Chapitre 2

La détection de nouveautés

2.1 Définition

La détection de nouveautés [90] (ou novelty detection) constitue une branche du machine learning se situant dans un cadre non supervisé. Elle est appliquée lorsque la base de données contient un label fortement majoritaire (souvent considéré comme normal) et un label minoritaire ou absent (pouvant être atypique). Ce type d'approche est défini comme la reconnaissance des données différant du comportement normal issu des donnés d'apprentissage [90]. La détection de nouveautés possède l'avantage de ne pas dépendre de la connaissance apriori ou de la présence de données atypiques.

La détection de nouveautés est pertinente lorsque l'apprentissage d'un modèle lié au label minoritaire n'est pas envisageable et s'apparente au problème de type one-class. Le principe consiste à construire un modèle de normalité à partir des données normales majoritairement présentes. Il s'agit de l'étape de caractérisation de la normalité. Des données de validation, non utilisées pour définir le modèle, sont comparées au modèle de normalité afin de définir un score et un seuil de nouveauté. Le score de nouveauté est appliqué à chaque donnée test. Plus ce score est élevé, plus la donnée peut être considérée comme n'étant pas issue de la même distribution que les données normales. Ce score est comparé à un seuil afin de classifier la donnée comme nouvelle ou non. Une description du processus est donnée en figure 2.1. Nos algorithmes correspondent à ce type d'approche. Les modèles mis en place doivent permettre la généralisation des caractéristiques normales des données tout en évitant un surapprentissage de celles-ci.

La détection de nouveautés s'apparente à la détection d'anomalies [26] et la détection d'outliers [25, 60]. La détection de nouveautés consiste à reconnaitre ce qui n'est pas observé dans la base de données normales. Les méthodes de détection d'outliers consistent à trouver les données ayant un comportement différent de celui attendu [25]. La détection d'anomalies cherche à mettre en évidence dans les données un fonctionnement anormal du système résultant ou pouvant entrainer des endommagements. Sa définition dans [26] est la même que celle donnée pour la

détection d'outliers dans [25]. La base de données normales est considérée comme caractérisant entièrement le comportement nominal du sytème et toute donnée s'en écartant provient d'une irrégularité ou d'un nouveau comportement du système. Ces approches ont pour but de détecter des patterns rares dans les données sans a priori sur ces dernières étant donné leur absence ou faible nombre dans la base de données.

2.2 État de l'art de la détection de nouveautés

Il existe différents états de l'art assez complets sur ces différentes approches [90, 26, 25, 60, 83, 84]. Nous donnons une brève description et intuition des différentes approches utilisées pour la détection de nouveautés. Nous renvoyons à ces articles et aux différentes références associées pour plus de précisions sur les différentes méthodes.

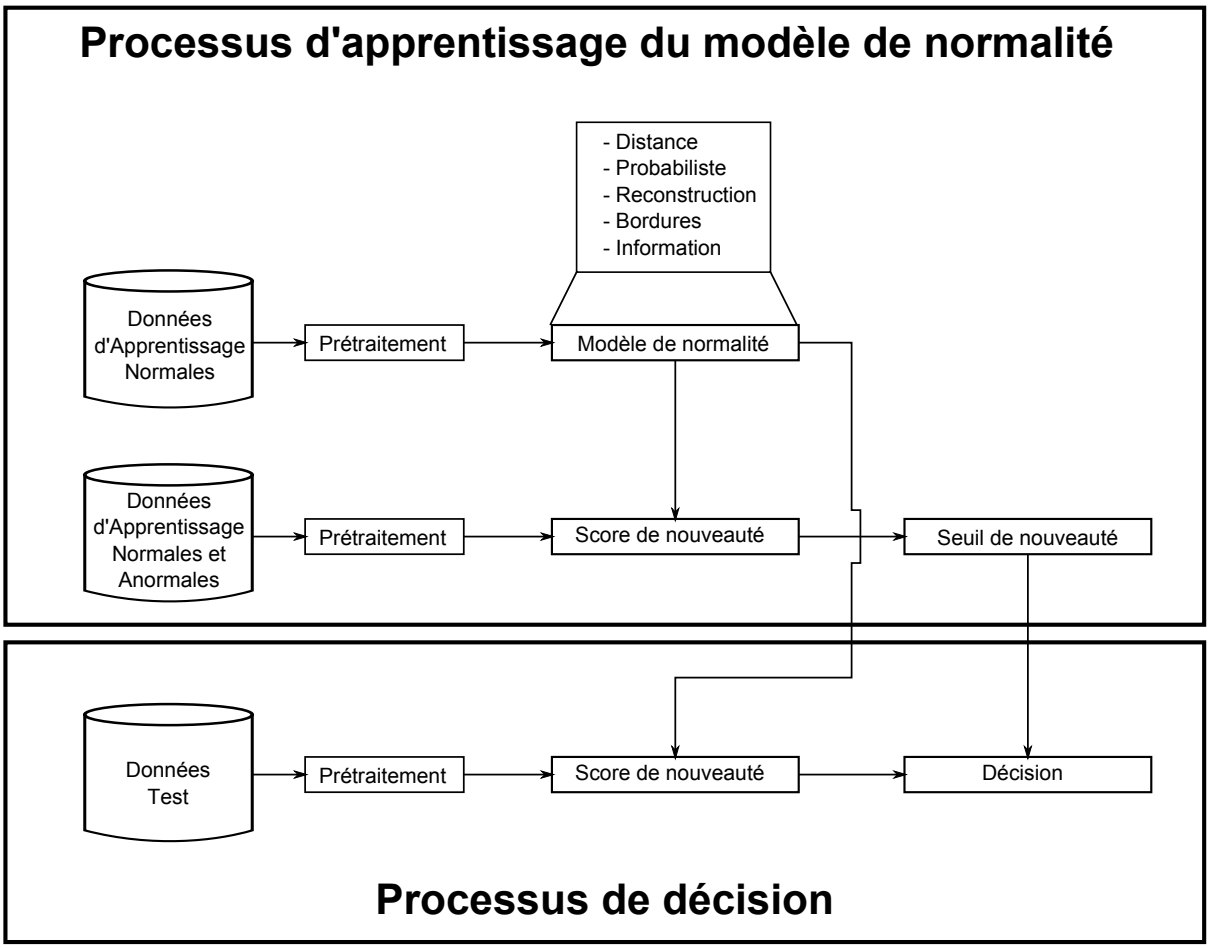

FIGURE 2.1 – Processus de détection de nouveautés/anomalies. La partie supérieure concerne l'apprentissage du modèle de normalité et la définition du seuil de détection à partir d'ensembles disjoints de données normales uniquement pour l'apprentissage du modèle, et de données normales et atypiques pour la définition du seuil de nouveauté. La partie inférieure porte sur la détection des données inconnues du système.

Différentes méthodes de détection de nouveautéss/anomalies/outliers permettent d'étudier et de classifier des ensembles de données dont le label atypique est très faiblement représenté. Ces méthodes passent toutes par la caractérisation des données normales, le calcul d'une mesure de nouveauté et la comparaison à un seuil. Ces techniques peuvent être répertoriés en 5 classes distinctes [90] :

— les approches probabilistes,
— les approches basées sur les distances,
— les approches basées sur la reconstruction de la donnée,
— les approches basées sur la caractérisation des limites des données normales,
— les approches basées sur la théorie de l'information.

Le choix d'une méthode par rapport à une autre est effectué selon les caractéristiques des données comme la dimension ou la structure du domaine des données normales.

2.2.1 Les approches probabilistes

Ces approches estiment la distribution sous-jacente des données normales, définissant ainsi un modèle de normalité. La distribution des données nominales n'est généralement pas connue. Elle peut alors être estimée à partir de mélanges de gaussiennes [15]. Il s'agit là d'une méthode paramétrique dont les paramètres principaux sont le nombre de gaussiennes du mélange et leurs caractéristiques (moyenne et covariance). Chaque gaussienne correspond à un comportement normal. Les paramètres du modèle sont estimés à partir du maximum de vraisemblance via l'algorithme Expectation-Maximisation (EM). Une seconde approche pour estimer la distribution sous-jacente est l'estimation de la densité par noyau [49]. Cette méthode non-paramétrique consiste en l'application d'un noyau probabiliste au niveau de chaque donnée d'apprentissage normale. Une explication plus détaillée de cette approche est donnée dans la section 5.3. Le principal paramètre de cette approche est l'échelle du noyau considérée. Si elle est trop importante, la distribution est fortement lissée. Si elle est trop faible, la distribution est trop fortement liée aux données d'apprentissage (surapprentissage) et ne généralise pas le modèle de normalité.

La nouveauté est détectée par la comparaison de la donnée test à la distribution du modèle de normalité. Dans un cas unimodal, c'est-à-dire lorsque la distribution possède un unique mode, les nouveautés se trouvent au niveau des queues de distribution et sont donc détectées en fixant un seuil sur cette distribution. La théorie des valeurs extrêmes [37] est une des méthodes permettant de fixer un seuil de détection, elle permet de modéliser la distribution du maximum d'un échantillon de taille fixe. D'après le théorème de Fisher-Tippett [48], dépendant de la loi des données, la distribution des valeurs extrêmes correspond à une loi de Gumbel, Fréchet ou Weibull. La caractérisation de ce maximum permet alors de définir le seuil de détection. Dans un cadre multi-modal, les nouveautés ne sont plus caractérisées uniquement par les extrêmes de la distribution, mais aussi par les points de faibles densités entre les différents modes. Le seuil de détection peut alors être défini par une valeur de densité en dessous de laquelle les points sont considérés comme improbables donc nouveaux [32].

Ces approches permettent de définir la normalité à travers la distribution des données et de détecter la nouveauté à partir d'un seuil sur la densité des données. Ces approches sont dépendantes du nombre de données utilisées pour estimer la distribution. Cependant, en grande dimension l'apprentissage de la densité demande un nombre très important de données pour caractériser la distribution au niveau de toutes les régions de l'espace. Ainsi plus la dimension est grande, plus le nombre de données nécessaires pour caractériser tout l'espace est important. Il s'agit du "fléau de la dimension". Ces approches ne sont donc pas adaptées aux problématiques en grande dimension.

2.2.2 Les approches basées sur les distances

Elles consistent en la mesure de la distance entre une donnée et ses homologues normales. Plus la distance est grande, plus cette donnée peut être considérée comme nouvelle. Il existe deux grandes catégories d'approches utilisant les distances. La première considère les plus proches voisins des données tests comme les k plus proches voisins (k-NN) [49]. Elle permet de classifier les données comme nouvelles si elles se trouvent éloignées des données normales. Cette méthode est paramétrée par le nombre k de voisins à prendre en compte et la distance considérée. Le temps de calcul est généralement long pour de grosses bases de données. Une seconde méthode basée sur les voisins est le Local Outlier Factor (LOF) [17]. Elle compare la densité de voisins autour des données étudiées dans un certain rayon de voisinage (constituant le principal paramètre de l'approche) avec cette même densité calculée pour leurs voisins. Une densité de voisins plus faible signifie une isolation de la donnée et donc son caractère nouveau. La seconde catégorie d'approches basées sur les distances correspond au clustering des données dont la méthode la plus répandue est le k-means [49], k correspondant cette fois-ci au nombre de clusters. La détection de nouveautés s'effectue dans ce cadre par rapport à la distance au plus proche cluster estimé sur les données normales. Il existe plusieurs choix de distances possibles avec des propriétés différentes constituant un paramètre supplémentaire de ces approches.

En grande dimension, la notion de distance est mal établie, ces approches ne sont donc pas efficaces. De plus, la mise en place de ces approches nécessitent un très grand nombre de données pour caractériser correctement l'espace des données normales et pour éviter les fausses détections.

2.2.3 Les approches basées sur la reconstruction des données

Ces approches définissent un modèle caractérisant la normalité au sein des données à partir duquel il est possible de donner une estimation normale de la donnée test. La détection s'effectue alors sur le résidu de la reconstruction. Dans le cas d'une donnée normale, la reconstruction de cette dernière est proche de celle-ci et donc entraine des résidus de reconstruction faibles. Pour une donnée contenant des nouveautés ou des anomalies, les résidus de la reconstruction sont plus importants. Pour ce type d'approche, l'élément déterminant est le nouvel espace de représentation, caractérisant la normalité des données et dans lequel les données sont représen-

tées [13]. Cet espace correspond à des dictionnaires pouvant être appris sur les données ou à des réseaux de neurones [68]. Les méthodes de représentation par dictionnaires décomposent la donnée dans un système linéaire avec des contraintes. Dans la suite nous donnons une explication plus détaillée de la représentation par dictionnaire. L'ACP [49] est la méthode la plus classique de représentation des données dans un dictionnaire permettant de maximiser la variance des données normales. La reconstruction est effectuée à partir des premières composantes principales modélisant la variabilité normale des données. Dans le cadre de la détection de nouveautés, il est également possible de projeter les données sur les dernières composantes principales n'ayant que peu d'information normale et détectant ainsi des outliers. Le principal paramètre de ce type d'approche est la nouvelle dimension dans laquelle les données sont représentées. Dépendant de la nouvelle dimension, ce type d'approche ne nécessite pas nécessairement un très grand nombre de données.

Les réseaux de neurones [68] ont été particulièrement utilisés pour des tâches de classification supervisée mais également de détection de nouveautés. Les Replicator Neural Network (RNN) [53] et les autoencodeurs [109] sont des réseaux de neurones apprenant des sous-espaces permettant la reconstruction de la donnée d'entrée. Les réseaux sont donc appris sur des données normales afin de caractériser cet espace de normalité, la détection de nouveautés s'effectue également sur les résidus de la reconstruction. Les principaux paramètres de ces approches sont le pas du gradient mais également la structure et la composition des couches du réseau de neurones. La construction de ces réseaux nécessitent un très grand nombre de données pour éviter le surapprentissage.

Ces approches fonctionnent sur des données en grande dimension telles que des images. Les espaces dans lesquelles les données sont projetées permettent d'apprendre la structure et le comportement normal des données. Les paramètres de ces différentes approches correspondent aux paramètres des espaces dans lesquels les données sont projetées comme la nouvelle dimension ou la structure du réseau.

2.2.4 Les approches basées sur la caractérisation des limites des données normales

Ces approches caractérisent les bordures de la normalité dans l'espace des données. Dans ce sens, elles ne tiennent pas compte de l'ensemble des données mais uniquement de celles se trouvant proches de ces bordures. La détection de nouveautés correspond alors à la comparaison de la donnée aux limites du domaine normal. Ces méthodes sont apparentées aux support vector machines (SVM) [49]. Un one-class SVM [99] permet de trouver un hyperplan dans un espace transformé des données d'apprentissage normales définissant de vastes marges par rapport à l'origine. Les principaux paramètres de ces approches sont le nombre de points définissant les bordures et un pourcentage de points pouvant être mal classifiés pour éviter le surapprentissage.

Le support vector data description (SVDD) [107] définit un domaine caractérisé par un centre et un rayon autour des données normales minimisant le volume de cette hypersphère. Cette

caractérisation s'effectue également dans un espace transformé. La nouveauté est détectée par l'appartenance ou la non-appartenance de la donnée à l'hypersphère. Les principaux paramètres sont similaires à ceux du one-class SVM.

Il existe d'autre approches caractérisant les bordures de l'espace de normalité non liées au SVM comme les one-class random forests [38]. Ces approches sont également intéressantes sur des données en grande dimension car seuls les points aux bordures sont étudiés. Elles requièrent tout de même un grand nombre de points pour la calibration du modèle.

2.2.5 Les approches basées sur la théorie de l'information

Ces approches proviennent de l'intuition que la présence d'une donnée nouvelle ou atypique au sein d'un ensemble de données normales entraine un changement de la quantité d'information estimée sur toutes les données. La mesure de l'information [15] permet alors de détecter des nouveautés en calculant l'entropie (possible mesure de l'information) pour des sous-ensembles des données. Lorsque l'entropie diminue, cela signifie que la donnée retirée est différente du reste [56].

Ces approches de théorie de l'information ne nécessitent aucun apriori sur les données, seule la mesure de l'information des données est considérée. Cependant il est indispensable de définir la mesure considérée pour l'information. De plus la détection s'effectue uniquement lorsque plusieurs observations sont atypiques dans le jeu de données. Dans le cas contraire, le retrait d'une seule observation peut ne pas être suffisant pour entrainer une modification significative de la mesure d'information. De plus, ces mesures passent souvent par la définition d'une densité de probabilité. Ce type d'approche n'est donc pas adapté en grande dimension.

2.3 La détection de nouveautés appliquée aux données vibratoires

2.3.1 Application aux données temporelles et fréquentielles

Les idées de détection de nouveautés et d'anomalies sont courantes en traitement du signal. Les signaux récupérés sur des systèmes industriels complexes sont en grande majorité des données considérées comme normales. La complexité des systèmes rend impossible de considérer toutes les signatures atypiques possibles. Une pratique standard du traitement du signal consiste à calculer le spectre à partir des données vibratoires temporelles (lorsque ces dernières sont stationnaires) et de comparer le pics fréquentiels obtenus avec ceux issus de comportements normaux. Il s'agit bien d'une approche de détection de nouveautés où les irrégularités sont caractérisées par l'apparition de pics inconnus. La mesure de nouveauté peut alors être obtenue par un calcul résiduel ou par une matrice de correspondance des fréquences d'apparition des différents pics. Les signaux temporels ou fréquentiels sur lesquels les méthodes de détection de nouveautés sont appliquées

sont généralement stationnaires.

Dans [55], une approche de type peak-over-threshold [1](POT) (approche liée aux valeurs extrêmes) est étudiée pour la détection de défauts de roulements à partir de mesures vibratoires sur des périodogrammes. Il s'agit de la transformée de Fourier discrète appliquée à des signaux temporels en régime stationnaire. Les maximums des périodogrammes sans signature atypique sont calculés pour chaque fréquence, ces derniers sont comparés à d'autres périodogrammes de la base d'apprentissage. Les points supérieurs aux maximums calculés sur les périodogrammes sont modélisés par une loi de Pareto sur laquelle le seuil de détection est déterminé.

Une approche de type one-class SVM est utilisée dans [23] pour la détection d'anomalies sur des données vibratoires temporelles. Différents indicateurs du signal temporel sont calculés et projetés dans un espace de dimension réduite sur lequel le one-class SVM est appris.

Une comparaison de différentes approches de détection de nouveautés a été réalisée dans [111] à partir de différents indicateurs classiques des signaux temporels et en ordre. Les différentes approches comparées sont l'ACP, l'estimation de densité par noyau gaussien, les k-means, les k plus proches voisins, les one-class SVM et les autoencodeurs. Les résultats montrent un comportement varié des différentes méthodes suivant le type d'anomalie présent dans les données.

Ces méthodes sont établies dans le domaine temporel ou angulaire du signal en supposant la stationnarité de ce dernier. Notre problématique consiste à étudier des représentations temps-fréquence de ces signaux permettant de tenir compte du caractère non-stationnaire.

2.3.2 Application sur les harmoniques du signal

Les études citées ci-dessus sont effectuées sur des signaux en régime stationnaire sur lesquels la transformée de Fourier peut être appliquée. Cependant dans les signaux non-issus de banc d'essai, la stationnarité est rarement vérifiée. Des méthodes classiques comme la transformée de Fourier ou des indicateurs sur les signaux temporels ne sont plus pertinents. Il est tout de même possible de construire des représentations temps-fréquence afin d'étudier ces signaux à partir de la STFT, des transformées en ondelettes,... Cela permet de prendre en compte l'aspect fréquentiel du signal (donc les éléments périodiques), mais aussi la non-stationnarité à travers l'aspect temporel. Il est également possible d'étudier juste quelques ordres ou harmoniques spécifiques du signal, c'est sur les premières harmoniques qu'apparaissent certains défauts comme les problèmes d'équilibrage.

Dans [31, 34], des méthodes de détection de nouveautés basées sur une estimation de la distribution des valeurs extrêmes, à partir d'un mélange de gaussiennes, ont été étudiées pour la détection d'anomalies de moteurs d'avions dans un cadre multivarié et multimodal. Les valeurs extrêmes dans ce cadre sont définies comme les valeurs les moins probables, c'est-à-dire celles dont les densités de probabilité sont les plus faibles. Les valeurs extrêmes sont calculées sur banc

1. méthode des excès

d'essai [31] pour les premiers ordres des spectrogrammes vibratoires normaux de chaque arbre du moteur, et en vol [34] à partir des points de plus fortes intensités.

Une approche de détection de nouveautés à partir du one-class SVM a été développée dans [54]. Les intensités vibratoires d'ordres prédéfinis des spectrogrammes normaux sont récupérées afin d'apprendre le domaine des données saines.

Notre étude consiste à détecter toute trace des signatures vibratoires inusuelles sur les spectrogrammes. Il est donc indispensable de définir des modèles sur les spectrogrammes et non pas sur quelques ordres de ce dernier.

2.3.3 Application aux spectrogrammes

Un masque normal des spectrogrammes [52] a été présenté dans le chapitre 1 permettant la suppression de l'information normale des spectrogrammes et de comptabiliser les éléments restants. Cette approche s'apparente à des méthodes de détection d'anomalies basées sur une distance correspondant à la concordance de l'information vibratoire présente entre les spectrogrammes normaux et le spectrogramme étudié.

Une approche similaire a été développée dans [33]. Les spectrogrammes ont été subdivisées en sous-zones sur lesquelles une normalisation est effectuée en modélisant le bruit par une loi Gamma afin de l'homogénéiser. Cette normalisation est effectuée intra-zone et entraine des artéfacts sur les bords des sous-zones définies. La distribution des valeurs extrêmes est estimée pour chaque sous-zone. Le nombre de points dépassant le seuil de détection défini à partir de la distribution estimée est comptabilisé pour chaque sous-zone des spectrogrammes normaux. L'approche est réitérée pour chaque spectrogramme et le nombre de points dépassant les seuils de détections sont comparés. Une anomalie est déclarée lorsqu'une des zones du spectrogramme étudié possède significativement plus de points détectés que ses homologues des spectrogrammes normaux.

Dans [63], une référence (définie par la moyenne et la variance) est apprise pour tous les points des spectrogrammes en ordre à partir de données normales. Les spectrogrammes tests sont alors comparés à cette référence pour déterminer les points anormaux. Afin que les points soient considérés comme anormaux, il est nécessaire qu'ils définissent une région continue avec un nombre de points et une surface suffisants. Pour que le spectrogramme soit déclaré comme anormal, la surface totale doit être supérieure à un seuil. Il est également possible de récupérer les raies le long de certains ordres particuliers en collectant les voisinages des maximas locaux au niveau des points détectés [64].

Notre travail est complémentaire à ces différentes approches. Nous cherchons à détecter finement toutes signatures atypiques sur les spectrogrammes à partir de notre base de données construite, c'est-à-dire déterminer l'ensemble des points composant les signatures inusuelles.

2.4 Caractérisation de la base de données de spectrogrammes construite

2.4.1 Répartition des données en sous-ensembles

La figure 2.1 présentait le besoin de plusieurs bases de données disjointes pour mettre en place et tester nos modèles :

— une base d'**apprentissage** comportant uniquement des données normales sans signature inusuelle pour mettre en place le modèle de normalité (Ω_{App}) ;
— une base de **validation** contenant des données avec et sans signatures inusuelles pour calibrer le seuil de détection et les paramètres optimaux du modèle par cross-validation (Ω_{Val}) ;
— une base de **test** pour présenter les résultats et évaluer les performances de nos approches, (Ω_{Test}).

Les données de la base d'apprentissage et des données normales de validation forment un même ensemble dans lequel nous sélectionnons aléatoirement la base d'apprentissage pour définir le modèle et les données normales de la base de validation. Les données atypiques de la base de validation restent identiques, nous pouvons en sélectionner un sous-ensemble pour nos études. La base de test reste la même pour toutes les différentes approches afin de pouvoir les comparer.

Nous avons présenté (section 1.4) la base de données ($\mathcal{B}_0$) construite et indéxée à partir de l'extraction automatique des zones atypiques sur les données textuelles d'annotations manuelles des experts. Nous avons également présenté nos propres annotations de quelques points en différentes classes ($\mathcal{B}_1$) sur quelques spectrogrammes (section 1.5.3) afin d'obtenir une vérité terrain plus fine que sur l'échelle d'un patch. Toutes les données se trouvant dans Ω_{Test} correspondent à des spectrogrammes annotés sur quelques points (donc $\Omega_{Test} \subsetneq \mathcal{B}_1$). Nous souhaitons nous servir d'eux pour donner des résultats numériques. Les spectrogrammes annotés ponctuellement mais ne faisant pas partie de Ω_{Test} font partie de la base de validation Ω_{Val} afin de calibrer les méta-paramètres des différentes approches comme les seuils de détection. Les moteurs de Ω_{Test} restent les mêmes tout au long de l'étude.

La répartition des données au niveau de chaque patch commence par une classification de ces derniers suivant qu'ils sont normaux ou atypiques. Chaque patch $\left\{Z_{\mathcal{K}_j}\right\}_{j=1,\ldots,\mathrm{card}\,(\mathcal{K})}$ est comparé à la base de données en vérifiant l'intersection entre le patch du spectrogramme étudié et les zones atypiques extraites de ce même spectrogramme. Les patchs dont la surface d'intersection est supérieure à un seuil sont considérés comme atypiques et sont envoyés dans la base de validation Ω^j_{Val} du patch j correspondant. Les patchs dont la surface d'intersection est inférieure à un seuil sont envoyés aléatoirement dans la base d'apprentissage Ω^j_{App} ou dans la base de validation Ω^j_{Val} du patch j correspondant. Le choix d'un seuil non nul de la surface d'intersection est dû à la récupération d'information normale lors de l'extraction des zones atypiques. Les données d'apprentissage et de validation sont donc différentes selon le patch étudié tandis que les données

de test restent les mêmes sur tous les patchs. Ce processus de répartition des données est détaillé dans l'algorithme 5. Pour les études ponctuelles (point à point) des spectrogrammes, nous utilisons la base $\mathcal{B}_1$ des annotations ponctuelles. La base de test reste la même (contenant déjà les annotations ponctuelles) et la base de validation correspond aux spectrogrammes de $\mathcal{B}_1$ non utilisés dans la base de test. La base d'apprentissage reste sélectionnée de la même manière.

Algorithme 5 : Répartition des données en base d'apprentissage et base de validation

Données : Base de données Ω, base de données $\mathcal{B}_1$ des spectrogrammes annotés ponctuellement, la subdivision $\mathcal{K}^{128}$, la surface minimale d'intersection S_{min}, le pourcentage de patchs normaux sélectionnés pour la base d'apprentissage τ_{App}, les données test Ω_{Test}

Résultat : Les données réparties Ω^j_{normal}, Ω^j_{ano}, Ω^j_{App}, Ω^j_{Val}

Initialisation : $\forall j \;\; \Omega^j_{normal} = \emptyset, \;\; \Omega^j_{ano} = \emptyset, \;\; \Omega^j_{App} = \emptyset, \;\; \Omega^j_{Val} = \mathcal{B}_1 \backslash \Omega_{Test}$;

pour *j in 1,...,card(*$\mathcal{K}^{128}$*)* **faire**

  pour chaque *moteur i de* Ω **faire**

    si $\exists k \; zone^i_{ano}(k) \; : \; Surface(Z_j, zone^i_{ano}(k)) > S_{min}$ **alors**

      Ajout de i à Ω^j_{ano}

    sinon

      Ajout de i à Ω^j_{normal}

    fin

  fin

  Sélection aléatoire parmi $\Omega^j_{normal} \backslash \Omega_{Test}$ de τ_{App} moteurs pour Ω^j_{App}

  $\Omega^j_{Val} = \Omega^j_{normal} \backslash \{\Omega_{Test} \cup \Omega^j_{App}\} \bigcup \Omega^j_{ano} \backslash \Omega_{Test}$

fin

retourner Ω^j_{normal}, Ω^j_{ano}, Ω^j_{App}, Ω^j_{Val}, $j \in \{1, ..., \text{card}\,(\mathcal{K}^{128})\}$

$zone^i_{ano}(k)$ correspond aux zones atypiques présentes dans la base de données annoté (Figure 1.8).

La figure 2.2 présente la répartition entre spectrogrammes normaux et atypiques pour chaque patch parmi les $n = 493$ moteurs. Chaque sous-rectangle (encadré en noir) correspond à un des patchs, sa position correspond à la position dans le spectrogramme. L'axe des N_2 a donc été divisé en 3 intervalles et l'axe des fréquences en 18 intervalles à partir de la décomposition en patchs carrés de taille 128 pixels. Le coin inférieur droit du sous-rectangle contient le taux de patchs normaux dans la base de données, et le coin supérieur gauche le taux de patchs atypiques. Comme nous l'avions énoncé dans le chapitre précédent (Figure 1.9), les données ne sont pas équilibrées. La grande majorité des patchs possèdent très peu de données atypiques. Les approches de type one-class/détection d'anomalies, permettant de caractériser le comportement normal des patchs, sont donc adaptées à notre problèmatique. Deux patchs présentent un nombre de données atypiques supérieur à celui des données normales. Il s'agit des patchs les plus complexes où de nombreux types de signatures inusuelles apparaissent. Utiliser des approches supervisées pour caractériser les données atypiques pourrait paraitre plus efficace sur ces patchs. Cependant la parcimonie des signatures inusuelles au sein des patchs et la grande variabilité de