Analyse Exploratoire Approfondie du Dataset PVF-10 pour la Détection des Défauts dans les Panneaux Photovoltaïques

Cette analyse exploratoire approfondie du dataset PVF-10 a permis de caractériser et valider un corpus substantiel de 5579 images thermiques de panneaux photovoltaïques, classifiées en 10 catégories (9 types de défauts et une classe de panneaux sains). L'exploration a révélé des signatures thermiques distinctives selon les types de défauts, validées par des tests statistiques rigoureux. Les caractéristiques de texture et les distributions thermiques présentent un potentiel discriminant significatif pour la modélisation ultérieure. Cette analyse pose les fondations méthodologiques solides pour le développement de systèmes de détection automatique des anomalies dans les installations photovoltaïques.

Introduction et contexte du projet

L'analyse et la détection précises des défauts dans les panneaux photovoltaïques (PV) représentent un enjeu critique pour l'optimisation des performances et la maintenance préventive des installations solaires. Dans le contexte énergétique actuel, avec plus d'un million d'installations en France totalisant 23.7 GW de puissance installée fin 2024[1], la fiabilité et l'efficacité des systèmes photovoltaïques deviennent primordiales.

Importance technique et économique

La thermographie, technique non destructive de capture des variations de température, constitue une approche privilégiée pour cette détection. Les drones équipés de caméras thermiques fournissent une méthode particulièrement efficace, permettant l'inspection rapide et exhaustive de grandes installations photovoltaïques. Sur le plan économique, détecter précocement ces défauts évite des pertes importantes et réduit significativement les coûts d'exploitation et de maintenance.

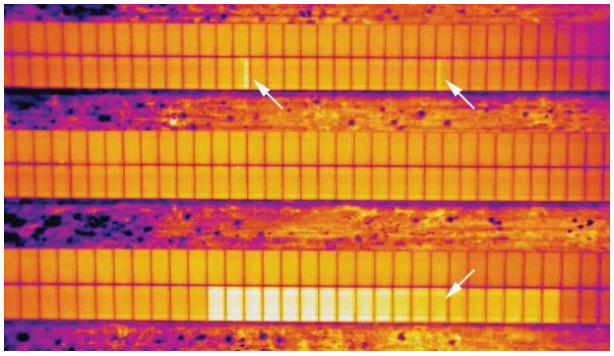

Figure 1 - Exemple d'image thermographique obtenue par drone sur une centrale photovoltaïque. Certains défauts sont mis en évidence (flèches blanches).

[1] https://analysesetdonnees.rte-france.com/production/solaire

Source: FLIR - A guide to inspecting solar fields with thermal imaging drones

Objectifs du projet

Ce projet propose d'approfondir l'analyse des données thermiques via une méthodologie hybride combinant approches statistiques et techniques d'intelligence artificielle. Spécifiquement, les objectifs sont de:

1. Développer un modèle fiable pour détecter et classer précisément les défauts thermiques à partir du dataset PVF-10
2. Optimiser le pré-traitement et l'extraction des caractéristiques essentielles des images thermiques
3. Combiner l'interprétabilité des méthodes statistiques avec la robustesse des approches par Machine Learning
4. L'expertise mobilisée couvre la thermographie, l'analyse statistique approfondie, et les techniques avancées en Machine Learning et Deep Learning pour maximiser la précision et la fiabilité du système de détection développé.

Structure et caractéristiques du dataset PVF-10

Composition globale

Le dataset PVF-10 comprend 5579 images thermiques uniques de modules photovoltaïques individuels, chacune disponible dans trois formats distincts, initialement structurées selon une arborescence hiérarchique: Format > Ensemble (train/test) > Classe. Ces images sont réparties en ensembles d'entraînement et de test (90% et 10% respectivement), avec une distribution homogène des classes entre ces ensembles.

Formats d'images et caractéristiques techniques

Chaque image est déclinée en trois formats distincts, présentant des caractéristiques spécifiques:

- Format original (.tif): Images de dimensions variables (entre 20×20 et 200×200 pixels), reflétant la résolution native de l'acquisition par drone. Ces images correspondent à l'extraction (segmentation) de panneaux individuels identifiés dans des images(telles que celle de la Figure 1) prises par drone lors de l'inspection de 8 centrales photovoltaïques différentes.
- Format 110×60 (.png): Format standardisé rectangulaire correspondant à la forme typique des panneaux. Cependant, environ 4% des images présentent des dimensions différentes, généralement carrées (≈60×60).
- Format 112×112 (.png): Format carré parfaitement homogène, optimisé pour l'entrée des réseaux de neurones convolutifs.

Toutes les images sont en couleur (3 canaux RGB), avec une dominance marquée de la composante rouge et une faible composante bleue, caractéristique des images thermiques en pseudo-couleur. On notera cependant que l'affichage en pseudo-couleur n'est qu'un moyen de rendre plus visible à l'oeil humain les différences de température à la surface des panneaux : une image thermographique est en réalité une image dont la valeur de chaque pixel est équivalente à une valeur de température (en K ou °C), telle que mesurée par le capteur thermique de la caméra dont est équipé le drone utilisé pour l'inspection.

Image DJI_20230223132111_0227_T_000001 : taille originale 212x170

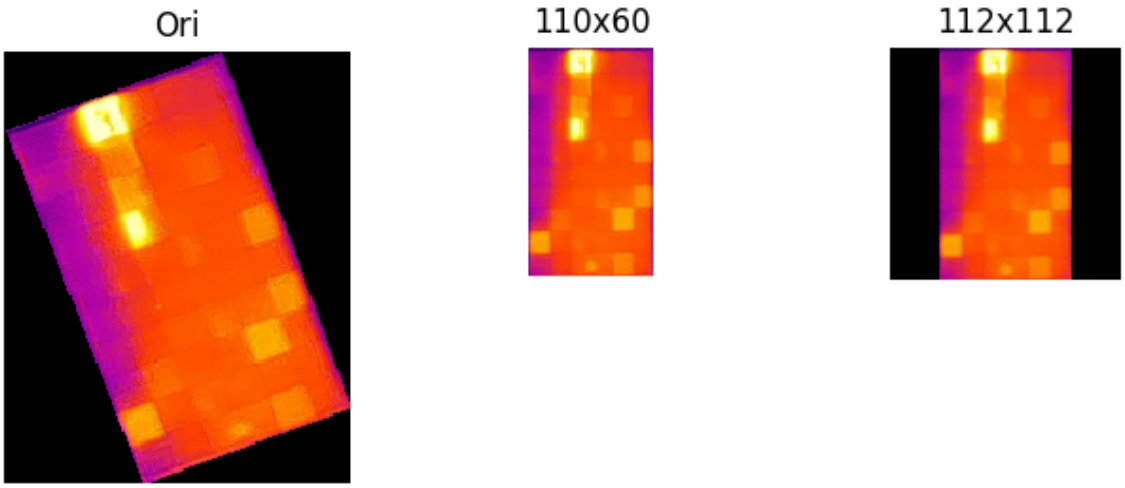

Image DJI_20230223130819_0183_T_000002 : taille originale 211x165

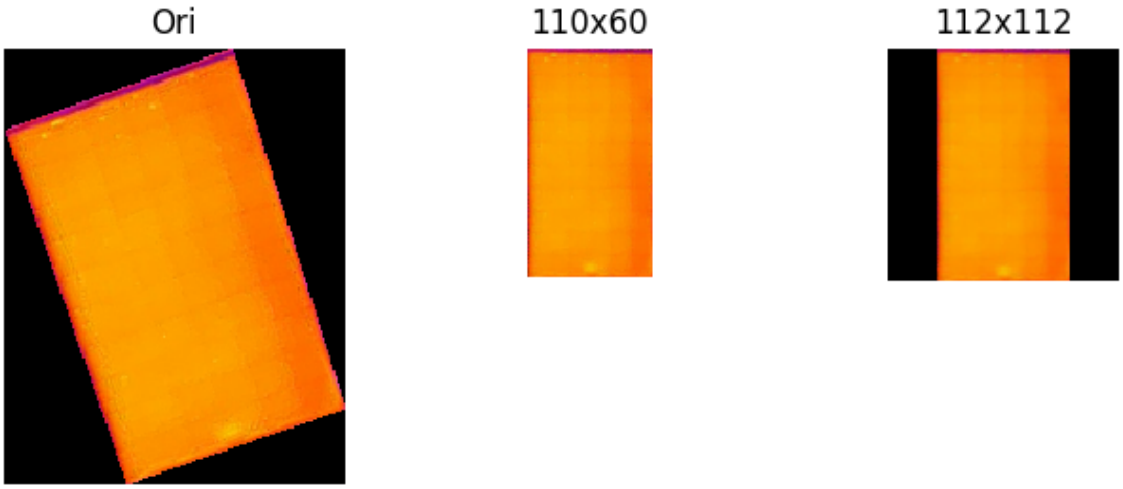

Figure 2 - Exemple d'images dans leur format original, 110x60 et 112x112

Distribution des classes et anomalies

Le dataset est structuré en 10 classes représentatives des défauts les plus couramment rencontrés en conditions d'exploitation réelles des panneaux photovoltaïques:

- **bottom dirt ;** Accumulation de salissures (poussière, boue, sable) sur le bas du panneau. **Impact :** diminution locale de l'irradiance, légère surchauffe des zones propres.
- **break :** Fissure ou rupture visible d'une cellule ou d'un module. **Impact :** forte surchauffe locale, danger de points chauds. Défaut critique.
- **debris cover :** Présence d'un objet étranger sur la surface (feuilles, plastiques, etc.). **Impact :** ombrage irrégulier entraînant une élévation thermique hétérogène.
- **junction box heat :** Surchauffe localisée au niveau de la boîte de jonction. **Impact :** défaut électrique potentiellement dangereux, perte d'efficacité.
- **hot cell :** Cellule ou groupe de cellules présentant une température anormalement élevée. **Impact :** défaut thermique ponctuel souvent causé par une mauvaise connexion ou une cellule défectueuse.
- **shadow :** Ombrage partiel dû à des éléments extérieurs (branches, câbles, etc.). **Impact :** baisse de rendement temporaire, souvent visible en bandes froides.
- **short circuit panel :** Court-circuit généralisé affectant l'ensemble du panneau. **Impact :** très forte surchauffe homogène, risque de dégradation accélérée.
- **string short circuit :** Court-circuit affectant une chaîne de cellules. **Impact :** surchauffe linéaire visible dans une zone continue du panneau.
- **substring open circuit :** Ouverture du circuit dans une sous-chaîne de cellules. **Impact :** surchauffe isolée, comportement thermique anormal sur une ligne.
- **healthy panel :** Panneau sans défaut thermique ou structurel. Référence de fonctionnement normal, utilisée pour la comparaison et l'apprentissage supervisé.

La classe 'healthy panel' est légèrement surreprésentée (~27% des données), tandis que les classes 'break' et 'string short circuit' sont les moins fréquentes (2,35% et 1,27% respectivement). Cette distribution relativement équilibrée permet d'envisager une modélisation robuste, moyennant des techniques appropriées pour les classes minoritaires.

Méthodologie d'exploration et pré-processing

Approche structurée d'exploration

L'analyse exploratoire a été menée selon une méthodologie rigoureuse combinant:

1. Analyse structurelle: examen de l'organisation du dataset, vérification des métadonnées, analyse des formats et dimensions
2. Analyse visuelle: visualisation représentative des différentes classes, formats et caractéristiques thermiques
3. Analyse statistique: quantification des propriétés thermiques, tests de différenciation entre classes, modélisation des distributions
4. Analyse spatiale et texturale: extraction de caractéristiques avancées, évaluation des structures spatiales des anomalies thermiques

Cette approche multi-dimensionnelle a permis une caractérisation complète du dataset et l'identification des axes de prétraitement prioritaires.

Prétraitement et feature engineering

Le prétraitement des données a comporté plusieurs étapes essentielles:

1. Vérification de l'intégrité: toutes les images ont été analysées pour confirmer leur lisibilité et l'absence de corruption.
2. Identification des anomalies dimensionnelles: détection des images de format 110×60 non conformes aux dimensions attendues (~4% du dataset).
3. Détection et suppression des doublons: identification de 21 doublons exacts (7 images uniques dupliquées dans chaque format) et suppression pour garantir l'intégrité du dataset.
4. Conversion et normalisation: transformation en niveaux de gris pour l'analyse thermique, avec préservation de l'information de température.

🔍 **Limitation majeure sur la reconstruction des températures réelles :**
Un point essentiel à noter est que **nous ne disposons d'aucune information sur l'algorithme utilisé pour convertir les températures en niveaux de gris, puis en pseudo-couleurs RGB**. À aucun moment la publication ne détaille cette étape, ce qui rend toute tentative de reconstruction des températures physiques **impossible**.

Bien qu'on puisse supposer que la palette "Inferno" classique ait été utilisée pour la conversion RGB, **rien ne le confirme** formellement. Et même si c'était le cas, **la transformation initiale température → grayscale reste indéchiffrable**, rendant l'opération non inversible.

✅ Cela dit, **dans le cadre strict du dataset PVF10**, ce n'est pas bloquant : on peut raisonnablement supposer que **la correspondance entre valeurs de pixels et températures est cohérente d'une image à l'autre**.
❗ En revanche, **si l'on souhaitait appliquer notre approche à un autre dataset**, cette opacité sur la conversion pourrait devenir un **frein majeur à la généralisation** de notre méthode.

L'extraction de caractéristiques (feature engineering) a été particulièrement approfondie, incluant:

- Extraction d'indicateurs statistiques (températures minimales, maximales, moyennes, médianes, écarts-types)
- Calcul des percentiles pour caractériser finement les distributions thermiques
- Analyse d'histogrammes de niveaux de gris pour capturer les signatures de défauts

Ces caractéristiques constituent une base solide pour la modélisation ultérieure, permettant de capturer diverses dimensions des anomalies thermiques.

Analyse statistique et visualisations approfondies

Visualisation des signatures de défauts

Les visualisations réalisées ont permis de mettre en évidence des caractéristiques visuelles distinctives pour chaque classe:

- Les histogrammes de niveaux de gris globaux montrent des profils clairement différenciés selon les classes.
- Les cartes d'entropie révèlent des patterns de complexité locale spécifiques à certains types de défauts.
- Les résultats du filtrage de Canny illustrent des structures de contours caractéristiques, notamment pour les défauts impliquant des fractures ou des court-circuits.

Ces visualisations constituent non seulement des outils d'analyse précieux, mais également des supports de validation et d'interprétation des modèles qui seront développés.

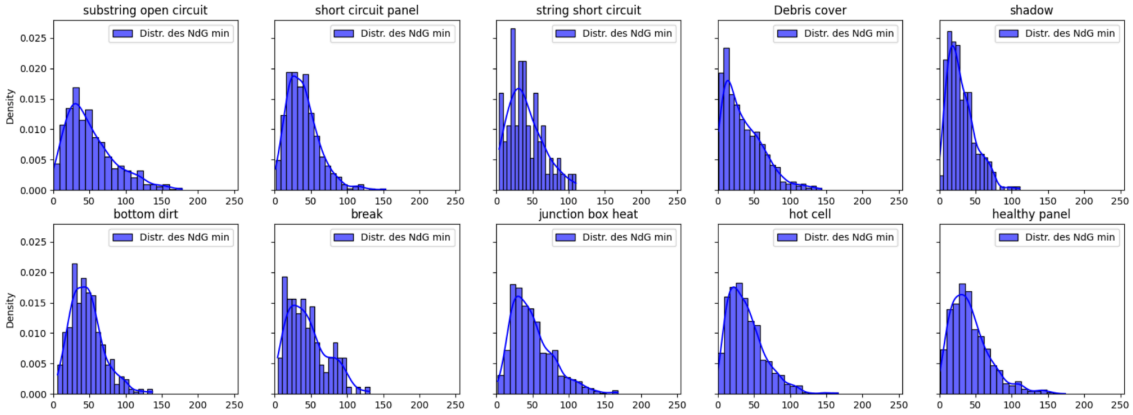

Figure 3 - Histogramme de la distribution des valeurs minimales de niveaux de gris par classe

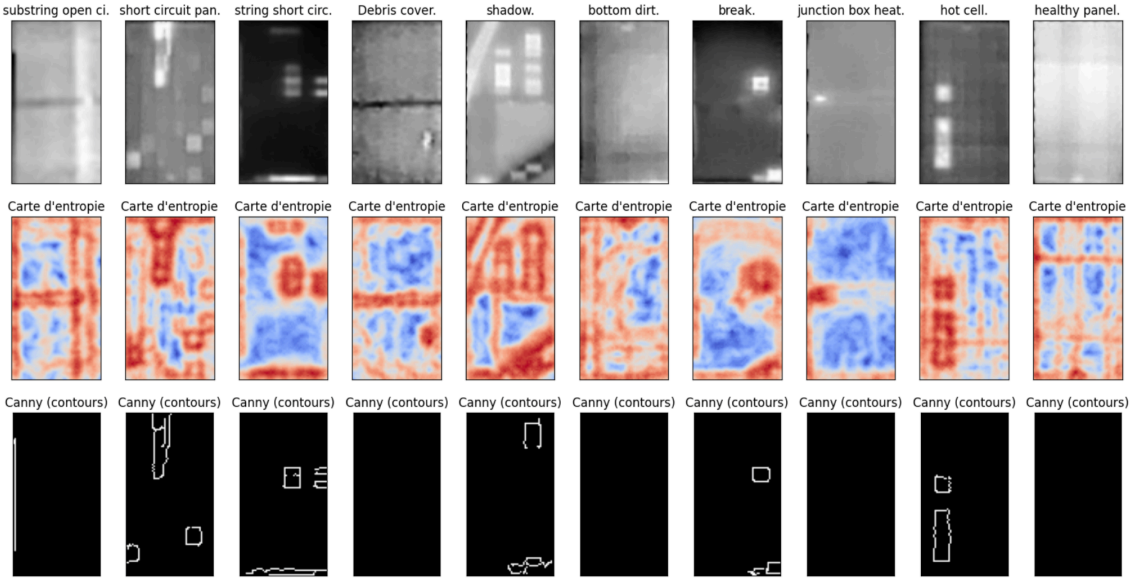

Figure 4 - Cartes d'entropie et filtres de Canny par classe

Caractérisation thermique des défauts

L'analyse statistique approfondie des distributions thermiques a révélé des signatures distinctives selon les classes:

- Les défauts de type 'break' et 'short circuit panel' présentent des températures maximales significativement plus élevées, témoignant de points chauds intenses résultant d'un effet résistif localisé.
- Les catégories 'shadow' et 'bottom dirt' montrent des distributions plus plates et des températures minimales plus basses, caractéristiques de leur impact thermique spécifique.
- La classe 'healthy panel' présente une distribution thermique plus homogène, avec une variance réduite, indiquant une conversion énergétique uniforme et efficiente.

Des tests statistiques (Kruskal-Wallis + test de Dunn-Bonferroni) ont confirmé une différence entre les classes vis-à-vis des indicateurs statistiques de la distribution thermique, différence plus marquée en ce qui concerne le maximum et la variance de la distribution.

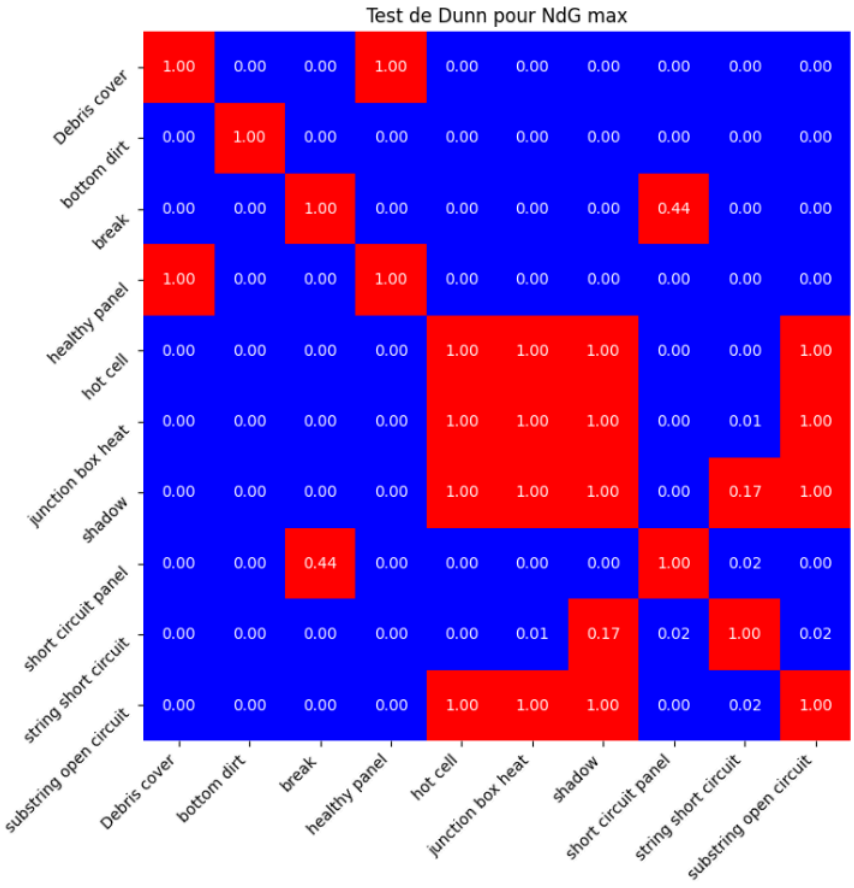

Figure 5 - Test statistique de Dunn (post Kruskal-Wallis) sur les valeurs maximales des niveaux de gris. Le bleu indique une différence significative (< 0.05) pour chaque paire de classe

Analyse texturale et spatiale

Au-delà des indicateurs statistiques simples, l'analyse texturale a apporté une dimension complémentaire essentielle:

- La densité de contours, mesurée après application du filtre de Canny, s'est révélée significativement plus élevée pour les défauts 'break', 'short circuit panel' et 'string short circuit', correspondant à des transitions thermiques abruptes.
- L'analyse d'entropie a mis en évidence des niveaux de complexité locale variables selon les classes, avec une complexité accrue pour certains défauts comme 'short circuit panel'.
- Les caractéristiques GLCM (contraste, homogénéité, énergie, corrélation)[2] ont démontré un pouvoir discriminant significatif, particulièrement le contraste et l'énergie.

[2] L'analyse GLCM (Gray-Level Co-occurence Matrix, ou matrice des niveaux de gris en francais) est une technique utilisée en traitement d'images pour analyser la texture d'une image. Elle repose sur la manière dont les niveaux de gris des pixels sont répartis et associés spatialement.

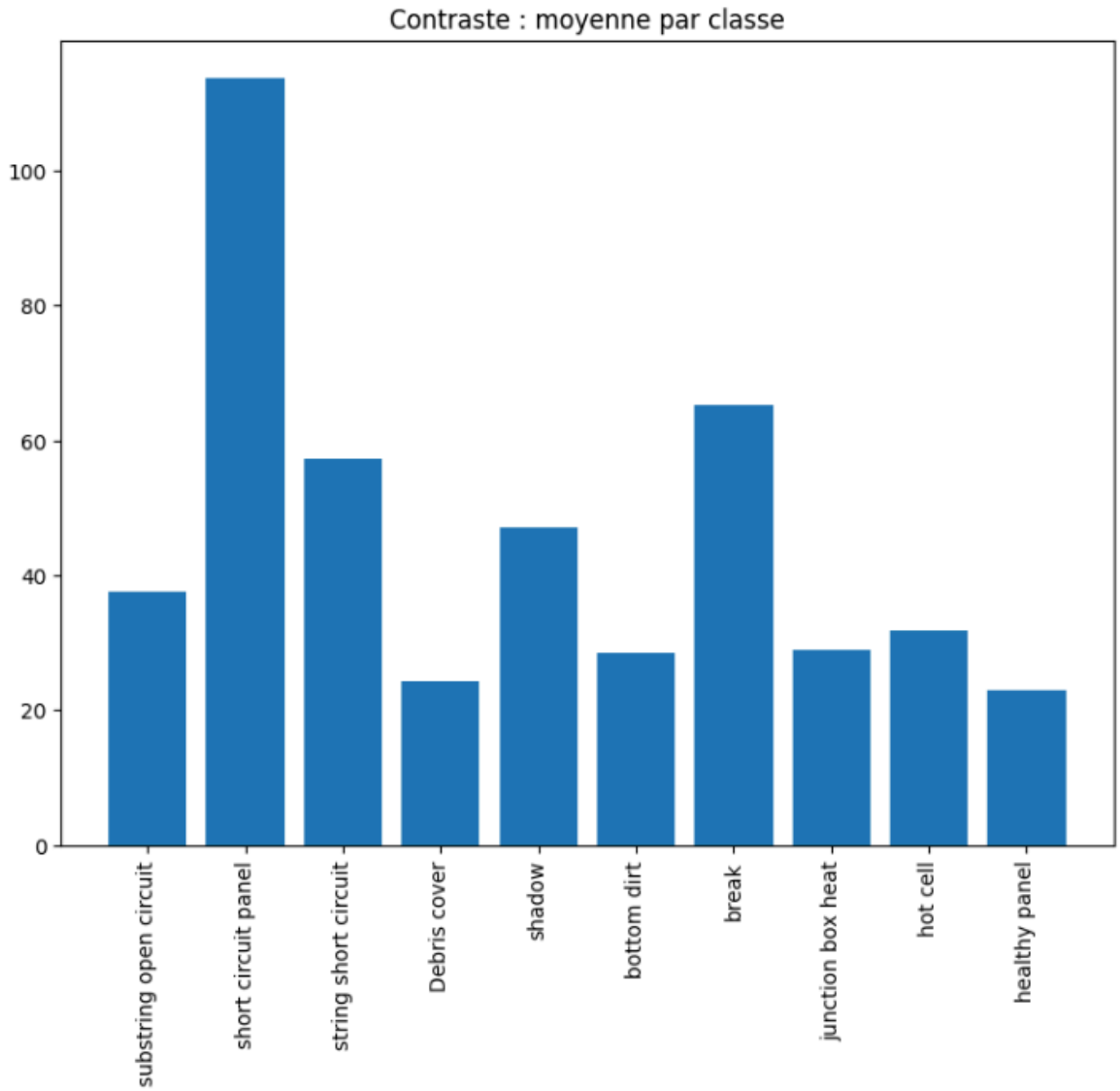

Figure 6 - Moyenne des contrastes pour chaque classe, calculés à partir des matrices GLCM

Les tests statistiques ont confirmé que ces différences texturales sont statistiquement significatives (p-value < 0.05), offrant une dimension complémentaire pour la classification des défauts.

Points particuliers et anomalies identifiées

Anomalies dimensionnelles

Un point d'attention significatif concerne les images du format 110×60 dont environ 4,19% présentent des dimensions réelles différentes, généralement proches du carré (60×60). L'analyse a révélé que ces images proviennent d'images originales déjà proches du carré, pour lesquelles l'étape de redimensionnement a probablement été omise lors de la conversion.

Cette anomalie constitue un point d'attention pour la modélisation, nécessitant soit:

- Un redimensionnement préalable pour homogénéiser les entrées
- Une exclusion de ces images pour garantir la cohérence du dataset

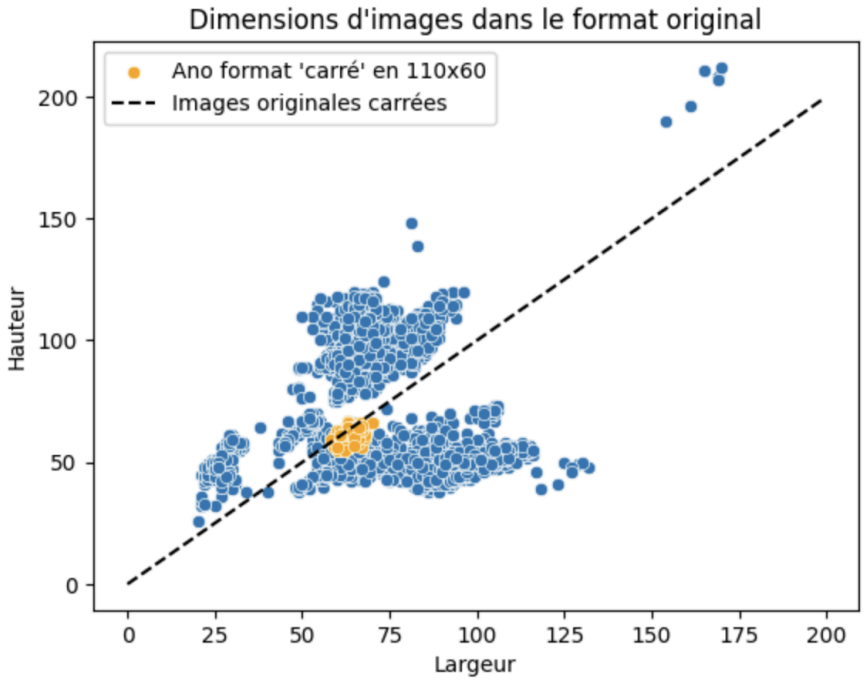

Figure 7 - illustration anomalies images 110x60 au format carré

Présence de doublons

L'analyse a identifié 21 doublons exacts répartis dans les trois formats, correspondant à 7 images uniques dupliquées dans chaque format. Ces doublons ont été supprimés pour éviter tout biais d'apprentissage, réduisant légèrement la taille du dataset (de 5579 à 5572 images uniques par format).

Doublon 1 :

	Chemin	Format	Train_Test	Classe	Nom	Type	Largeur	Hauteur	Canaux	Doute_Carre
2482	PVF-10\PVF_10_110x60\train\05hot cell\DJI_2023...	110x60	train	hot cell	DJI_20230227163409_0290_T_000001	.png	60	110	3	False
2484	PVF-10\PVF_10_110x60\train\05hot cell\DJI_2023...	110x60	train	hot cell	DJI_20230227163409_0290_T_000005	.png	60	110	3	False

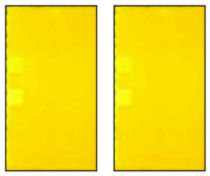

Figure 8 - Exemple de doublon

Distribution des classes et stratégies d'équilibrage

Bien que relativement équilibrée, la distribution des classes présente une légère surreprésentation de la classe 'healthy panel' (~27%) et une sous-représentation des classes 'break' et 'string short circuit' (respectivement 2,35% et 1,27%).

Pour pallier ce déséquilibre modéré, plusieurs stratégies ont été envisagées:

- Pondération des classes lors de l'entraînement
- Techniques de suréchantillonnage pour les classes minoritaires
- Approche d'apprentissage hiérarchique pour les classes difficiles à différencier

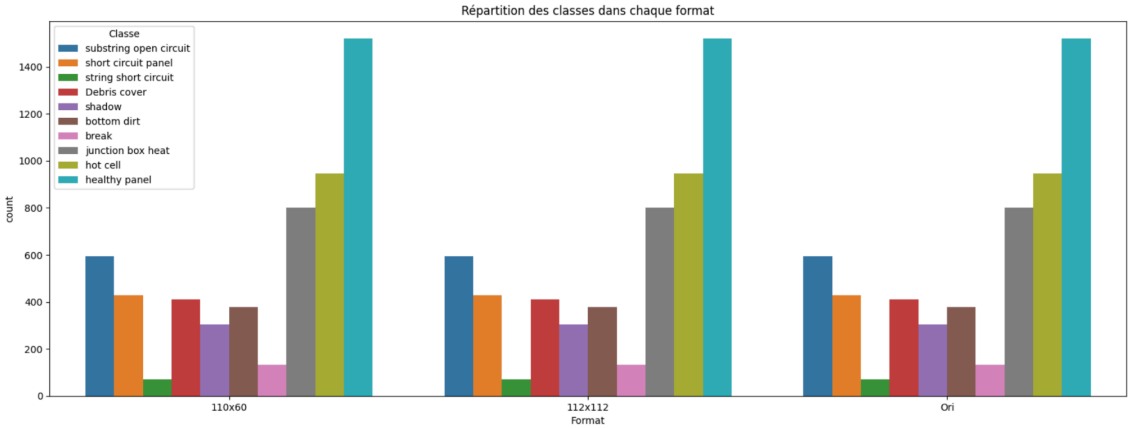

Figure 9 - Répartition des classes dans chaque format

Stratégies de modélisation recommandées

Approches hybrides et complémentaires

L'exploration approfondie du dataset PVF-10 suggère l'adoption d'une stratégie de modélisation hybride combinant:

1. Modèles statistiques de référence: utilisés comme baseline et pour leur interprétabilité, exploitant les indicateurs thermiques extraits (températures, percentiles, moments statistiques).
2. Réseaux de neurones convolutifs (CNN): particulièrement adaptés pour capturer les motifs visuels complexes directement à partir des images standardisées, notamment dans le format 112×112.
3. Pipelines séquentiels: combinant la filtration initiale par méthodes statistiques, suivie d'une analyse fine via deep learning pour les cas ambigus ou complexes.

Cette approche permettra d'exploiter la complémentarité entre interprétabilité des méthodes statistiques et puissance discriminative des réseaux de neurones.

Recommandations spécifiques pour le prétraitement

Pour optimiser les performances des modèles, plusieurs recommandations de prétraitement ont été formulées:

- Filtrer ou redimensionner les images hors norme du format 110×60
- Normaliser les intensités thermiques pour standardiser l'entrée des modèles
- Extraire des caractéristiques avancées (gradient, moments, descripteurs de texture)
- Développer des stratégies d'augmentation de données pour enrichir l'apprentissage, particulièrement pour les classes sous-représentées

Ces prétraitements devront être intégrés dans un pipeline cohérent garantissant la reproductibilité et la robustesse de l'approche.

Conclusion et perspectives

Synthèse des découvertes

L'analyse exploratoire approfondie du dataset PVF-10 a permis de:

1. Valider la structure et la qualité du dataset, avec l'identification et le traitement des anomalies (dimensions irrégulières, doublons)
2. Établir des signatures thermiques et texturales distinctives pour chaque classe de défaut, validées par des tests statistiques rigoureux
3. Développer une méthodologie robuste d'extraction de caractéristiques multidimensionnelles, combinant aspects thermiques et texturaux
4. Proposer des stratégies de modélisation adaptées exploitant la complémentarité des approches statistiques et de deep learning

Le dataset PVF-10 démontre un potentiel élevé pour une classification précise des défauts photovoltaïques grâce aux caractéristiques thermiques extraites et aux méthodologies de pré-processing développées.

Perspectives futures

Les prochaines étapes comprendront:

1. L'implémentation des modèles statistiques simples (Random Forest, KNN, etc.) exploitant les caractéristiques extraites
2. Le développement de modèles de deep learning adaptés aux spécificités des images thermiques
3. L'évaluation comparative des différentes approches et leur optimisation

À terme, cette recherche contribuera à l'optimisation de la performance des installations photovoltaïques existantes, représentant un levier majeur pour maximiser la contribution de l'énergie solaire à la transition énergétique.

Références

Bo Wang, Qi Chen, Mengmeng Wang, Yuntian Chen, Zhengjia Zhang, Xiuguo Liu, Wei Gao, Yanzhen Zhang, Haoran Zhang, **PVF-10: A high-resolution unmanned aerial vehicle thermal infrared image dataset for fine-grained photovoltaic fault classification**, Applied Energy, Volume 376, Part A, 2024, 124187, ISSN 0306-2619, https://www.sciencedirect.com/science/article/pii/S0306261924015708