de ne pas se limiter à un unique type de réseau et de combiner une branche convolutive et une branche récurrente plutôt que de les opposer.

4.3 Classifieur MLP sur descripteurs d'enveloppe optimisé par algorithme génétique (Ait Laasri et al., 2025)

À la différence des deux méthodes précédentes, fondées sur des représentations image du signal traitées par convolution, Ait Laasri et al. (2025) proposent une approche à base de descripteurs : des caractéristiques d'enveloppe temporelle, extraites par transformée de Hilbert, et d'enveloppe spectrale, extraites par une transformée de Fourier lissée, sont calculées pour chaque événement puis fournies à un perceptron multicouche dont la sélection des descripteurs et l'optimisation des hyperparamètres sont pilotées par un algorithme génétique. Cette approche, plus légère qu'un réseau convolutif profond, a été ré-implémentée durant le stage en suivant le protocole de validation croisée décrit par les auteurs, et a servi de point de comparaison : elle permet notamment d'évaluer dans

quelle mesure l'apprentissage de représentations directement à partir du signal (par CNN et BiLSTM) apporte un gain par rapport à une extraction de descripteurs manuelle, même optimisée automatiquement.

4.4 Synthèse et positionnement du stage

Ces trois méthodes de référence occupent des points distincts dans l'espace des approches possibles : apprentissage de représentation par convolution sur image (Curilem et al., 2018), comparaison empirique entre architectures convolutives, récurrentes et non supervisées (Canário et al., 2020), et approche par descripteurs manuels optimisés (Ait Laasri et al., 2025). Le travail mené durant le stage se positionne dans le prolongement des deux premières : il conserve l'apprentissage de représentation par réseau de neurones plutôt que l'extraction de descripteurs manuels, mais cherche à dépasser l'opposition entre convolution et récurrence observée dans Canário et al. (2020) en les faisant coopérer au sein d'une architecture unique plutôt qu'en les comparant isolément, c'est précisément l'objet de la démarche en cinq étapes présentée en Section 6.

5 Données et prétraitement

5.1 Le jeu de données du volcan Llaima

Le volcan Llaima se situe sur le versant occidental des Andes, dans la région d'Araucanie au Chili (38°41' S, 71°44' O) : c'est l'un des édifices les plus actifs d'Amérique du Sud. Les signaux utilisés proviennent de la composante verticale (Z) de la station LAV, sélectionnée par les experts de l'OVDAS pour la longueur de son historique d'enregistrement et son bon rapport signal/bruit, sur la période 2010–2016 (Canário et al., 2020b). Chaque trace est échantillonnée à 100 Hz, filtrée par un filtre passe-bande de Butterworth d'ordre 10 sur la bande utile (environ 1–10 Hz), puis ramenée à 6 000 échantillons (60 secondes) par remplissage de zéros (*zero-padding*) si nécessaire. Les événements ont été individuellement inspectés, segmentés et étiquetés par des experts.

Le jeu de données comprend **3 592 signaux**, répartis en quatre classes de façon fortement déséquilibrée (Tableau 1) : les classes VT et TR, minoritaires, sont aussi les plus difficiles à apprendre.

Classe	Type d'événement	Effectif	Proportion
LP	Longue période	1 310	36,5 %
TR	Tremor	490	13,6 %
VT	Volcano-tectonique	304	8,5 %
TC	Tectonique	1 488	41,4 %
Total		**3 592**	**100 %**

Table 1: Répartition des 3 592 signaux du jeu de données Llaima par classe d'événement.

Une conséquence directe du protocole de segmentation mérite d'être signalée : la durée effective des événements varie fortement, et le reste de la fenêtre de 60 s est rempli de zéros. Les événements LP et surtout VT, souvent brefs, contiennent donc une portion « silencieuse » importante, visible sur les scalogrammes de la Figure 2 sous forme de larges zones de très faible énergie.

5.2 Représentation temps-fréquence par transformée en ondelettes continue

Chaque signal est d'abord normalisé linéairement dans l'intervalle $[0, 1]$ par une transformation min-max appliquée trace par trace, ce qui rend la représentation insensible au gain absolu de l'enregistrement (fortement dépendant de la distance à la source, et donc peu fiable comme descripteur de classe).

Plutôt que de travailler sur le spectrogramme classique utilisé par Curilem et al. (2018), le prétraitement retenu repose ensuite sur une transformée en ondelettes continue (CWT) à ondelette de Morlet complexe de paramètres `cmor1.5-1.0` (largeur de bande 1,5 ; fréquence centrale 1,0 Hz), qui offre un meilleur compromis résolution temporelle / résolution fréquentielle que la transformée de Fourier à fenêtre glissante, en particulier aux basses fréquences qui concentrent une grande partie du contenu discriminant entre classes d'événements volcaniques. Les échelles d'analyse sont réparties géométriquement pour couvrir uniformément la bande 0,5–20 Hz sur un axe logarithmique, avec $K = 32$ échelles :

$$a_k = \frac{f_c f_s}{f_k}, \qquad f_k = f_{\min} \left(\frac{f_{\max}}{f_{\min}} \right)^{k/(K-1)}, \qquad k = 0, \ldots, K-1,$$

avec $f_c = 1{,}0$ Hz, $f_s = 100$ Hz, $f_{\min} = 0{,}5$ Hz et $f_{\max} = 20$ Hz : les échelles vont ainsi de $a = 200$ (0,5 Hz) à $a = 5$ (20 Hz), avec un pas géométrique de 12,6 % d'une échelle à l'autre (environ 5,8 échelles par octave sur les 5,3 octaves couvertes).

Seul le module des coefficients complexes est conservé (la phase, peu robuste au bruit

dans cette représentation, est écartée), puis converti en décibels :

$$S(a,t) = 20\log_{10}\Big(|W(a,t)| + \varepsilon\Big), \qquad \varepsilon = 10^{-12}.$$

Cette compression logarithmique est un choix méthodologique important : les scalogrammes sismiques s'étendent sur plusieurs ordres de grandeur entre le pic d'énergie de l'arrivée et la coda ou le bruit de fond ; sur une échelle linéaire, une part dominante de l'information, précisément celle qui caractérise la décroissance lente des événements LP ou l'étalement énergétique du tremor, est écrasée près de zéro et devient invisible au réseau. L'échelle décibel rend ces structures de faible amplitude comparables, en dynamique, aux structures de forte amplitude.

La dimension temporelle est enfin réduite de 6 000 à 64 colonnes par moyennage sur des blocs contigus de $\lfloor 6000/64 \rfloor = 93$ échantillons, ce qui utilise 5 952 des 6 000 échantillons (les 48 derniers, soit 0,48 s, sont ignorés). Chaque colonne du scalogramme représente ainsi 0,93 s de signal. Le résultat est, pour chaque signal, une matrice CWT de taille 64×32 (temps $\times$ échelles), à laquelle une dimension de canal est ajoutée pour former le tenseur d'entrée $(64, 32, 1)$ de la branche convolutive 2D. Par rapport au spectrogramme à fenêtre fixe des approches antérieures (Curilem et al., 2018), cette représentation offre une résolution adaptative : fine en temps aux hautes fréquences (utile pour les arrivées impulsives des événements VT et TC) et fine en fréquence aux basses fréquences (utile pour séparer LP de TR).

Cette matrice CWT constitue la représentation d'entrée commune à l'ensemble des architectures présentées dans ce rapport :

- directement, sous forme de séquence $64 \times 32 \times 1$, pour les branches convolutives 2D et récurrentes des architectures hybrides;
- après normalisation et passage à travers une `colormap jet`, sous forme d'image RGB $32 \times 32 \times 3$ (contre 20×20 dans Curilem et al. (2018)), pour le modèle CNN de référence ré-implémenté en Section 6.8.

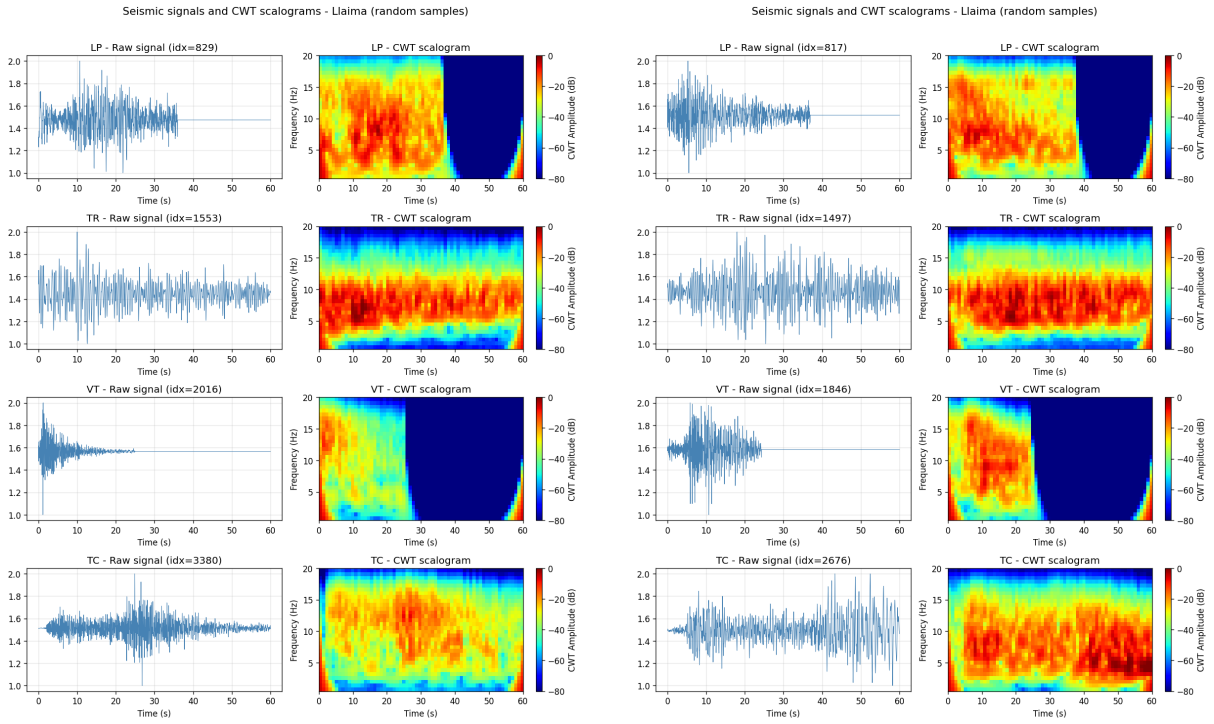


Figure 2: Deux exemples pour chaque classe (LP, TR, VT, TC) : signal brut (gauche) et scalogramme CWT en décibels (droite). Les zones bleu foncé du scalogramme correspondent au remplissage de zéros pour les événements plus courts que 60 s.

5.3 Le signal brut : seconde modalité

Le second point d'entrée du réseau final (Sections 6.5 et 6.6) est le signal lui-même, normalisé par la même transformation min-max que celle appliquée en amont de la CWT. Les deux branches décrivent ainsi exactement le même signal, sur la même échelle d'amplitude, sans que le réseau n'ait à concilier deux conventions de normalisation différentes. Aucune réduction préalable n'est appliquée : le signal brut est fourni à sa résolution native de 100 Hz (tenseur $(6000, 1)$). C'est précisément l'intérêt de cette modalité car après un moyennage sur 0,93 s, le scalogramme ne peut plus rendre compte de ce qui se joue à l'échelle de quelques centièmes de seconde. Or c'est exactement à cette échelle que se joue la distinction entre une arrivée impulsive (VT, TC) et une arrivée émergente (LP, TR).

5.4 Augmentation de données et régularisation

Afin de limiter le risque de surapprentissage sur un jeu de données de taille modeste et déséquilibré entre classes, une augmentation légère est appliquée aux données d'entraînement uniquement, une copie augmentée étant ajoutée au sein de chaque pli (facteur 1, soit un doublement de l'ensemble d'entraînement) :

- **scalogramme :** bruit gaussien additif ($\sigma = 0{,}01$) et décalage temporel circulaire aléatoire de ± 5 colonnes, soit $\pm 4{,}7$ s ;

- **signal brut :** bruit gaussien additif ($\sigma = 0{,}01$) et décalage temporel circulaire de ± 300 échantillons, soit ± 3 s.

Les amplitudes de décalage des deux modalités ont été choisies proportionnelles à leurs longueurs respectives ($5/64 \approx 300/6000$), afin que la perturbation temporelle reste d'ampleur physique comparable dans les deux vues. Les étiquettes sont par ailleurs lissées (*label smoothing*, coefficient 0,1) dans la fonction de coût, ce qui limite la sur-confiance du modèle et améliore sa calibration. Il s'agit d'une propriété utile pour un outil d'aide à la décision, dont la probabilité de sortie est destinée à être lue par un analyste et non simplement maximisée.

6 Démarche méthodologique : conception itérative de l'architecture

Cette section retrace, dans l'ordre chronologique où elles ont été conduites, les cinq étapes de conception qui ont mené de la ré-implémentation initiale des méthodes de référence à l'architecture originale présentée dans l'article scientifique produit au terme du stage. Chaque étape est motivée par une limite identifiée dans l'étape précédente : cette logique de raffinement successif est volontairement mise en avant dans ce rapport, conformément à la consigne de présenter le stage « tel qu'il est devenu a posteriori » plutôt que comme une trajectoire linéaire prévue dès le départ.

Distinction importante : prototypes personnels vs. méthodes de référence Les étapes 1 et 2 ci-dessous s'appuient sur des *prototypes personnels* de CNN et de BiLSTM, construits en tout début de stage pour prendre en main le problème et explorer des stratégies de fusion. Ces prototypes sont distincts des ré-implémentations rigoureuses des trois méthodes de référence (Curilem et al., 2018; Canário et al., 2020; Ait Laasri et al., 2025) présentées en Section 6.8, qui suivent au plus près les architectures et protocoles décrits par leurs auteurs respectifs et dont les résultats, obtenus dans les mêmes conditions expérimentales que le modèle final, figurent dans le tableau comparatif de Section 7. Autrement dit : les étapes 1–2 racontent comment l'intuition d'une architecture à deux branches complémentaires a émergé ; Section 6.8 documente la comparaison scientifique rigoureuse retenue dans l'article.

6.1 Vue d'ensemble

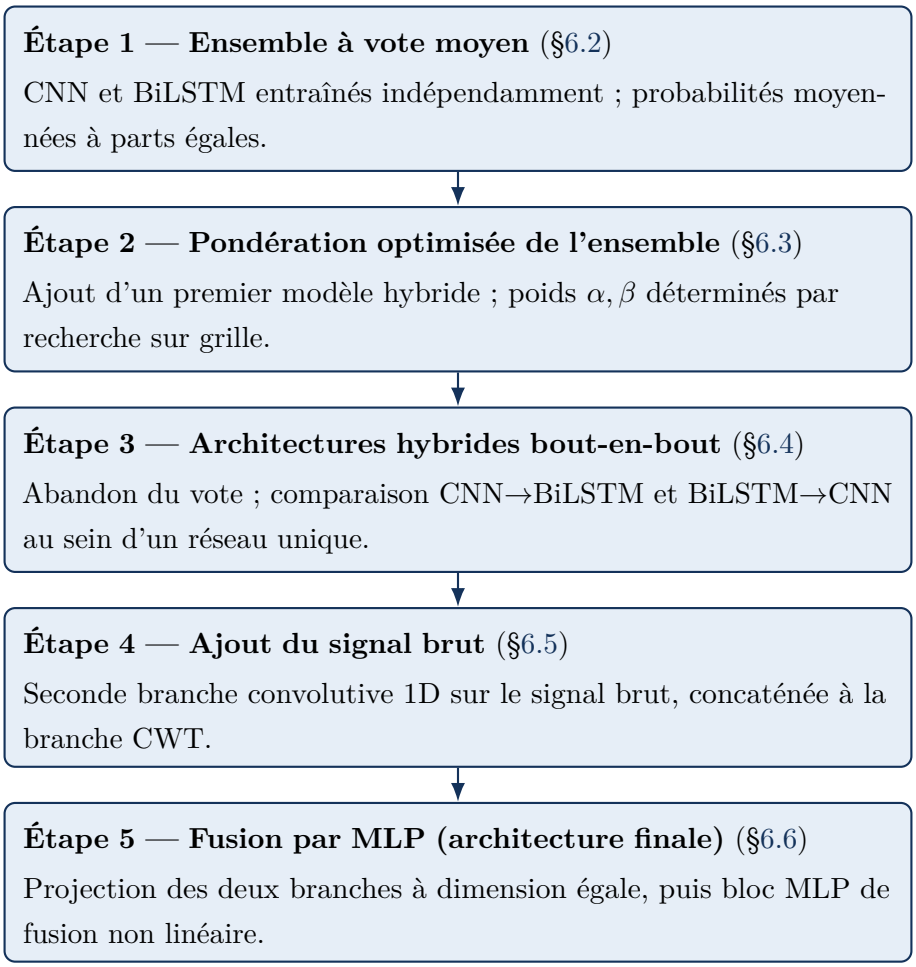


Figure 3: Vue d'ensemble de la démarche de conception itérative suivie durant le stage.

6.2 Étape 1 — Ensemble CNN + BiLSTM à vote moyen

Le tout premier prototype, construit pour prendre en main le problème avant toute recherche de méthode originale, s'inspire des deux familles d'architectures identifiées dans l'état de l'art (Section 4) : un CNN appliqué à l'image CWT et un BiLSTM appliqué à la séquence CWT, entraînés *indépendamment* l'un de l'autre, puis combinés par une simple moyenne de leurs probabilités de sortie (vote soft, à parts égales).

- **Branche CNN** — 3 blocs convolutifs (32, 64, 128 filtres 3×3), chacun suivi d'une normalisation par lot (BN) et, pour les deux premiers, d'un *max-pooling* et d'un *dropout* (0,15 puis 0,20) ; le troisième bloc est suivi d'un *global average pooling* plutôt que d'un aplatissement classique, afin de limiter le nombre de paramètres ; tête de classification : Dense(512)+BN+Dropout(0,50) → Dense(4, *softmax*).
- **Branche BiLSTM** — 2 couches BiLSTM (64 puis 32 unités), chacune suivie d'une normalisation par lot et d'une activation CReLU, qui double la dimension du vecteur de sortie en concaténant les activations positives et négatives ; un mécanisme d'auto-

attention à produit scalaire met ensuite en évidence les instants les plus discriminants de la séquence ; tête de classification :
Flatten → Dense(380)+BN+CReLU+Dropout(0,50) → Dense(4, *softmax*).

Ce premier prototype a mis en évidence un déséquilibre marqué du nombre de paramètres entre les deux branches : de l'ordre de 150 000 paramètres pour le CNN, contre plus de 3 millions pour le BiLSTM. L'origine de ce déséquilibre a pu être identifiée précisément dans le code : il provient de la couche `Dense(380)` placée après l'aplatissement (`Flatten`) de la sortie de l'attention, qui est un tenseur de forme $(64, 128)$ une fois la dimension doublée par l'activation CReLU utilisée dans ce prototype, soit 8192 valeurs en entrée de cette seule couche dense. Ce constat, obtenu en examinant la structure du réseau plutôt qu'en cherchant à équilibrer arbitrairement les deux branches, s'est révélé important pour la suite : il montre qu'un déséquilibre de paramètres peut être un simple effet de construction architecturale (un aplatissement suivi d'une couche dense large) plutôt qu'un choix de capacité délibéré, et qu'il peut être résolu en révisant la tête du réseau plutôt qu'en réduisant la taille des couches récurrentes elles-mêmes. Ce principe est repris dès l'étape 3, où l'aplatissement est remplacé par un *global average pooling*, et systématisé dans l'architecture finale (Section 6.6), qui ne comptera au total que 808 036 paramètres pour *deux* modalités d'entrée.

6.3 Étape 2 — Pondération optimisée de l'ensemble

Un simple moyennage à parts égales suppose implicitement que les deux branches sont également fiables sur l'ensemble des classes, ce qui n'a pas de raison d'être vrai en général. La deuxième étape introduit donc, une procédure de pondération : plutôt qu'une moyenne fixe, les coefficients de combinaison α (CNN), β (BiLSTM) sont déterminés par une recherche exhaustive sur grille, avec un pas de 0,05, en retenant le couple qui maximise l'exactitude sur un ensemble de validation isolé au sein de chaque pli d'entraînement.

Une alternative par apprentissage empilé (*stacking*) a également été implémentée et testée à ce stade : un petit perceptron (Dense(16) → Dense(4)) entraîné à combiner les probabilités des branches CNN et BiLSTM, plutôt que d'utiliser des poids linéaires fixes.

6.4 Étape 3 — Architectures hybrides bout-en-bout

La démarche par ensemble, bien qu'elle améliore les résultats par rapport à un modèle unique, présente une limite conceptuelle : chaque branche est entraînée séparément, sans jamais échanger d'information avec les autres avant la combinaison finale des probabilités. La troisième étape marque un changement de paradigme : plutôt que de combiner des classifieurs entraînés indépendamment, il s'agit de faire coopérer convolution et récurrence *au sein d'un seul réseau, entraîné de bout en bout.* Le vote et la recherche de poids de l'étape 2 sont abandonnés.

Deux ordres de composition sont implémentés et comparés directement, avec le même protocole expérimental, afin de trancher empiriquement plutôt que par intuition :

- **CNN → BiLSTM** — deux blocs Conv2D (32, 64 filtres) extraient des motifs locaux du scalogramme CWT, en ne réduisant la résolution que sur l'axe des fréquences (*max-pooling* $(1, 2)$) afin de préserver intégralement l'axe temporel ; les cartes de caractéristiques obtenues sont ensuite réorganisées (*reshape*) en une séquence de 64 pas de temps, transmise à deux couches BiLSTM puis à un module d'auto-attention.
- **BiLSTM → CNN** *(ordre inversé, introduit spécifiquement pour cette comparaison)* — les deux couches BiLSTM traitent directement la séquence CWT brute ; la séquence de sortie (64 pas × 64 caractéristiques, du fait de la bidirection) est réorganisée en « pseudo-image » et transmise à deux blocs de convolution 2D, suivis d'un *global average pooling*.

Les deux variantes sont évaluées face aux modèles CNN et BiLSTM autonomes de l'étape 1, sans plus aucun mécanisme de vote. C'est la variante **CNN → BiLSTM** qui a été retenue pour la suite du stage : le découpage en deux blocs de convolution suivis de deux couches BiLSTM, avec préservation stricte de l'axe temporel, est directement celui que l'on retrouve dans l'architecture finale décrite en Section 6.6. Cette étape a par ailleurs résolu, comme effet secondaire de la refonte de la tête de classification (*global average pooling* au lieu de *Flatten*+Dense large), le déséquilibre de paramètres identifié à l'étape 1.

6.5 Étape 4 — Intégration du signal brut

Sur suggestion de Piya Limcharoen, mon tuteur de stage, la piste suivante a consisté à ne plus se limiter à la représentation CWT du signal, mais à fournir également au réseau le *signal brut normalisé*, sous forme d'une seconde entrée traitée par une branche convolutive 1D dédiée. Cette branche applique quatre convolutions 1D à pas 2 (*stride*), qui réduisent progressivement la résolution temporelle de 6000 à 375 échantillons, avant un redimensionnement exact vers les 64 pas de temps de la branche CWT par interpolation bilinéaire. C'est une étape nécessaire car 6000 n'est pas un multiple entier de 64. La séquence ainsi obtenue est concaténée, sur l'axe des caractéristiques, à la sortie de la branche convolutive CWT, avant d'être transmise à la tête BiLSTM partagée, commune au modèle « CWT seule » de l'étape 3 et à ce nouveau modèle à deux entrées.

L'intérêt d'ajouter le signal brut en complément de sa transformée en ondelettes est double : la CWT, bien que riche en information temps-fréquence, implique une perte d'information par construction (passage en décibels, sous-échantillonnage temporel à 64 pas), alors que le signal brut conserve l'intégralité de la dynamique temporelle fine du phénomène ; à l'inverse, le signal brut seul ne rend pas explicite le contenu fréquentiel, ce que la branche CWT apporte. Les deux représentations sont donc complémentaires plutôt que redondantes.

6.6 Étape 5 — Fusion par bloc MLP explicite (architecture finale)

La dernière étape, qui constitue l'architecture décrite dans l'article scientifique produit à l'issue du stage, part d'un second constat de déséquilibre : à l'étape 4, la branche CWT produit, à chaque pas de temps, un vecteur de 512 caractéristiques, contre seulement 64 pour la branche signal brut, un rapport de 8 pour 1 qui, comme à l'étape 1, n'a rien d'un choix délibéré mais résulte simplement des largeurs de couches choisies indépendamment pour chaque branche. Une simple concaténation dans ces conditions revient à laisser la branche CWT dominer très largement le vecteur fusionné.

Deux modifications sont apportées pour y remédier et pour aller plus loin dans la fusion des deux modalités :

1. **Équilibrage par projection** — chaque branche est désormais suivie d'une couche Dense de projection vers une dimension commune de 256 (avec normalisation par lot), avant la concaténation. Les deux branches contribuent ainsi à parts égales au vecteur fusionné $(64, 512)$.
2. **Bloc de fusion non linéaire** — plutôt que de transmettre directement le résultat de la concaténation à la tête BiLSTM, comme à l'étape 4, un bloc de type perceptron multicouche est inséré : Dense(512) $\rightarrow$ Dense(256), chacune suivie d'une normalisation par lot et d'un *dropout* (0,20), appliqué indépendamment à chaque pas de temps (l'équivalent d'un MLP partagé, de type *TimeDistributed*). L'objectif est de permettre au réseau d'apprendre un mélange non linéaire explicite des deux modalités *avant* la modélisation temporelle par le BiLSTM, plutôt que de laisser cette dernière apprendre seule, implicitement, à la fois le mélange et la dynamique temporelle.

La Figure 4 détaille cette architecture. Le modèle complet compte **808 036 paramètres** (dont 804 292 entraînables, les 3 744 restants étant les moyennes et variances glissantes des 13 couches de normalisation par lot), répartis très inégalement entre ses quatre grands modules (Tableau 2) : le bloc de fusion MLP à lui seul en concentre près de la moitié, ce qui confirme qu'il ne s'agit pas d'un simple adaptateur dimensionnel mais bien du principal organe de calcul appris du modèle. À titre de comparaison, la branche signal brut ne représente que 5,4 % des paramètres : l'ajout d'une seconde modalité a donc un coût paramétrique marginal, l'essentiel du budget étant consacré non pas à la lecture du signal mais à sa mise en relation avec la représentation temps-fréquence.

Module	Paramètres	Part
Branche CWT (Conv2D + projection)	151 552	18,8 %
Branche signal brut (Conv1D + projection)	43 744	5,4 %
Bloc de fusion MLP	397 056	49,1 %
Tête BiLSTM + attention + classifieur	215 684	26,7 %
Total	**808 036**	**100 %**

Table 2: Répartition des paramètres de l'architecture finale par module.

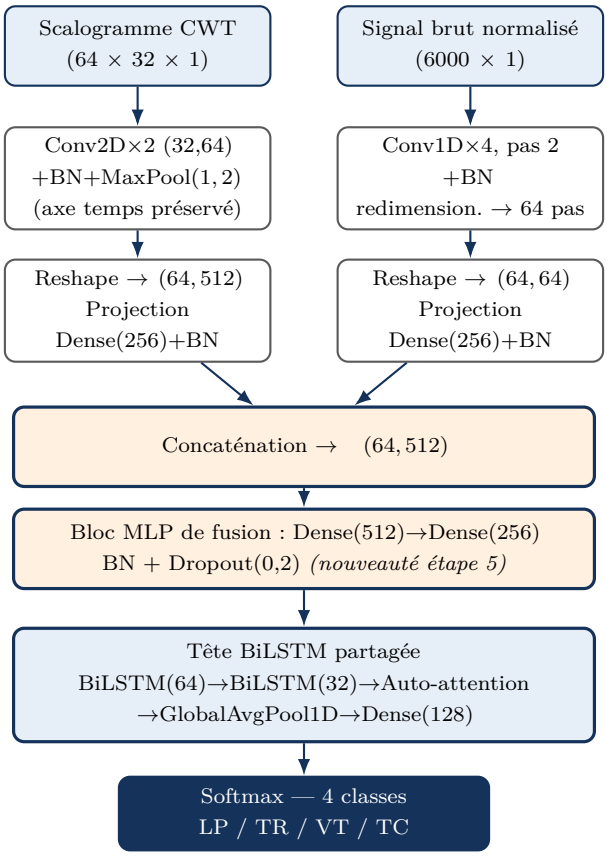


Figure 4: Architecture finale à fusion multimodale (étape 5).

6.7 Protocole expérimental commun

L'architecture finale, tout comme les trois méthodes de référence ré-implémentées pour la comparaison (Section 6.8), est évaluée selon un protocole identique : une validation croisée stratifiée à $k = 5$ plis (graine aléatoire fixée à 42), qui préserve les proportions de classes dans chaque pli et fournit, en plus de la moyenne, l'écart-type des indicateurs.

Le découpage par pli est organisé ainsi : sur les 3 592 signaux, chaque pli met de côté environ 719 signaux de test, jamais vus et jamais augmentés, et environ 2 873 signaux

d'entraînement. Ces derniers sont doublés par augmentation (≈5 747 exemples), mélangés, puis répartis en 90 % d'entraînement effectif (≈5 172) et 10 % de validation interne (≈575), cette dernière n'étant utilisée que pour piloter l'arrêt précoce et la réduction du taux d'apprentissage. Le pli de test n'intervient qu'une seule fois, en toute fin d'entraînement, pour la prédiction.

- **Optimisation** — Adam, taux d'apprentissage initial 10^{-3}, entropie croisée catégorielle avec lissage d'étiquettes (*label smoothing*) à 0,1, mini-lots de 64 exemples, jusqu'à 40 époques.
- **Callbacks** — réduction du taux d'apprentissage sur plateau (facteur 0,5, patience 5, plancher 10^{-6}) et arrêt précoce (patience 10, restauration des meilleurs poids), tous deux supervisés sur la perte de validation interne.
- **Métriques** — exactitude (*accuracy*), F1-score macro (moyenne non pondérée par classe, particulièrement informative ici puisqu'elle donne le même poids aux 304 événements VT qu'aux 1 488 événements TC), kappa de Cohen (qui corrige l'accord dû au hasard et reste interprétable malgré le déséquilibre des classes), et matrice de confusion cumulée sur les 5 plis.

6.8 Ré-implémentation rigoureuse des trois méthodes de référence

Indépendamment de la démarche exploratoire ci-dessus, et pour permettre une comparaison scientifiquement valide, les trois méthodes de référence de Section 4 ont fait l'objet d'une ré-implémentation fidèle, évaluée sur le *même* jeu de données, selon le *même* protocole de validation croisée à 5 plis et la même graine aléatoire que l'architecture finale : un CNN 2D sur image CWT 32×32 RGB, fidèle à Curilem et al. (2018) ; un BiLSTM sur séquence CWT 64×32, fidèle au meilleur pipeline CWT+LSTM de Canário et al. (2020) ; et un perceptron multicouche sur descripteurs d'enveloppe temporels et spectraux optimisés par algorithme génétique, fidèle à Ait Laasri et al. (2025). Le Tableau 3 compare le nombre de paramètres entraînables de ces trois méthodes à celui de l'architecture proposée : le CNN et le BiLSTM de référence comptent respectivement 1,4 et 1,9 fois plus de paramètres que le modèle final, pour une seule modalité d'entrée chacun, tandis que le modèle à descripteurs d'enveloppe n'en compte que 356, reflet d'une représentation manuelle compacte plutôt qu'apprise.

Rang	Méthode	Paramètres	× vs. proposé
1	Hybride bimodal, fusion MLP (**proposé**)	**808 036**	**1,0**×
2	CWT + BiLSTM (Canário et al., 2020)	1 569 872	1,9×
3	CWT + CNN 2D (Curilem et al., 2018; Canário et al., 2020)	1 142 896	1,4×
4	Enveloppes + MLP (Ait Laasri et al., 2025)	356	0,0004×

Table 3: Paramètres entraînables des méthodes comparées.

7 Résultats

Cette section présente les résultats de l'architecture finale (étape 5) et leur comparaison avec les trois méthodes de référence ré-implémentées (Section 6.8).

7.1 Résultats par pli

Le Tableau 4 détaille, pour chacun des 5 plis, l'accuracy et le F1-score macro de l'architecture proposée. La performance est stable d'un pli à l'autre, l'étendue totale ne dépassant pas 1 point d'accuracy et 1,1 point de F1 : le pli 5 est systématiquement le plus favorable sur les deux métriques à la fois, tandis que les plis 3 et 4 sont à égalité sur l'accuracy la plus faible (le pli 3 étant seul en retrait sur le F1 macro). Cette variation n'est donc pas attribuable à un déséquilibre de classes local, mais à la difficulté intrinsèque des signaux tirés dans ces plis.

Pli	Accuracy (%)	F1-macro (%)
1	97,64	96,79
2	98,19	97,18
3	97,35	96,19
4	97,35	96,44
5	98,33	97,24
Moyenne	**97,77 ± 0,41**	**96,77 ± 0,41**

Table 4: Accuracy et F1-score macro (%) par pli pour l'architecture proposée.

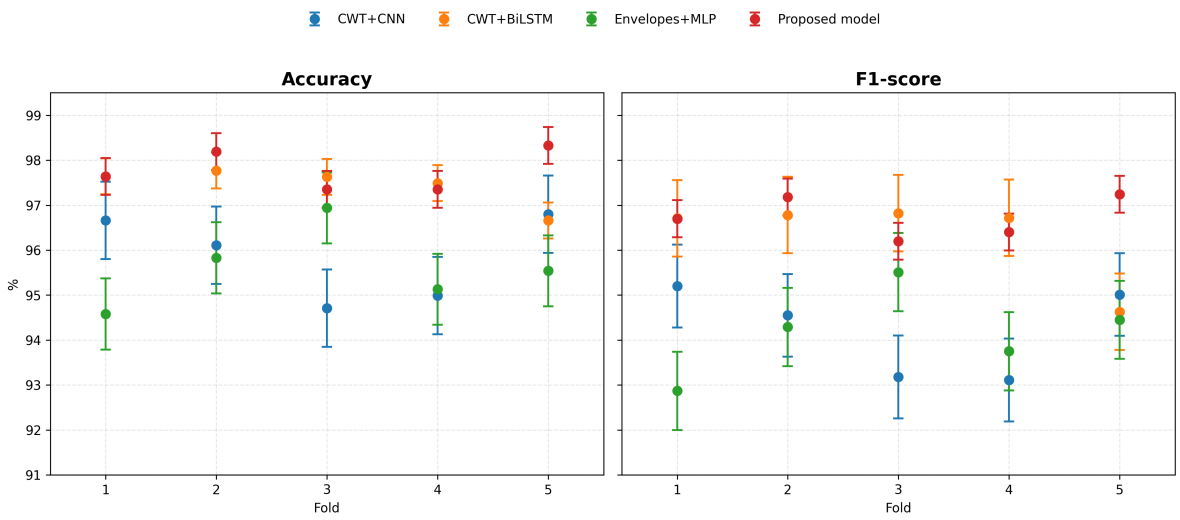


Figure 5: Accuracy (gauche) et F1-score (droite) par pli pour l'architecture proposée (rouge) et les trois méthodes de référence ré-implémentées (Section 6.8).

7.2 Résultats agrégés

Sur l'ensemble des 5 plis, le modèle atteint une **accuracy de 97,77 % ± 0,41**, un **F1-score macro de 96,77 % ± 0,41** et un **kappa de Cohen de 0,967** (Tableau 5), pour 3 512 prédictions correctes sur les 3 592 signaux évalués. L'écart-type inter-plis, inférieur à un demi-point, confirme la robustesse du système au tirage de la partition.

Le kappa de 0,967 mérite un commentaire : avec des classes de proportions 36,5 / 13,6 / 8,5 / 41,4 %, l'accord dû au hasard s'élève à 33,0 % ; le kappa mesure donc la fraction de la marge restante (67,0 %) effectivement récupérée par le classifieur. Un kappa de 0,967 signifie que 96,4 % de cette marge est captée, bien au-delà du seuil conventionnel de 0,81 qui caractérise un accord « presque parfait ».

Métrique	Valeur
Accuracy	97,77 % ± 0,41
F1-score macro	96,77 % ± 0,41
Kappa de Cohen	0,967
Signaux évalués	3 592
Prédictions correctes (cumulées)	3 512

Table 5: Performance moyenne (± écart-type) sur la validation croisée à 5 plis.

7.3 Matrice de confusion

La Figure 6 présente la matrice de confusion normalisée du modèle, cumulée sur les 5 plis ; le Tableau 6 en donne les effectifs bruts.

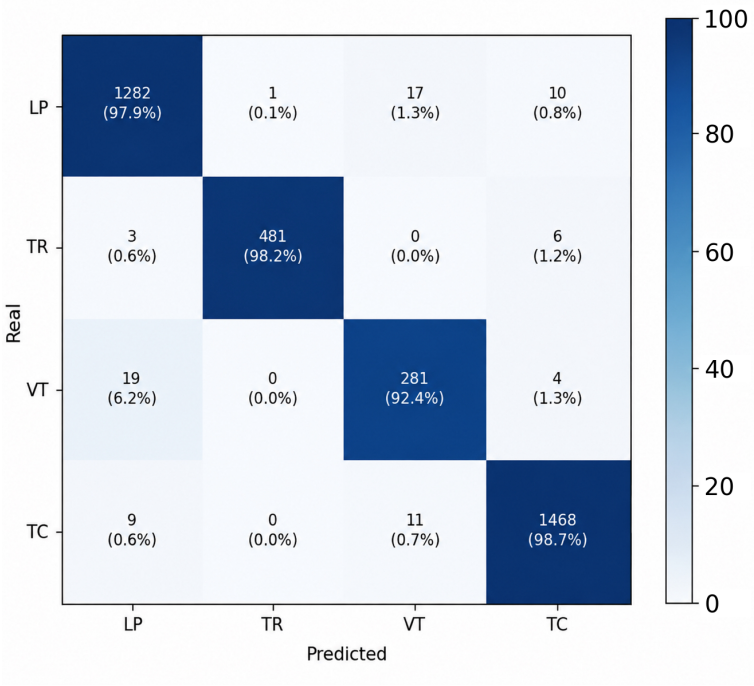


Figure 6: Matrice de confusion normalisée de l'architecture proposée, cumulée sur les 5 plis.

Réel \ Prédit	LP	TR	VT	TC	Total
LP	**1 282**	1	17	10	1 310
TR	3	**481**	0	6	490
VT	19	0	**281**	4	304
TC	9	0	11	**1 468**	1 488
Total prédit	1 313	482	309	1 488	3 592

Table 6: Matrice de confusion cumulée (effectifs ; lignes : classe réelle, colonnes : classe prédite ; diagonale en gras : prédictions correctes).

La structure des erreurs est nette et cohérente avec la physique des sources.

La classe TR est très bien séparée. Un seul signal (un événement LP) est prédit comme tremor : la colonne TR ne contient que les 481 vrais tremors plus cette unique erreur, soit une précision de 99,79 % et une spécificité de 99,97 %. Ce résultat s'explique par la nature même du tremor, seul événement quasi stationnaire du jeu de données : son énergie s'étale sur toute la fenêtre dans une bande étroite, une signature que presque aucune arrivée transitoire ne peut imiter. Les neuf tremors manqués se répartissent vers LP (3) et TC (6), vraisemblablement des segments courts ou à faible rapport signal/bruit.

Les erreurs restantes forment un triangle LP–VT–TC, qui concentre 70 des 80 erreurs totales. Le sous-graphe le plus chargé est LP $\leftrightarrow$ VT ($17 + 19 = 36$ erreurs) : c'est le recouvrement attendu entre un événement de source fluide à basses fréquences et

un événement de rupture dont la coda, une fois atténuée, prend une allure comparable. Vient ensuite LP $\leftrightarrow$ TC ($10 + 9 = 19$ erreurs), cohérent avec le fait que des événements tectoniques distants, filtrés par la propagation, perdent leurs hautes fréquences et se rapprochent spectralement des LP. Enfin, VT $\leftrightarrow$ TC ($4 + 11 = 15$ erreurs) reflète la proximité physique de deux mécanismes de rupture qui ne diffèrent essentiellement que par leur localisation.

L'asymétrie des confusions est instructive : VT est prédit comme LP dans 19 cas alors que LP n'est prédit comme VT que dans 17 cas, alors même que les tailles de classes diffèrent d'un facteur 4,3. Rapporté à la taille de classe, 6,25 % des événements VT sont classés LP contre seulement 1,30 % des LP classés VT : en cas de doute, le modèle penche vers la classe majoritaire. Il s'agit d'un biais attendu pour un classifieur entraîné sans repondération explicite des classes.

7.4 Performance par classe

Le Tableau 7 détaille les indicateurs par classe, calculés sur la matrice cumulée. Trois classes sur quatre dépassent 97 % de F1, et la classe TR atteint 98,97 %. La classe VT demeure la difficulté résiduelle, avec un F1 de 91,68 % — près de six points sous les trois autres.

Classe	Rappel	Précision	F1	Spécificité	Taux d'erreur
LP	97,86 %	97,64 %	97,75 %	98,64 %	1,64 %
TR	98,16 %	99,79 %	98,97 %	99,97 %	0,28 %
VT	92,43 %	90,94 %	91,68 %	99,15 %	1,42 %
TC	98,66 %	98,66 %	98,66 %	99,05 %	1,11 %
Macro	**96,78 %**	**96,76 %**	**96,77 %**	**99,20 %**	—

Table 7: Performance par classe de l'architecture proposée, cumulée sur les 5 plis.

Ce déficit sur VT a une double explication. D'abord la rareté : avec 304 exemplaires (8,5 % du jeu de données), la classe VT ne dispose que d'environ 243 signaux d'entraînement par pli avant augmentation, un ordre de grandeur de moins que TC. Ensuite l'ambiguïté physique : contrairement à TR, dont la signature quasi stationnaire est unique, un événement VT partage son mécanisme de source (rupture cassante) avec les événements TC et sa coda basse fréquence avec les événements LP : la classe VT occupe ainsi précisément l'intersection des deux autres régions du plan temps-fréquence, une position structurellement défavorable. Rappel (92,43 %) et précision (90,94 %) y sont à des niveaux proches : le modèle ne sacrifie pas l'un pour l'autre, il manque simplement de séparabilité sur cette classe.

7.5 Comparaison avec les méthodes de référence

Le Tableau 8 rassemble les résultats de l'architecture proposée et des trois méthodes de référence ré-implémentées dans les mêmes conditions (Section 6.8).

Méthode	Accuracy	F1-macro	Kappa
CWT + CNN 2D (Curilem et al., 2018; Canário et al., 2020)	95,85 %	94,23 %	0,938
CWT + BiLSTM (Canário et al., 2020)	97,44 %	96,33 %	0,962
Enveloppes + MLP (Ait Laasri et al., 2025)	95,60 %	94,17 %	0,934
Hybride bimodal, fusion MLP (proposé)	**97,77 %**	**96,77 %**	**0,967**

Table 8: Comparaison avec les méthodes de référence ré-implémentées, même jeu de données et même validation croisée à 5 plis.

Le modèle proposé surpasse les trois méthodes de référence sur les deux métriques. L'écart est net face au classifieur à descripteurs d'enveloppe, limité par le pouvoir descriptif d'une représentation manuelle ; il est plus modeste face au CNN sur scalogramme.

Cette dissymétrie entre les deux métriques est le résultat le plus instructif du tableau : le F1 macro progresse nettement plus que l'accuracy par rapport à la meilleure référence. Or l'accuracy est dominée par les classes majoritaires (LP et TC représentent ensemble 78 % du jeu de données) alors que le F1 macro donne un poids égal à chaque classe. Un gain concentré sur le F1 macro signifie donc mécaniquement que l'amélioration porte sur les classes minoritaires, VT et TR. C'est-à-dire précisément là où une représentation unique atteint ses limites et où une seconde vue du signal est la plus décisive. L'apport du signal brut ne se lit pas dans le nombre total de prédictions corrigées, mais dans leur distribution.

7.6 Discussion

La complémentarité des modalités est à l'origine du gain. Le scalogramme et le signal brut ne sont pas deux encodages redondants de la même information : le premier est un module, il écarte la phase. Il est moyenné sur des fenêtres de 0,93 s ; il décrit où se situe l'énergie dans le plan temps-fréquence mais reste aveugle à la structure fine du signal à l'intérieur de chaque colonne. Le second conserve intégralement cette structure fine, polarité et raideur de l'arrivée comprises, mais n'expose aucune organisation fréquentielle. Que le gain se concentre sur le F1 macro plutôt que sur l'accuracy est cohérent avec cette lecture : les classes majoritaires étaient déjà bien séparées par la seule représentation temps-fréquence, tandis que les classes rares et ambiguës bénéficient de l'information supplémentaire.

L'alignement temporel est une nécessité, pas un détail d'implémentation. Concaténer deux séquences n'a de sens que si leurs indices désignent le même instant physique : c'est ce qui justifie le *max-pooling* asymétrique $(1, 2)$ de la branche CWT, qui interdit toute réduction de l'axe temporel, et le redimensionnement bilinéaire de la branche 1D, préféré à un moyennage par blocs impossible puisque 375 n'est pas divisible par 64. Sans ces deux précautions, la concaténation mélangerait des instants décalés et le bloc de fusion apprendrait des corrélations parasites.

L'équilibrage dimensionnel conditionne l'exploitation effective de la seconde modalité. Le rapport naturel de 512 pour 64 entre les deux branches aurait relégué le signal brut au rang de perturbation marginale ; les projections dans un espace commun à 256 dimensions garantissent que la fusion s'opère entre deux modalités de statut équivalent. Alors même que la branche signal brut ne représente que 5,4 % des paramètres : son influence tient à sa position dans l'architecture, non à sa capacité.

Séparer fusion inter-modale et intégration temporelle est efficace. En confiant le mélange non linéaire des modalités à un MLP partagé, et l'intégration temporelle à la tête BiLSTM, le problème d'apprentissage se décompose en deux sous-problèmes, chacun avec sa propre capacité et sa propre régularisation. Le MLP de fusion mobilise à lui seul 49 % des paramètres du réseau, confirmant qu'il s'agit du principal organe de calcul appris, non d'un simple adaptateur.

L'analyse par classe localise précisément la difficulté résiduelle. La classe VT, la plus rare (8,5 %) et structurellement intermédiaire entre LP et TC, concentre les erreurs rapportées à sa taille et présente le seul F1 sous 97 %. À l'inverse, la classe TR atteint 99,79 % de précision et 99,97 % de spécificité : presque aucun événement des autres familles n'est jamais confondu avec un tremor, confirmant que sa signature quasi stationnaire n'a pratiquement pas d'équivalent dans le jeu de données.

Pistes d'amélioration identifiées : (i) un rééquilibrage explicite des classes, par pondération de la fonction de coût ou par augmentation ciblée des événements VT ; (ii) l'ajout d'une troisième modalité, par exemple les descripteurs d'enveloppe de Ait Laasri et al. (2025), injectable dans le même bloc de fusion MLP sans modification structurelle ; (iii) le remplacement de l'auto-attention sans paramètres par une attention à projections apprises, dont le surcoût resterait modeste au regard des 808 036 paramètres actuels. Une limite de portée est également assumée : l'évaluation porte sur une seule station d'un seul volcan ; la transférabilité à d'autres réseaux de surveillance, où les effets de site et les conditions d'enregistrement diffèrent, reste à établir.

8 Conclusion

8.1 Bilan au regard des enjeux du stage

Le stage avait pour point de départ un objectif circonscrit : ré-implémenter et améliorer trois méthodes publiées de classification d'événements sismiques du volcan Llaima (Curilem et al., 2018; Canário et al., 2020; Ait Laasri et al., 2025). Cet objectif a été atteint, et a servi de socle à un travail plus personnel : la conception, en cinq étapes documentées en Section 6, d'une architecture originale de fusion multimodale entre représentation temps-fréquence (CWT) et signal brut, dont les résultats sont discutés en Section 7. Cette trajectoire, passer d'un simple ensemble à vote moyen à une architecture bout-en-bout où les deux modalités sont explicitement mélangées par un bloc non linéaire avant modélisation temporelle, répond directement à l'enjeu technologique central identifié en Section 3 : exploiter conjointement l'information locale temps-fréquence et les dépendances temporelles longues, plutôt que de les traiter comme deux problèmes séparés.

Rapportés aux enjeux posés en Section 3, les résultats obtenus : 97,77 % d'accuracy, 96,77 % de F1 macro et un kappa de 0,967, pour seulement 808 036 paramètres répondent concrètement à l'enjeu technologique : le gain par rapport aux trois méthodes de référence se concentre sur les classes minoritaires et physiquement ambiguës (VT, TR), c'est-à-dire exactement les cas où l'enjeu humain défini en Section 3 est le plus sensible. En effet, une sous-détection de ces événements étant celle qui retarderait le plus une alerte pertinente. Le fait que l'architecture finale reste quatre fois plus légère que la référence BiLSTM seule, malgré deux modalités d'entrée, va par ailleurs dans le sens de l'enjeu organisationnel d'un système déployable en continu plutôt que d'un prototype de laboratoire.

8.2 Apports personnels et professionnels

Sur le plan technique, ce stage a permis d'approfondir la conception et l'entraînement de réseaux de neurones profonds, au-delà de l'application de recettes standards : construction de mécanismes personnalisés (activation CReLU dans les prototypes initiaux, module d'auto-attention repris jusqu'à l'architecture finale), diagnostic de problèmes d'architecture à partir du nombre de paramètres et de la structure du réseau plutôt que par essais-erreurs : la lecture du déséquilibre CNN/BiLSTM de l'étape 1 (Section 6.2) en est l'exemple le plus direct, et mise en place rigoureuse d'un protocole de validation croisée et de métriques adaptées à un problème multi-classes déséquilibré. La nécessité de comparer objectivement plusieurs variantes architecturales (Section 6.4) a également renforcé une démarche expérimentale rigoureuse : isoler une seule variable à la fois (le sens de composition CNN/BiLSTM, puis l'ajout du signal brut, puis le mode de fusion) plutôt que de faire évoluer plusieurs choix simultanément, et ré-implémenter les méthodes de référence dans des conditions strictement comparables (Section 6.8) plutôt que de se fier à des chiffres publiés obtenus sur d'autres protocoles.

Le passage d'un travail de ré-implémentation à la conception d'une méthode originale,

puis à sa formalisation sous la forme d'un article scientifique co-signé avec mon tuteur de stage, constitue par ailleurs une première expérience complète de recherche appliquée : elle a notamment nécessité de savoir situer une contribution technique par rapport à l'état de l'art (Section 4), de conduire une comparaison expérimentale rigoureuse (Section 7), et d'en communiquer la démarche de façon structurée, des compétences directement réinvesties dans la rédaction de ce rapport.

Références

M. Curilem, J. P. Canário, L. Franco, R. A. Rios. *Using CNN to Classify Spectrograms of Seismic Events From Llaima Volcano (Chile)*. In *2018 International Joint Conference on Neural Networks (IJCNN)*, p. 1–8, IEEE, 2018. https://doi.org/10.1109/IJCNN.2018.8489285

J. P. Canário, R. Mello, M. Curilem, F. Huenupan, R. Rios. *In-depth comparison of deep artificial neural network architectures on seismic events classification*. Journal of Volcanology and Geothermal Research, 401, article 106881, 2020. https://doi.org/10.1016/j.jvolgeores.2020.106881

J. P. Canário, R. F. de Mello, M. Curilem, F. Huenupan, R. A. Rios. *Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification*. Data in Brief, 30, article 105627, 2020.

E. H. Ait Laasri, A. Atmani, E.-S. Akhouayri, D. Agliz. *A simple and effective MLP-based seismic signal classifier using temporal and spectral envelope features with genetic algorithm-optimization*. Measurement, 247, article 116776, 2025.

A Rapport d'évaluation du tuteur de stage

RAPPORT D'EVALUATION
ASSESSMENT REPORT

Merci de retourner ce rapport par courrier ou par voie électronique en fin du stage à :
At the end of the internship, please return this report via mail or email to:

ENSTA – Bureau des stages - 2 rue François Verny - 29806 BREST cedex 9 – FRANCE
☎ 00.33 (0) 2.98.34.87.70 / stages.brest@ensta.fr

I - ORGANISME / *HOST ORGANISATION*

NOM / *Name*_____ Kasetsart University_____

Adresse / *Address*_____Department of Computer Engineering, Kasetsart University 50 Ngamwongwan Rd, Chatuchak Bangkok 10900 Thailand_____

Tél / *Phone (including country and area code)*_____Tel. (+66) 0-2797-0999 ext 1403-04 1450___

Nom du superviseur / *Name of internship supervisor*
_____Piya Limcharoen_____
Fonction / *Function*_____Lecture_____

Adresse e-mail / *E-mail address*_____piya.lim@ku.th_____

Nom du stagiaire accueilli / *Name of intern* Juliette Lourdault

II - EVALUATION / *ASSESSMENT*

Veuillez attribuer une note, en encerclant la lettre appropriée, pour chacune des caractéristiques suivantes. Cette note devra se situer entre **A (très bien)** et **F (très faible)**
*Please attribute a mark from **A (excellent)** to **F (very weak).***

.1 MISSION / *TASK*

❖ La mission de départ a-t-elle été remplie ? A **B** C D E F
Was the initial contract carried out to your satisfaction?

❖ Manquait-il au stagiaire des connaissances ? oui/*yes* **non/*no***
Was the intern lacking skills?

Si oui, lesquelles ? / *If so, which skills?* _____

.2 ESPRIT D'EQUIPE / *TEAM SPIRIT*

❖ Le stagiaire s'est-il bien intégré dans l'organisme d'accueil (disponible, sérieux, s'est adapté au travail en groupe) / *Did the intern easily integrate the host organisation? (flexible, conscientious, adapted to team work)*
A **B** C D E F

Souhaitez-vous nous faire part d'observations ou suggestions ? / *If you wish to comment or make a suggestion, please do so here*_____
_____

Version du 05/04/2019

.3 COMPORTEMENT AU TRAVAIL / *BEHAVIOUR TOWARDS WORK*

Le comportement du stagiaire était-il conforme à vos attentes (Ponctuel, ordonné, respectueux, soucieux de participer et d'acquérir de nouvelles connaissances) ?
Did the intern live up to expectations? (*Punctual, methodical, responsive to management instructions, attentive to quality, concerned with acquiring new skills)?*

A **B** C D E F

Souhaitez-vous nous faire part d'observations ou suggestions ? / *If you wish to comment or make a suggestion, please do so here* _____
_____

.4 INITIATIVE – AUTONOMIE / *INITIATIVE – AUTONOMY*

Le stagiaire s'est –il rapidement adapté à de nouvelles situations ? A B C D E F
(Proposition de solutions aux problèmes rencontrés, autonomie dans le travail, etc.)

Did the intern adapt well to new situations? A **B** C D E F
(eg. suggested solutions to problems encountered, demonstrated autonomy in his/her job, etc.)

Souhaitez-vous nous faire part d'observations ou suggestions ? / *If you wish to comment or make a suggestion, please do so here* _____
_____

.5 CULTUREL – COMMUNICATION / *CULTURAL – COMMUNICATION*

Le stagiaire était-il ouvert, d'une manière générale, à la communication ? A **B** C D E F
Was the intern open to listening and expressing himself /herself?

Souhaitez-vous nous faire part d'observations ou suggestions ? / *If you wish to comment or make a suggestion, please do so here* _____
_____

.6 OPINION GLOBALE / *OVERALL ASSESSMENT*

❖ La valeur technique du stagiaire était : A **B** C D E F
Please evaluate the technical skills of the intern:

III - PARTENARIAT FUTUR / *FUTURE PARTNERSHIP*

❖ Etes-vous prêt à accueillir un autre stagiaire l'an prochain ?

Would you be willing to host another intern next year? **oui/*yes*** non/*no*

Fait à _____, le _____
In _Kasetsart University, Bangkok__, on__Starting from April (same period as this year)__

Signature Entreprise Piya Digitally signed by Piya Limcharoen DN: CN=Piya Limcharoen, O=CPE KU, OU=CPE KU, E=piya.lim@ku.th Reason: I am the author of this documen Location: Date: 2026.09.08 09:25:37+07'00' Foxit PDF Editor Version: 2026.1.3 Signature stagiaire
Company stamp _____ *Intern's signature* J Lourdault

Merci pour votre coopération
We thank you very much for your cooperation

B Article scientifique

L'article scientifique produit à l'issue du stage, *Automatic Classification of Seismo-Volcanic Events with a Bimodal CNN–BiLSTM Network and MLP Fusion: Continuous Wavelet Scalogram and Raw Waveform — Application to Llaima Volcano (Chile)*, co-signé avec Piya Limcharoen, est joint en annexe dans son intégralité.

Automatic Classification of Seismo-Volcanic Events with a Bimodal CNN–BiLSTM Network and MLP Fusion: Continuous Wavelet Scalogram and Raw Waveform — Application to Llaima Volcano (Chile)

Juliette Lourdault Piya Limcharoen

Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
ENSTA, Institut Polytechnique de Paris, 29200 Brest, France
Corresponding author: juliette.lourdault@ensta.fr

4 September 2026

Abstract

Monitoring active volcanoes relies on the classification of a steadily growing volume of seismic recordings, a task that the sheer data rate makes impossible to perform manually in real time. We present an automatic classifier for the four event families recorded at Llaima volcano (Chile) — long-period (LP), tremor (TR), volcano-tectonic (VT) and tectonic (TC) — based on a *bimodal* deep network that jointly processes two complementary representations of the same recording. The first is a scalogram obtained from the continuous wavelet transform (CWT) with a complex Morlet wavelet, expressed in decibels; the second is the normalised raw waveform at its native 100 Hz resolution. Each modality is encoded by a dedicated convolutional branch — two-dimensional convolutions with a preserved time axis for the scalogram, strided one-dimensional convolutions for the waveform — and then projected onto a common 256-dimensional space, so that both branches contribute equally to the fusion. The two sequences, aligned on 64 time steps, are concatenated and passed through a *time-distributed multilayer perceptron (MLP) fusion block* that performs an explicit non-linear mixing of the two modalities before any temporal modelling takes place. The fused sequence finally feeds a two-layer bidirectional long short-term memory (BiLSTM) head, augmented with scaled dot-product self-attention and global average pooling. Evaluated by stratified five-fold cross-validation on 3,592 signals, the model reaches an accuracy of 97.77 %, a macro F1-score of 96.77 % and a Cohen's kappa of 0.967, using only 808,036 parameters. It outperforms the three re-implemented baseline methods (CWT + CNN: 95.85 %; CWT + BiLSTM: 97.44 %; envelopes + MLP: 95.60 %), and the margin in macro F1 is substantially larger than the margin in accuracy, indicating that the gain is concentrated on the minority classes.

Keywords: volcano monitoring; seismic event classification; continuous wavelet transform; multimodal fusion; 1D convolutional network; BiLSTM; attention mechanism; Llaima volcano.

Contents

1 Introduction **2**

2 Dataset **3**

3 Methodology **4**
3.1 Time–frequency representation by wavelets 6
3.2 Raw waveform: the second modality 7
3.3 2D convolutional branch on the scalogram 7
3.4 1D convolutional branch on the raw waveform 8
3.5 Branch balancing and concatenation 9
3.6 MLP fusion block 9
3.7 BiLSTM head with self-attention 10
3.8 Parameter budget 11
3.9 Regularisation and training 12

4 Experimental protocol **12**

5 Results **13**
5.1 Per-fold results 13
5.2 Aggregated results 14
5.3 Confusion matrix 14
5.4 Per-class performance 16
5.5 Comparison with baseline methods 16

6 Discussion **17**

7 Conclusion **18**

1 Introduction

Volcanic activity manifests itself through physical processes of fluid and energy transport, part of which is released as discrete seismic oscillations. Recognising the signatures associated with these oscillations is an essential source of information for the remote monitoring of a volcanic edifice. This recognition nonetheless remains delicate: source geometry, ground properties and site effects shift the phase, delay, amplify or attenuate the frequency components of the recorded signals, blurring boundaries that are already tenuous between event families.

At Llaima volcano, the *Observatorio Vulcanológico de los Andes Sur* (OVDAS) distinguishes four main event families: long-period (LP), tremor (TR), volcano-tectonic (VT) and tectonic (TC). Manual sorting of these signals by analysts does not scale to the continuous stream produced by the stations, hence the interest of a reliable automatic classifier able to support decision-making in near real time.

Two broad families of approaches emerge from the literature. The first extracts descriptors (amplitude statistics, spectral and phase parameters, wavelet coefficients, envelopes) that feed a conventional classifier; the second lets a deep network learn the discriminant patterns directly from the raw signal or from a time–frequency representation. On the Llaima dataset, three works provide natural points of comparison: the CNN classification of spectrograms by Curilem et al. (IJCNN 2018) [2], the in-depth comparative study of deep architectures by Canário et al. (JVGR 2020) [3], which notably retains a CWT + LSTM pipeline, and the envelope-descriptor MLP classifier optimised by a genetic algorithm of Ait Laasri et al. (Measurement 2025) [4].

All of these approaches share an implicit assumption: that a single representation of the signal is sufficient. Yet the two usual domains do not carry the same information, nor do they carry it in the same way. The CWT scalogram explicitly renders the distribution of energy in the time–scale plane, but it relies on the modulus of the coefficients and therefore *discards the phase*; it further results from a smoothing operation (block averaging in time) that attenuates the shortest transients. The raw waveform, conversely, retains the entire signal shape — polarity, steepness of the onset, fine structure of the first arrival, phase coherence between successive oscillations — but exposes no frequency structure directly usable by a classifier. The two views are therefore complementary by construction, not redundant.

We propose to process them jointly within a *single bimodal network trained end to end*, rather than combining a posteriori the decisions of independent models. The contribution of this work is fourfold:

1. **A bimodal architecture with explicit temporal alignment.** The scalogram is encoded by a 2D convolutional branch whose pooling operations act on the scale axis only, preserving all 64 time steps; the raw waveform is encoded by a strided 1D convolutional branch and then resampled onto exactly the same 64 steps. Both sequences then describe, time step by time step, the same physical segment of the signal, which makes their fusion semantically coherent.

2. **Dimensional balancing of the two modalities.** Without care, the CWT branch naturally produces 512 features per time step against 64 for the waveform branch, an eight-to-one ratio that would relegate the latter to a marginal role. A shared dense projection brings both branches to 256 dimensions before concatenation, guaranteeing an equal contribution.

3. **A time-distributed MLP fusion block**, inserted between the concatenation and the recurrent head. This module learns an explicit non-linear mixing of the two modalities *before* temporal modelling, instead of leaving the BiLSTM to handle cross-modal fusion and temporal integration simultaneously.

4. **A compact recurrent head with self-attention**, whose global average pooling replaces sequence flattening: the complete model has only 808,036 parameters, roughly one quarter of a comparably sized BiLSTM with a flattened head.

The whole system is evaluated by stratified five-fold cross-validation and compared, on the same dataset and under the same protocol, with the three baseline methods above.

The remainder of the paper describes the dataset (Section 2), the proposed methodology (Section 3), the experimental protocol (Section 4), the results (Section 5), and then discusses them (Section 6) before concluding (Section 7).

2 Dataset

The signals come from Llaima volcano, located on the western slope of the Andes in the Araucanía region of Chile (S 38°41′ – W 71°44′), one of the most active edifices in South America. They were recorded on the vertical component (Z) of station LAV, selected by OVDAS specialists for its long operational record and good signal-to-noise ratio, over the period 2010–2016 [1]. Each trace is sampled at 100 Hz, filtered by a 10th-order Butterworth band-pass filter over the useful band (approximately 1–10 Hz), and then brought to 6,000 samples, i.e. 60 seconds, by zero-padding where necessary. Events were individually inspected, segmented and labelled by experts.

The four families exhibit distinct time–frequency signatures. Volcano-tectonic (VT) events are associated with rock fracturing in the conduits; their broadband spectrum may reach 10 Hz, with an impulsive onset and an exponential decay. Long-period (LP) events, related to the transit of magmatic and hydrothermal fluids, display a narrower spectrum, mainly between 0.5

and 5 Hz, with a weakly impulsive onset and a slow decay. Tremor (TR) is characterised by a long, regularly shaped signal with a narrow spectrum (typically 0.5–3 Hz) and slow attenuation. Finally, tectonic (TC) events, unrelated to volcanic activity, result from fault dynamics; their spectral content, close to that of VT events, generally carries more energy, and depending on epicentral proximity they may be confused with LP or VT events.

The dataset comprises 3,592 signals whose distribution (Table 1) is markedly imbalanced: the VT and TR classes are in the minority, which makes them the hardest cases to learn.

Table 1: Distribution of the 3,592 signals of the Llaima dataset by event class.

Class	Event type	Count	Proportion
LP	Long-period	1,310	36.5 %
TR	Tremor	490	13.6 %
VT	Volcano-tectonic	304	8.5 %
TC	Tectonic	1,488	41.4 %
Total		**3,592**	**100 %**

One direct consequence of the segmentation protocol deserves to be flagged at this stage: the effective duration of the events varies widely, and the remainder of the 60-second window is filled with zeros. LP and especially VT events, often short, therefore contain a substantial silent portion, clearly visible in the scalograms of Fig. 2 as broad regions of very low energy. This point is revisited in Section 6.

3 Methodology

The proposed method chains five stages: (i) construction of the two input modalities from the raw trace, (ii) their encoding by two specialised convolutional branches, (iii) their projection onto a common space and temporal concatenation, (iv) their non-linear mixing by a time-distributed MLP fusion block, and (v) temporal modelling and classification by a BiLSTM head with self-attention. The whole forms a single network trained end to end by backpropagation. Fig. 1 gives an overview.

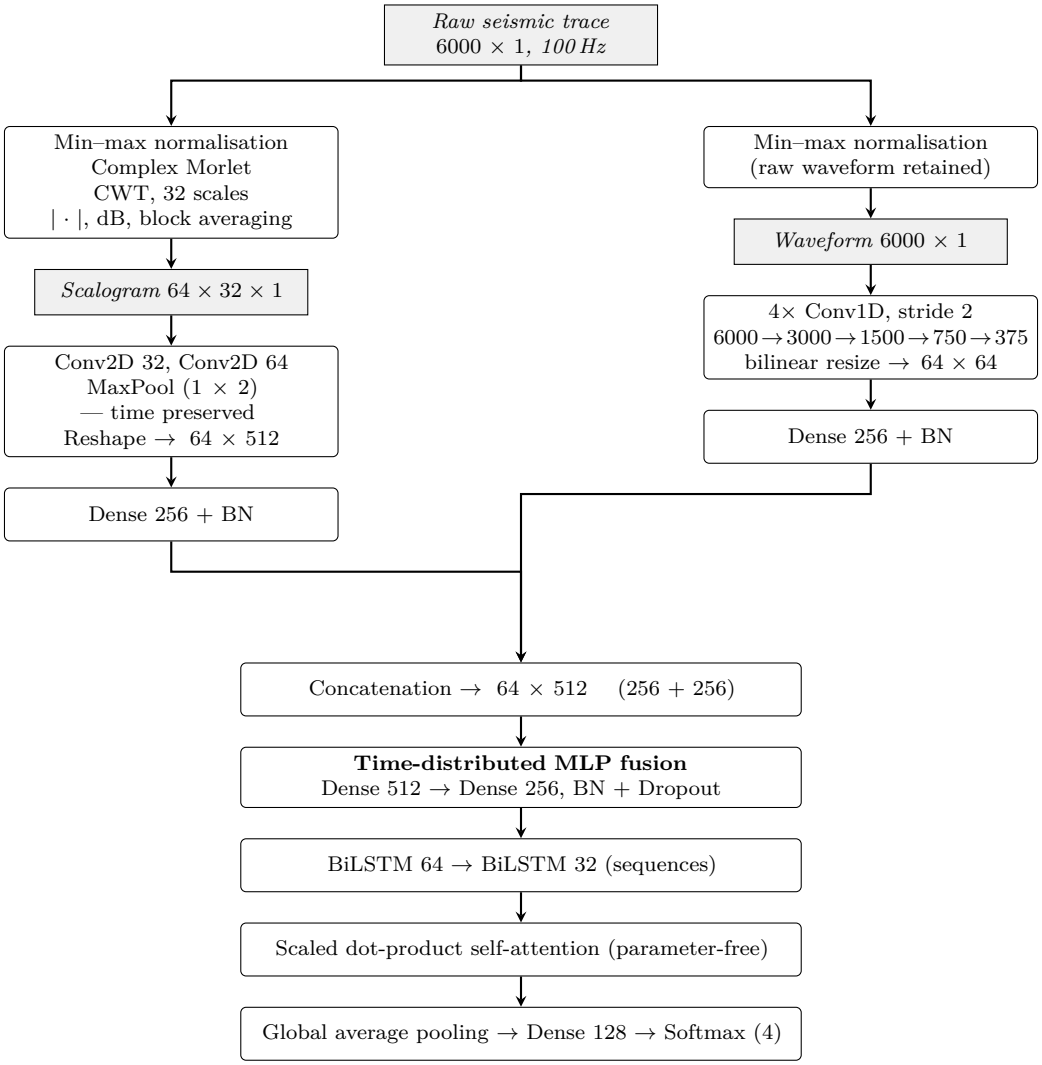


Figure 1: Overview of the proposed bimodal architecture. The two branches are aligned on the same 64 time steps and projected onto the same 256-dimensional space before fusion. BN: batch normalisation.

Fig. 2 illustrates, for one representative of each class, the raw signal (input of the 1D branch) and the corresponding CWT scalogram (input of the 2D branch). The third column, showing the 32×32 RGB rendering in the *jet* colormap, *is not used by the proposed model*: it is reproduced for reference only, as it constitutes the input of baseline method [2] re-implemented in Section 5.5.

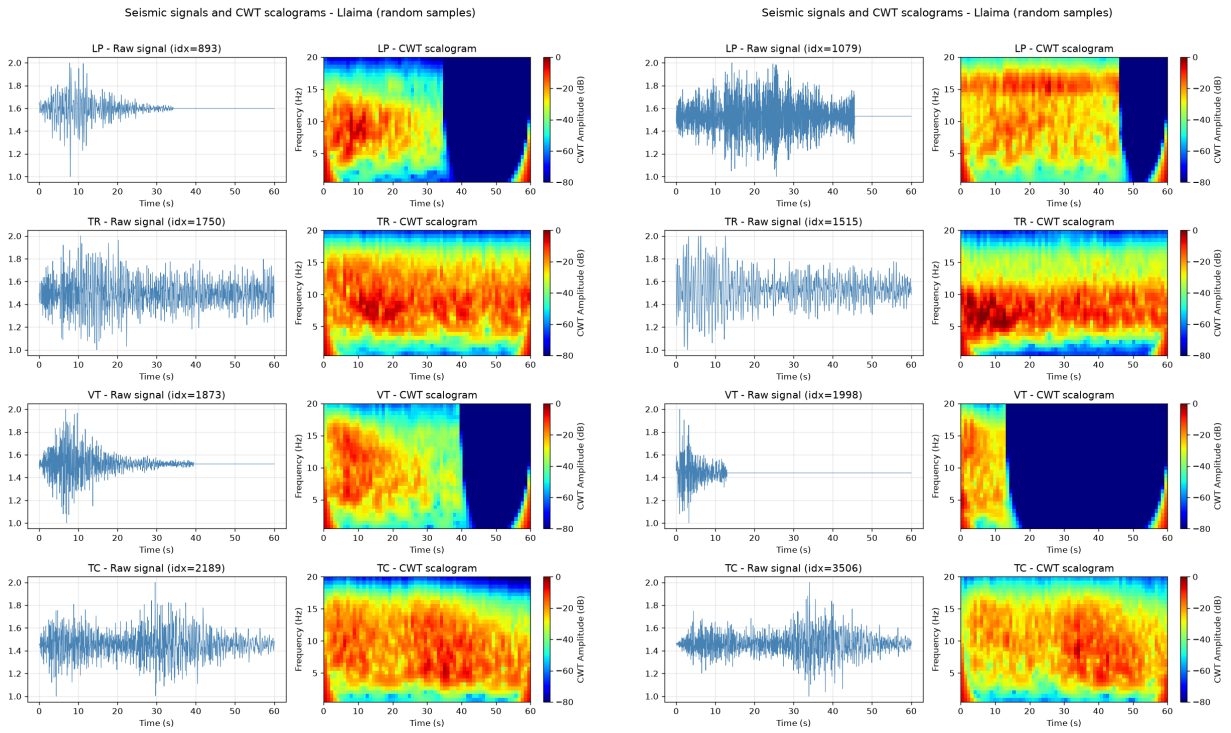


Figure 2: For each class (LP, TR, VT, TC), two example : raw signal (left), CWT scalogram with a complex Morlet wavelet in decibels (centre).

3.1 Time–frequency representation by wavelets

Each signal is first linearly normalised to the interval $[0, 1]$ by a min–max transformation applied trace by trace. This per-sample normalisation makes the representation insensitive to the absolute gain of the recording, which depends strongly on the source distance and is therefore not a reliable class descriptor.

The continuous wavelet transform is then applied with a complex Morlet wavelet of parameters `cmor1.5-1.0` (bandwidth 1.5; centre frequency 1.0 Hz). The analysis scales are geometrically distributed so as to cover the 0.5–20 Hz band uniformly on a logarithmic axis with 32 scales. With the scale–frequency relation

$$a_k = \frac{f_c\, f_s}{f_k}, \qquad f_k = f_{\min}\left(\frac{f_{\max}}{f_{\min}}\right)^{k/(K-1)}, \qquad k = 0, \ldots, K-1, \tag{1}$$

where $f_c = 1.0$ Hz, $f_s = 100$ Hz, $f_{\min} = 0.5$ Hz, $f_{\max} = 20$ Hz and $K = 32$, the scales span from $a = 200$ (0.5 Hz) to $a = 5$ (20 Hz), with a geometric step of 12.6 % from one scale to the next, i.e. about 5.8 scales per octave over the 5.3 octaves covered.

Only the *modulus* of the complex coefficients is retained. It measures the local energy in time and frequency and is free of the phase, which is poorly robust to noise in this representation. The modulus is then converted to decibels:

$$S(a, t) = 20 \log_{10}(|W(a, t)| + \varepsilon), \qquad \varepsilon = 10^{-12}. \tag{2}$$

This logarithmic compression is an important methodological choice. Seismic scalograms span several orders of magnitude between the energy peak of the arrival and the coda or background noise; on a linear scale, a dominant fraction of the information — precisely the part that characterises the slow decay of LP events or the energy spreading of tremor — is crushed near zero and becomes invisible to the network. The decibel scale makes these low-amplitude structures comparable, in dynamic range, to the high-amplitude ones.

The time dimension is finally reduced from 6,000 to 64 columns by averaging over contiguous blocks of $\lfloor 6000/64 \rfloor = 93$ samples, which uses 5,952 of the 6,000 samples (the last 48, i.e. 0.48 s, are discarded). Each scalogram column therefore represents 0.93 s of signal. This averaging smooths high-frequency noise while preserving the energy envelope of the event.

For each signal this yields a matrix of size 64×32 (time $\times$ scale), to which a channel dimension is added to form the input tensor $(64, 32, 1)$ of the 2D convolutional branch. Compared with the fixed-window spectrogram of earlier approaches [2], this representation offers an adaptive resolution: fine in time at high frequencies (useful for the impulsive onset of VT and TC events) and fine in frequency at low frequencies (useful for separating LP from TR).

3.2 Raw waveform: the second modality

The second input of the network is the trace itself, normalised by the *same* min–max transformation as the one applied upstream of the CWT. This consistency is deliberate: both branches then describe exactly the same signal, on the same amplitude scale, and the network does not have to reconcile two different normalisation conventions. The input tensor of the 1D branch has shape $(6000, 1)$.

No prior reduction is applied: the waveform is provided at its native 100 Hz resolution. This is precisely the point of this modality. After averaging over 0.93 s, the scalogram can no longer account for what happens on the scale of a few hundredths of a second — yet it is exactly at that scale that the distinction between an impulsive onset (VT, TC) and an emergent onset (LP, TR) is played out, and that the fine structure of the first arrival can be read. The 1D branch preserves this information and makes it available to the classifier.

3.3 2D convolutional branch on the scalogram

The CWT branch processes the $(64, 32, 1)$ tensor through two successive convolutional blocks of 32 and then 64 filters of size 3×3 (`same` padding, ReLU activation), each followed by batch normalisation, max-pooling and a dropout of 15 %.

The key point is the *shape of the pooling kernel:* $(1, 2)$. Downsampling acts on the scale axis only, never on the time axis. The frequency dimension therefore goes from 32 to 16 and then to 8, while all 64 time steps are preserved from end to end of the branch. This choice follows directly from the fusion objective: to concatenate the output of this branch with that of the 1D branch, both sequences must share the same clock. A conventional $(2, 2)$ pooling, as used in image classification CNNs, would have reduced the sequence to 16 time steps and destroyed the intended alignment.

At the output of the second block, the tensor of shape $(64, 8, 64)$ is flattened along its last two axes by a `Reshape` operation, producing a sequence of *64 time steps each described by* $8 \times 64 = 512$ *features*. Every time step therefore carries, in vector form, the joint description of all filtered frequency bands for the instant considered.

A dense layer of 256 units (ReLU, followed by batch normalisation) finally projects this sequence onto the common fusion space. This layer plays two roles: it halves the dimensionality, and above all it brings the output of the CWT branch to the same dimension as that of the waveform branch (Section 3.5). Layer-by-layer details are given in Table 2.

Table 2: 2D convolutional branch on the CWT scalogram. BN: batch normalisation.

Layer	Output shape	Parameters
Scalogram input	$(64, 32, 1)$	0
Conv2D 32 (3×3, `same`) + ReLU	$(64, 32, 32)$	320
BN	$(64, 32, 32)$	128
MaxPooling (1×2)	$(64, 16, 32)$	0
Dropout 15 %	$(64, 16, 32)$	0
Conv2D 64 (3×3, `same`) + ReLU	$(64, 16, 64)$	18,496
BN	$(64, 16, 64)$	256
MaxPooling (1×2)	$(64, 8, 64)$	0
Dropout 15 %	$(64, 8, 64)$	0
Reshape (time, scale $\times$ channels)	$(64, 512)$	0
Dense 256 + ReLU	$(64, 256)$	131,328
BN	$(64, 256)$	1,024
Total		**151,552**

3.4 1D convolutional branch on the raw waveform

The waveform branch processes the $(6000, 1)$ tensor through four *stride-2* 1D convolution layers of 16, 32, 64 and 64 filters, with kernel sizes 7, 5, 5 and 3 respectively, each followed by a ReLU activation and batch normalisation. The sequence length is halved at every layer: $6000 \to 3000 \to 1500 \to 750 \to 375$.

This sizing follows a logic of growing receptive field. The receptive field of one output step is successively 7, 15, 31 and then *47 samples*, i.e. 0.47 s at 100 Hz, with a jump of 16 samples (0.16 s). Each vector of the output sequence therefore summarises half a second of waveform, sampled every 0.16 s: a resolution sufficient to characterise the steepness of an onset and the structure of a wave train, without retaining the sample-by-sample detail, which would be neither informative nor robust.

Going from 375 to 64 time steps cannot be done by simple block averaging, since 375 is not an integer multiple of 64. An *exact bilinear resampling along the time axis* is therefore used, implemented with `tf.image.resize` applied to the sequence viewed as an image of height 1. The resulting sequence has 64 time steps of 64 channels and superimposes exactly, instant by instant, onto the sequence produced by the CWT branch: each scalogram column spans 0.93 s, and each step of the 1D branch after resampling spans 0.9375 s.

As for the CWT branch, a dense layer of 256 units (ReLU + BN) projects the sequence onto the fusion space. Details are given in Table 3.

Table 3: 1D convolutional branch on the raw waveform.

Layer	Output shape	Parameters
Raw waveform input	$(6000, 1)$	0
Conv1D 16 (k=7, s=2) + ReLU	$(3000, 16)$	128
BN	$(3000, 16)$	64
Conv1D 32 (k=5, s=2) + ReLU	$(1500, 32)$	2,592
BN	$(1500, 32)$	128
Conv1D 64 (k=5, s=2) + ReLU	$(750, 64)$	10,304
BN	$(750, 64)$	256
Conv1D 64 (k=3, s=2) + ReLU	$(375, 64)$	12,352
BN	$(375, 64)$	256
Bilinear temporal resampling	$(64, 64)$	0
Dense 256 + ReLU	$(64, 256)$	16,640
BN	$(64, 256)$	1,024
Total		**43,744**

3.5 Branch balancing and concatenation

The two branches do not naturally have the same output dimensionality: 512 features per time step for the CWT branch against 64 for the waveform branch. A direct concatenation would produce a 576-dimensional vector of which 89 % would come from a single modality. In such a regime, the downstream layers — whose weights are initialised isotropically — allocate their capacity in proportion to the number of available channels, and the minority modality risks being treated as additional noise rather than as a distinct source of information.

The dense projections described above correct this imbalance by bringing both branches to a *common dimension of 256*. Concatenation along the feature axis then produces a sequence of *64 time steps* $\times$ *512 features*, of which exactly half comes from each modality. The batch normalisation layers placed after each projection further guarantee that both halves have comparable activation statistics at the entrance of the fusion (per-channel normalised mean and variance), thereby complementing the dimensional balancing with a scale balancing.

3.6 MLP fusion block

Concatenation juxtaposes the two modalities but does not combine them: at this stage, no unit of the network computes a function involving simultaneously a CWT feature and a temporal feature. This is the role of the MLP fusion block.

The block consists of two successive dense layers of 512 and then 256 units, each with a ReLU activation, batch normalisation and a dropout of 20 %. The essential point is that a dense layer applied to a rank-3 tensor acts only on its *last axis*: the same weights are therefore applied independently to each of the 64 time steps, with no mixing between instants. The block is thus strictly equivalent to a shared `TimeDistributed` MLP, or equivalently to a stack of 1D convolutions with kernel size 1.

This strict separation of roles is the core of the contribution:

- the *MLP fusion block* is responsible for *cross-modal* mixing at a fixed instant: it learns, for each time step, which non-linear combinations of time–scale energy and local waveform shape are discriminant;
- the *BiLSTM head* is responsible for *temporal* mixing: it learns how these fused descriptors evolve over the course of the event.

Without this block, the BiLSTM would have to fulfil both functions simultaneously through its recurrent gates alone, which is possible but less direct: the input transformations of an LSTM cell are linear, and non-linear mixing between modalities could only emerge indirectly, by composition with the gate non-linearities. The fusion MLP makes this operation explicit, endows it with its own capacity (397,056 parameters, i.e. 49 % of the model) and with its own regularisation.

The output of the block is a sequence of *64 time steps* × *256 fused features*.

3.7 BiLSTM head with self-attention

The fused sequence passes through two bidirectional LSTM layers of 64 and then 32 units, returning full sequences, each followed by batch normalisation and a dropout of 20 %. Bidirectionality makes it possible to integrate the past and future context of each instant, whereas a unidirectional cell only has access to the past — a useful property for an offline-segmented signal, where the end of the coda retrospectively informs on the nature of the arrival.

A *scaled dot-product self-attention* mechanism then weights the time steps against one another. Query, key and value are here the sequence itself ($Q = K = V = X$), with no learned projection:

$$\mathrm{Att}(X) = \mathrm{softmax}\left(\frac{XX^{\top}}{\sqrt{d}}\right)X, \qquad d = 64. \tag{3}$$

The module is therefore *entirely parameter-free*: it adds no weights and cannot overfit by itself. Its effect is to recompute each instant as a weighted average of all instants that resemble it, which reinforces recurring patterns (the quasi-stationary oscillation of a tremor, whose instants all strongly resemble one another) and highlights singular instants (the impulsive onset of a VT event, which resembles no other).

The weighted sequence is reduced by *global average pooling* along the time axis, producing a single 64-dimensional vector. This choice replaces the sequence flattening used in earlier recurrent architectures: flattening would have produced a $64 \times 64 = 4096$-dimensional vector and made the following dense layer alone responsible for several million parameters. Average pooling further provides invariance to temporal translation, which is desirable since the position of the event within the 60-second window depends on the segmentation and not on the class.

A dense layer of 128 units (ReLU, batch-normalised, regularised by a dropout of 50 %) finally precedes the four-class softmax layer. Details of the fusion block and of the head are given in Table 4.

Table 4: MLP fusion block and recurrent head.

Layer	Output shape	Parameters
Concatenation (CWT ‖ waveform)	$(64, 512)$	0
Dense 512 + ReLU	$(64, 512)$	262,656
BN	$(64, 512)$	2,048
Dropout 20 %	$(64, 512)$	0
Dense 256 + ReLU	$(64, 256)$	131,328
BN	$(64, 256)$	1,024
Dropout 20 %	$(64, 256)$	0
BiLSTM 64 (sequences)	$(64, 128)$	164,352
BN	$(64, 128)$	512
Dropout 20 %	$(64, 128)$	0
BiLSTM 32 (sequences)	$(64, 64)$	41,216
BN	$(64, 64)$	256
Dropout 20 %	$(64, 64)$	0
Scaled dot-product self-attention	$(64, 64)$	0
Global average pooling 1D	(64)	0
Dense 128 + ReLU	(128)	8,320
BN	(128)	512
Dropout 50 %	(128)	0
Dense 4 (softmax)	(4)	516
Total		**612,740**

3.8 Parameter budget

The complete model has **808,036 parameters**, of which 804,292 are trainable; the remaining 3,744 are the running means and variances of the thirteen batch-normalisation layers. Table 5 gives the breakdown.

Table 5: Parameter breakdown by module.

Module	Parameters	Share
CWT branch (Conv2D + projection)	151,552	18.8 %
Waveform branch (Conv1D + projection)	43,744	5.4 %
MLP fusion block	397,056	49.1 %
BiLSTM head + attention + classifier	215,684	26.7 %
Total	**808,036**	**100 %**

Two observations are worth making. First, the MLP fusion block alone accounts for nearly half of the network's parameters: this is the price of explicit fusion, and it confirms that this module is not a mere adapter but the main learned computational organ. Second, the waveform branch represents only 5.4 % of the model. Adding a second modality therefore comes at a marginal parameter cost, the bulk of the budget being devoted not to *reading* the waveform but to *combining* it with the time–frequency representation.

Table 6: Trainable parameters per compared technique.

Rank	Technique	Parameters	× vs. proposed
1	Bimodal hybrid, MLP fusion (proposed)	808,036	1.0×
2	CWT + BiLSTM [3]	1,569,872	1.9×
3	CWT + 2D CNN [2, 3]	1,142,896	1.4×
4	Envelopes + MLP [4]	356	0.0004×

For comparison, the reference BiLSTM model of [3], reimplemented in this study, contains more than 1.57 million parameters for a single input modality. The proposed hybrid model is therefore significantly more compact than the reference BiLSTM, with approximately half as many parameters, while simultaneously exploiting both the Continuous Wavelet Transform (CWT) representation and the raw signal. In contrast, the MLP+GA model contains only a few hundred trainable parameters, as it operates on a compact set of manually engineered envelope features rather than learning directly from the input signals. The parameter comparison is summarized in Table 5.

3.9 Regularisation and training

To limit overfitting on an imbalanced dataset, light augmentation is applied to the training data only, one augmented copy being added within each fold (factor 1, i.e. a doubling of the training set):

- *scalogram*: additive Gaussian noise ($\sigma = 0.01$) and random circular time shift of ± 5 columns, i.e. ± 4.7 s;
- *raw waveform*: additive Gaussian noise ($\sigma = 0.01$) and circular time shift of ± 300 samples, i.e. ± 3 s.

The shift amplitudes of the two modalities were chosen proportional to their respective lengths ($5/64 \approx 300/6000$), so that the temporal perturbation is of comparable physical magnitude in both views.

The loss function is the categorical cross-entropy with *label smoothing* (factor 0.1), which discourages overconfident predictions and improves calibration — a valuable property for a decision-support tool, where the output probability is meant to be read by an analyst and not merely maximised.

The Adam optimiser is used with an initial learning rate of 10^{-3}, a batch size of 64 and a maximum of 40 epochs. Two callbacks regulate training: learning rate reduction on plateau (factor 0.5, patience 5 epochs, floor 10^{-6}) and early stopping with restoration of the best weights (patience 10 epochs), with 10 % of each fold's training data used as internal validation.

4 Experimental protocol

Performance is estimated by *stratified five-fold cross-validation* (random seed fixed at 42), which preserves the class proportions in each fold and provides, in addition to the mean, the standard deviation of the indicators.

The per-fold split is organised as follows. Out of the 3,592 signals, each fold sets aside approximately 719 test signals, never seen and never augmented, and approximately 2,873 training signals. The latter are doubled by augmentation ($\approx$ 5,747 examples), shuffled, and then split into 90 % effective training ($\approx$ 5,172) and 10 % internal validation ($\approx$ 575) used exclusively to drive the learning-rate schedule and early stopping. The test fold intervenes only once, at the end of training, for prediction.

Four metrics are reported:

- *accuracy*, the overall proportion of correct predictions;
- the *macro F1-score*, the unweighted mean of the per-class F1 scores, particularly informative here since it gives the same weight to the 304 VT events as to the 1,488 TC events;
- *Cohen's kappa*, which corrects for chance agreement and remains interpretable under class imbalance;
- the *confusion matrix*, accumulated over the five folds so as to aggregate predictions over the whole dataset.

The implementation relies on TensorFlow/Keras for the network and on PyWavelets for the continuous wavelet transform.

5 Results

5.1 Per-fold results

Table 7 details, for each of the five folds, the accuracy and macro F1-score of the proposed model. Performance is stable from one fold to another, the total range not exceeding 1 points of accuracy and 1.1 points of F1.

Table 7: Accuracy and macro F1-score (%) per fold for the bimodal hybrid model with MLP fusion.

Fold	Accuracy	Macro F1
1	97.64	96.79
2	98.19	97.18
3	97.35	96.19
4	97.35	96.44
5	98.33	97.24
Mean	$\mathbf{97.77 \pm 0.41}$	$\mathbf{96.77 \pm 0.41}$

Fold 5 is consistently the most favourable on both metrics at once, while fold 3 and fold 4 tie for the lowest accuracy; fold 3 alone has the lowest macro F1. The variation is therefore not attributable to a local class imbalance, but to the intrinsic difficulty of the signals drawn in those folds. Fig. 3 visualises this evolution.

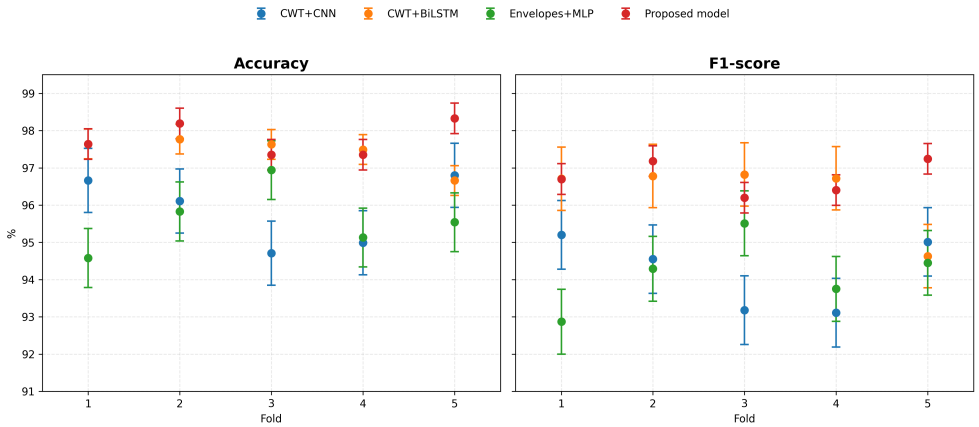


Figure 3: Accuracy (left) and F1-score (right) per fold. The proposed model is the red curve; the other curves are the re-implemented baselines (Section 5.5).

5.2 Aggregated results

Over the five folds, the model reaches **97.77 % ± 0.41 accuracy**, **96.77 % ± 0.41 macro F1-score** and a **Cohen's kappa of 0.967** (Table 8). The inter-fold standard deviation below half a point confirms the robustness of the system to the partition draw.

Table 8: Mean performance (± standard deviation) over five-fold cross-validation.

Metric	Value
Accuracy	97.77 % ± 0.41
Macro F1-score	96.77 % ± 0.41
Cohen's kappa	0.967
Signals evaluated	3,592
Correct predictions (cumulated)	3,512

The kappa of 0.967 deserves a comment. With four classes of proportions 36.5 / 13.6 / 8.5 / 41.4 %, chance agreement amounts to 33.0 %; kappa therefore measures the fraction of the remaining margin (67.0 %) actually recovered by the classifier. A kappa of 0.967 means that 96.4 % of that margin is captured, well beyond the conventional 0.81 threshold that characterises "almost perfect" agreement.

5.3 Confusion matrix

Fig. 4 shows the normalised confusion matrix of the model, accumulated over the five folds; Table 9 gives the raw counts.

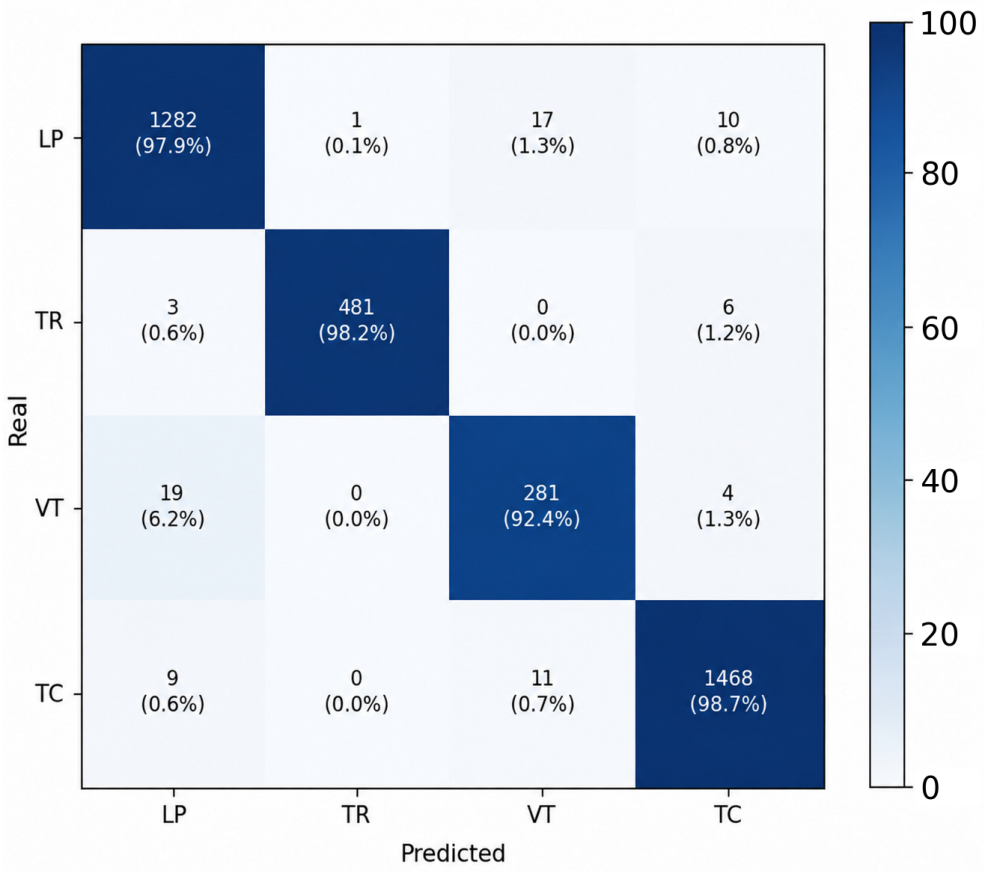


Figure 4: Normalised confusion matrix of the bimodal hybrid model with MLP fusion (cumulated over the five folds).

Table 9: Cumulated confusion matrix (counts; rows: true class, columns: predicted class; diagonal = correct predictions).

True \ Predicted	LP	TR	VT	TC	Total
LP	**1,282**	1	17	10	1,310
TR	3	**481**	0	6	490
VT	19	0	**281**	4	304
TC	9	0	11	**1,468**	1,488
Predicted total	1,313	482	309	1,488	3,592

The error structure is remarkably clean and consistent with source physics.

The TR class is very well separated. A single signal (one LP event) is ever predicted as tremor: the TR column contains the 482 true tremors plus this one error, giving a precision of 99.79 % and a specificity of 99.97 %. This result is explained by the very nature of tremor, the only quasi-stationary event in the dataset: its energy spreads over the entire window within a narrow band, a signature that almost no transient arrival can imitate. The nine missed tremors are scattered towards LP (3) and TC (6), presumably short segments or segments with a low signal-to-noise ratio.

The remaining errors form an LP–VT–TC triangle. These three families exchange 70 of the 80 total errors. The most loaded sub-graph is LP $\leftrightarrow$ VT ($17 + 19 = 36$ errors): this is the

expected overlap between a fluid-source event with low-frequency content and a fracturing event whose coda, once attenuated, takes on a comparable appearance. Next comes LP ↔ TC (10 + 9 = 19 errors), consistent with the fact that distant tectonic events, filtered by propagation, lose their high frequencies and move spectrally closer to LP events. Finally, VT ↔ TC (4 + 11 = 15 errors) reflects the physical proximity of two fracturing mechanisms that essentially differ only in their location.

The asymmetry of the confusions is informative. VT is predicted as LP in 19 cases while LP is predicted as VT in only 17 cases, even though the class sizes differ by a factor of 4.3. Relative to class size, 6.25 % of VT events are classified as LP against only 1.30 % of LP events classified as VT: in case of doubt, the model leans towards the majority class. This bias is the expected behaviour of a classifier trained without explicit class reweighting.

5.4 Per-class performance

Table 10 details the per-class indicators computed on the cumulated matrix.

Table 10: Per-class performance of the proposed model (cumulated over the five folds).

Class	Recall	Precision	F1	Specificity	Error rate
LP	97.86 %	97.64 %	97.75 %	98.64 %	1.64 %
TR	98.16 %	99.79 %	98.97 %	99.97 %	0.28 %
VT	92.43 %	90.94 %	91.68 %	99.15 %	1.42 %
TC	98.66 %	98.66 %	98.66 %	99.05 %	1.11 %
Macro	**96.78 %**	**96.76 %**	**96.77 %**	**99.20 %**	—

Three classes out of four exceed 97 % F1, and the TR class reaches 98.97 %. The *VT class remains the residual difficulty*, with an F1 of 91.69 % — nearly six points below the three others.

This deficit has a twofold explanation. First, rarity: with 304 exemplars (8.5 % of the dataset), the VT class has about 243 training signals per fold before augmentation, an order of magnitude fewer than TC. Second, physical ambiguity: unlike TR, whose quasi-stationary signature is unique, a VT event shares its source mechanism (brittle fracturing) with TC events and its low-frequency coda with LP events. The VT class therefore occupies precisely the intersection of the two other regions of the time–frequency plane, which places it in a structurally unfavourable position.

It should be noted that recall (92.43 %) and precision (90.94 %) are at similar levels here : the model does not sacrifice one for the other, it simply lacks separability on this class. Rebalancing by loss weighting would shift the trade-off without necessarily improving the F1.

5.5 Comparison with baseline methods

The three baseline methods were re-implemented and evaluated on the same 3,592 signals, with the same stratified five-fold cross-validation protocol and the same random seed, ensuring a fair comparison. Table 11 collects the results.

Table 11: Comparison with the re-implemented baseline methods (same dataset, same five-fold cross-validation).

Method	Representation	Modalities	Accuracy	Macro F1	κ
CWT + 2D CNN [2, 3]	RGB scalogram 32×32	1	95.85 %	94.23 %	0.938
CWT + BiLSTM [3]	2D sequence 64×32	1	97.44 %	96.33 %	0.962
Envelopes + MLP [4]	Envelope descriptors	1	95.60 %	94.17 %	0.934
Bimodal hybrid, MLP fusion	**CWT** 64×32 **+ waveform** 6000×1	**2**	**97.77 %**	**96.77 %**	**0.967**

The proposed model outperforms the three baselines on both metrics. The margin is clear against the envelope-descriptor classifier, which remains limited by the descriptive power of a hand-crafted representation. It is more modest against the scalogram CNN.

This dissymmetry between the two metrics is the most instructive result in the table. The macro F1 improves substantially more than the accuracy relative to the best baseline. Now, accuracy is dominated by the majority classes (LP and TC together account for 78 % of the dataset) whereas macro F1 gives equal weight to each class. A gain concentrated on macro F1 therefore mechanically means that the improvement bears on the minority classes — VT and TR — that is, precisely where a single representation reaches its limits and where a second view of the signal is most decisive. The contribution of the raw waveform is not to be read in the total number of corrected predictions, but in their distribution.

6 Discussion

Modality complementarity is the source of the gain. The scalogram and the waveform are not two redundant encodings of the same information. The former is a modulus — it discards the phase — averaged over 0.93 s windows; it describes *where* energy sits in the time–frequency plane but remains blind to the fine structure of the signal within each column. The latter fully retains that fine structure, including polarity and onset steepness, but exposes no frequency organisation. The fact that the gain is concentrated on macro F1 rather than on accuracy is consistent with this reading: the majority classes were already well separated by the time–frequency representation alone, whereas the rare and ambiguous classes benefit from the additional information.

Temporal alignment is a requirement, not an implementation detail. Concatenating two sequences only makes sense if their indices designate the same physical instant. This justifies two architectural choices that would otherwise look arbitrary in isolation: the asymmetric $(1, 2)$ pooling of the CWT branch, which forbids any reduction of the time axis, and the bilinear resampling of the 1D branch, preferred over block averaging, which would have been impossible since 375 is not divisible by 64. Without these two precautions, the concatenation would mix offset instants and the fusion MLP would learn spurious correlations.

Dimensional balancing conditions the effective exploitation of the second modality. The natural 512-to-64 feature ratio between the two branches would have relegated the waveform to the rank of a marginal perturbation. The dense projections to 256 dimensions, complemented by batch normalisations that also equalise activation scales, ensure that fusion operates between two modalities of equivalent status. This is all the more notable as the waveform branch accounts for only 5.4 % of the parameters: its influence stems from its position in the architecture, not from its capacity.

Separating cross-modal fusion from temporal integration is effective. By entrusting the non-linear mixing of modalities to a time-distributed MLP, and temporal integration to the

BiLSTM head, one decomposes a learning problem into two sub-problems, each with its own capacity and its own regularisation. The fusion MLP mobilises 49 % of the network's parameters, confirming that it is not a mere dimensional adapter but the main learned computational organ.

Per-class analysis locates the residual difficulty precisely. The VT class, the rarest (8.5 % of the dataset) and structurally intermediate between LP and TC, concentrates the errors relative to its size and shows the only F1 below 97 %. At the opposite end, the TR class attains 99.79 % precision and 99.97 % specificity: almost no event from the other families is ever confused with a tremor, confirming that its quasi-stationary signature has virtually no equivalent in the dataset.

One limitation concerns compactness rather than cost. With 808,036 parameters, the network is four times lighter than the reference BiLSTM with a flattened head, despite two input modalities. The dominant cost lies upstream, in computing the CWT over 32 scales for each 6,000-sample trace, an operation that will have to be optimised for continuous-stream deployment.

Three methodological limitations must be reported. First, zero-padding of events shorter than 60 s creates, after the decibel conversion, an extremely low energy floor in the silent portions — a floor clearly visible in the LP and VT scalograms of Fig. 2. The duration of this floor correlates with event duration, which itself is partly related to class. It therefore cannot be excluded that the network partly exploits a segmentation artefact rather than a physical signature. A check by explicit masking of the padded regions, or by clipping the scalogram at a fixed floor (for instance -80 dB) followed by renormalisation, would quantify this contribution.

Second, the augmentation noise magnitude ($\sigma = 0.01$) was calibrated for data in $[0, 1]$ and is negligible relative to the decibel dynamic range of the scalogram; the effective augmentation of the time–frequency view therefore reduces to the circular time shift. Scaling the noise to the dynamic range of the representation is a straightforward refinement.

Third, data augmentation is applied before the split between effective training and internal validation within each fold: the augmented copy of a signal may therefore end up on the validation side while the original is on the training side. It should be emphasised that *the test fold is never augmented and never seen during training*, so that the metrics reported in Section 5 are not contaminated; only the early-stopping criterion relies on a slightly optimistic validation set, which may lead to stopping training somewhat late. Correcting this point is a worthwhile improvement whose expected effect on the results remains small.

Finally, the evaluation covers a single station of a single volcano; transferability to other monitoring networks, where site effects and recording conditions differ, remains to be established.

Directions for improvement. Three directions appear promising: (i) explicit class rebalancing, by loss weighting or by targeted augmentation of VT events, whose F1 deficit concentrates most of the remaining margin; (ii) the addition of a third modality, for instance the envelope descriptors of [4], which could be injected into the same fusion MLP without structural modification of the architecture; (iii) replacing the parameter-free self-attention with attention using learned projections, whose overhead would remain modest relative to the current 808,036 parameters.

7 Conclusion

We have presented a seismo-volcanic event classifier based on a *bimodal* hybrid network that jointly processes, within a single end-to-end trained architecture, a continuous wavelet scalogram and the raw temporal waveform. Three architectural choices structure the method: an exact temporal alignment of the two branches on 64 time steps, obtained by asymmetric pooling on the scalogram side and bilinear resampling on the waveform side; a dimensional balancing of the two modalities by projection onto a common 256-dimensional space; and a time-distributed

MLP fusion block that takes charge of cross-modal mixing and leaves the BiLSTM head with self-attention solely responsible for temporal integration.

On the 3,592 signals of Llaima volcano and under stratified five-fold cross-validation, the model reaches **97.77 % accuracy, 96.77 % macro F1-score and a Cohen's kappa of 0.967**, with only 808,036 parameters. It outperforms the three baseline methods re-implemented under identical conditions, and its advantage in macro F1 is markedly larger than its advantage in accuracy — a sign that the benefit of bimodality is concentrated on the minority and spectrally ambiguous classes.

These results support the value of multimodal architectures with learned fusion for automated volcano monitoring, and pave the way for the integration of additional modalities as well as for multi-station validation.

References

[1] J. P. Canário, R. F. de Mello, M. Curilem, F. Huenupan, R. A. Rios, "Llaima volcano dataset: In-depth comparison of deep artificial neural network architectures on seismic events classification," *Data in Brief*, vol. 30, 105627, 2020.

[2] M. Curilem, J. P. Canário, L. Franco, R. A. Rios, "Using CNN to classify spectrograms of seismic events from Llaima volcano (Chile)," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.

[3] J. P. Canário, R. Mello, M. Curilem, F. Huenupán, R. Rios, "In-depth comparison of deep artificial neural network architectures on seismic events classification," *Journal of Volcanology and Geothermal Research*, vol. 401, 106881, 2020.

[4] E. H. Ait Laasri, A. Atmani, E.-S. Akhouayri, D. Agliz, "A simple and effective MLP-based seismic signal classifier using temporal and spectral envelope features with genetic algorithm-optimization," *Measurement*, vol. 247, 116776, 2025.