I. Description du projet :

Le projet a pour objectif de créer un modèle de classification multimodale capable de reconnaître et de classer des scènes à partir d'images et de fichiers audio. Il vise à extraire des caractéristiques pertinentes des données visuelles et auditives, en intégrant divers modèles d'apprentissage automatique et d'apprentissage profond, afin d'améliorer la précision de la classification. Ce projet s'inscrit dans le domaine de l'apprentissage automatique et de la vision par ordinateur, et utilise des techniques avancées de traitement d'images et de signaux audio.

II. Objectifs du Projet :

L'objectif principal de ce projet de classification multimodale d'images et d'audio est de développer des modèles capables de reconnaître et de classer des scènes à partir de données visuelles et auditives. Plus précisément, les objectifs incluent :

1. Extraction de Caractéristiques : Utiliser un modèle VGG19 pré-entrainé sur la base de données 'imagenet' et d'explorer d'autres alternatives notamment RESNET et INCEPTION pour extraire les caractéristiques des images, et le LSTM pour extraire les caractéristiques pertinentes à partir des coefficients MFCC des enregistrements audio.

2. Classification Précise : Développer des modèles de machine learning et de réseaux de neurones denses pour les images et des LSTM pour l'audio afin d'atteindre une classification précise des différentes scènes.

3. Analyse des Performances : Évaluer les modèles via des matrices de confusion et des métriques comme la précision, le rappel et rapport de classification pour identifier les forces et les faiblesses.

4. Intégration Multimodale : Explorer l'impact de l'utilisation de différentes architectures et méthodes de fusion des données, notamment la méthode tardive (late fusion), précoce (early fusion) et hybride sur la performance de notre système final.

III. Description de la base de données :

A. Contexte :

Le dataset "Scene Classification: Images and Audio" disponible sur Kaggle propose une collection de données multimodales pour la classification de scènes. Chaque scène est représentée par un dossier contenant des images (frames) capturées pour chaque seconde de la vidéo correspondante, ainsi que des caractéristiques MFCC de l'audio associé à chaque frame. Cette combinaison permet d'explorer comment l'intégration de données visuelles et auditives peut accroître la précision des modèles de classification.

B. Structure de la Base de Données:

Dans notre dataset, nous avons :

- **Dossier "images"** :
    - Contient plusieurs sous-dossiers, chacun représentant une scène spécifique.

- Les sous-dossiers représentent des catégories telles que "beach", "classroom1", "restaurant1", "restaurant2", etc.
- Chaque sous-dossier renferme plusieurs images correspondant à la scène.

- **Fichier "dataset.csv"** :
  - Contient des informations sur chaque enregistrement audio associé à une image.
  - Établit un lien entre les données visuelles et auditives.

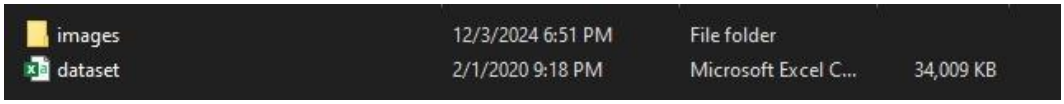

Figure 1:Structure du Dataset Multimodal : Images et Fichier CSV

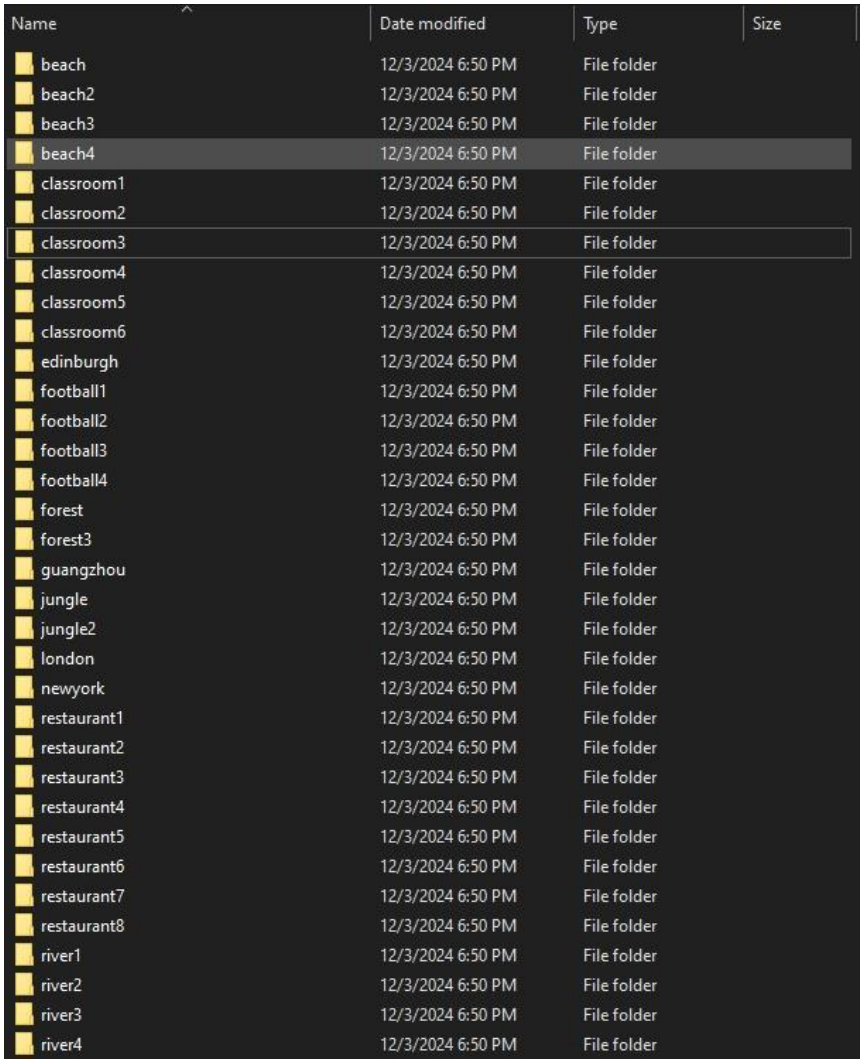

Figure 2: Sous-dossiers d'Images : Catégories de Scènes

C. Détails du Fichier CSV du Dataset:

- **Image :** cette colonne est un chemin qui relie les données visuelles aux données auditives, sous la forme de (Dossier racine (‘Images’)/ sous-catégorie/nom de l’image).

- **Mfcc_1 à mfcc_104 :** Ces colonnes représentent les coefficients cepstraux en fréquence Mel (MFCC), qui sont des caractéristiques acoustiques extraites pour chaque seconde de la vidéo de la scène (associée à l'image). Les MFCC sont largement utilisés dans le traitement du signal et l'analyse sonore, car ils capturent les propriétés spectrales de l'audio, essentielles pour la reconnaissance et la classification des sons.

- **CLASS1 et CLASS2** :
    - **CLASS1** : Classe générale indiquant le type d'environnement (par exemple, OUTDOORS, INDOORS).
    - **CLASS2** : Classe spécifique (la classe à prédire par nos modèles) représente la catégorie principale (par exemple, FOREST, GROCERY-STORE).

- **Taille et Dimensions :**
    - **Nombre de lignes** : La base contient 17 252 instances d'échantillons.
    - **Nombre de colonnes** : Au total, il y a 107 colonnes :
        - 1 colonne pour le chemin d'accès à l'image.
        - 104 colonnes pour les coefficients MFCC.
        - 2 colonnes pour les labels (CLASS1 et CLASS2).

```
[12] mfcc_data
```

	IMAGE	mfcc_1	mfcc_2	mfcc_3	mfcc_4	mfcc_5	mfcc_6	mfcc_7	mfcc_8	mfcc_9	...	mfcc_97	mfcc_98	mfcc_99	mfcc_100	mfcc_101	mfcc_102	mfcc_103	mfcc_104	CLASS1	CLASS2
0	images/forest/forest0.png	15.795384	-3.442518	-25.316836	-33.412104	2.447290	-46.981182	12.889984	-23.588534	-22.625879	...	-43.876462	20.697491	-22.793173	-9.417196	13.762870	-31.976786	18.461561	-13.140674	OUTDOORS	FOREST
1	images/forest/forest1.png	15.883880	-3.494075	-21.189490	-18.077115	4.284962	-27.014271	3.666955	-9.091312	-3.746509	...	-33.883092	17.223236	-24.985005	12.035913	8.321000	-16.249293	8.717523	0.743640	OUTDOORS	FOREST
2	images/forest/forest2.png	17.872629	-18.877467	-31.665319	-47.045579	1.813430	-45.899877	14.975982	-24.462396	-1.812962	...	-34.456028	21.433239	-14.190274	-8.629235	1.035640	-20.703358	5.986662	-14.644013	OUTDOORS	FOREST
3	images/forest/forest3.png	16.843997	-3.527753	-21.282970	-24.248141	27.201589	-18.787674	30.093938	-1.922008	10.156418	...	-36.410615	19.949251	-5.466172	6.480569	13.070739	-14.853299	10.243606	-17.983957	OUTDOORS	FOREST
4	images/forest/forest4.png	16.128583	-4.267328	-25.608325	-20.231084	15.922823	-35.703313	16.307644	-3.547505	4.804142	...	-41.548915	15.697646	-20.615005	-11.942869	5.421639	-27.445147	9.060233	-15.077528	OUTDOORS	FOREST
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17247	images/store4/store4-1026.png	14.239928	24.677429	-16.642587	-16.333397	28.834652	-48.712410	45.642448	-21.947515	-3.759865	...	-40.159954	30.120032	-13.060528	-9.219403	12.475126	-22.801245	17.750249	-15.017793	INDOORS	GROCERY-STORE
17248	images/store4/store4-1027.png	14.718721	1.950969	-12.385012	-11.159338	-9.408496	-6.195532	-5.524698	-3.498479	1.615768	...	-52.602034	24.761429	-27.102315	-3.128097	27.192930	-32.536263	24.921955	-13.841536	INDOORS	GROCERY-STORE
17249	images/store4/store4-1028.png	14.600713	4.169370	-8.191752	-10.247319	-17.536838	-12.626182	-10.417944	2.356240	-1.213015	...	-31.669516	23.663856	-8.380984	5.988465	6.474730	-20.457587	-16.182687	-10.610574	INDOORS	GROCERY-STORE
17250	images/store4/store4-1029.png	14.319269	18.059742	-13.402903	-13.233159	14.580749	-42.086968	14.006702	-18.285120	-4.757876	...	-20.178121	22.866183	-14.766197	-4.440472	13.921726	-16.048833	4.011102	0.031063	INDOORS	GROCERY-STORE
17251	images/store4/store4-1030.png	14.150456	2.506630	-8.436129	-8.280871	-12.001908	2.700066	0.029298	-4.737745	-4.447416	...	-47.612023	37.086938	-26.782493	-17.047327	6.179108	-20.238054	19.269190	-5.813571	INDOORS	GROCERY-STORE

17252 rows × 107 columns

Figure 3: Structure du Fichier CSV du Dataset

IV. Protocole Expérimental :

Le protocole expérimental comprend plusieurs étapes clés :

1. Prétraitement des Images :

- **Redimensionnement :** redimensionnement des images pour qu'elles soient adéquates aux modèles d'extraction des caractéristiques (VGG19, RESNET, INCEPTION).
- **Réduction de Bruit :** Appliquez un filtre médian avec une fenêtre de taille 3 pour réduire le bruit dans les images.
- **Augmentation du Contraste :** Amplifiez le contraste de l'image avec un facteur de 1.5 pour améliorer la visibilité des détails.

- **Normalisation :** Normalisez les pixels de l'image dans l'intervalle [0, 1] pour standardiser les valeurs.

2. Extraction des Caractéristiques des Images (cas de VGG19):

- **Chargement du Modèle VGG19** : Utilisez le modèle VGG19 préentraîné sur la base de données « Imagenet » pour extraire les caractéristiques.
- **Extraction de Caractéristiques :** Prenez les caractéristiques de la couche "fc2", qui est l'une des dernières couches denses avant la sortie finale.
- **Stockage dans un DataFrame :** Créez un DataFrame nommé scenes_data pour stocker le nom de la scène, le nom de l'image et ses 4096 caractéristiques extraites.

	subclass	IMAGE	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	...	C_4086
0	beach	beach0.png	0.538444	0.0	0.463962	0.028086	1.296627	0.921913	0.0	0.0	...	0.038777
1	beach	beach1.png	0.478749	0.0	0.473211	0.054321	1.282259	0.920566	0.0	0.0	...	0.085440

Figure 4: DataFrame scenes_data

3. Prétraitement des Caractéristiques MFCC (Audio):

- **Importation des Données Auditives :** Chargez le DataFrame contenant les caractéristiques MFCC, avec la colonne IMAGE indiquant le chemin des images traitées.
- **Vérification des Valeurs Manquantes :** Assurez-vous qu'il n'y a pas de valeurs manquantes (NaN) ou de chaînes vides. Si des valeurs manquantes sont trouvées, la fonction doit renvoyer True.
- **Modification de la Colonne IMAGE :** Conservez seulement le nom de l'image dans la colonne IMAGE et ajoutez une nouvelle colonne subclass avec le nom de la scène.
- **Normalisation des Valeurs :** Normalisez les valeurs numériques dans les colonnes MFCC pour assurer une échelle uniforme.
- **Suppression de la Colonne CLASS1 :** Éliminez la colonne CLASS1, car elle n'est pas nécessaire pour l'analyse.
- **Label Encoding de CLASS2 :** Associez à chaque valeur unique dans la colonne CLASS2 un entier unique à l'aide du Label Encoding.

	subclass	IMAGE	mfcc_1	mfcc_2	mfcc_3	mfcc_4	mfcc_5	mfcc_6	mfcc_7
439	beach	beach0.png	-0.174343	0.251045	0.576211	-0.453543	0.603207	-0.573730	0.824263
440	beach	beach1.png	0.020074	0.139168	0.661423	0.138959	0.779691	-0.801382	0.634700
449	beach	beach10.png	-0.636062	0.157091	0.759928	0.163945	0.389139	-0.161064	0.438196
539	beach	beach100.png	0.549627	-0.567954	0.052918	-0.580181	-0.203623	-0.444691	0.115973
540	beach	beach101.png	-0.146405	-0.722738	1.182417	0.380163	-0.139631	-0.019968	-0.499439

Figure 5:Prétraitement des Données Auditives (MFCC)

4. Les expériences:

Le protocole expérimental décrit ci-dessus a été mis en œuvre dans le but de développer et d'évaluer plusieurs méthodes de classification multimodale, notamment les méthodes tardives, précoces et hybrides.

A. Méthode tardive:

a. Division des Données:

- **Séparation des Données :** division des données visuelles (les caractéristiques des images) et auditives (les caractéristiques MFCC) en ensembles d'entraînement (2/3) et de test (1/3), en préservant la proportion des classes dans chaque sous-catégorie.

b. Entraînement du Modèle de Réseau de Neurones Dense:

- **Définition et Compilation du Modèle :** Création d'un modèle de réseau de neurones dense, et son compilation avec une fonction de perte appropriée et un optimiseur.
- **Entraînement :** Entraînement du modèle sur les données d'entraînement visuelles.

c. Entraînement d'un Modèle LSTM:

- **Définition du Modèle :** Création et compilation du modèle LSTM en spécifiant les paramètres appropriés.
- **Entraînement :** Entraînement du modèle LSTM sur les données MFCC d'entraînement.

Principe de la fusion tardive :

- **Traitement indépendant des modalités :** Chaque modèle traite son propre type de données de manière indépendante.
- Réseau de neurones dense s'entraine sur les caractéristiques des images.
- LSTM s'entraine sur les caractéristiques MFCC.
- **Génération des prédictions individuelles :** Chaque modèle produit une sortie (une vecteur de probabilités). Ces sorties sont généralement homogènes.
- **Combinaison des sorties** : Les prédictions des modèles sont ensuite combinées pour produire une décision finale.

La combinaison peut se faire de différentes manières :

d. Fusion par Moyenne :

- **Calcul des Probabilités :** Calculez la moyenne des probabilités prédites par le modèle de réseau de neurones dense et le modèle LSTM pour chaque échantillon.
- **Classe Maximale :** Identifiez la classe correspondant à la probabilité maximale à partir des moyennes.

e. Bagging à base de RandomForset:

- **Obtenir les Prédictions :** Utilisez les modèles pour obtenir les prédictions sur le jeu d'entraînement.
- **Création de Nouvelles Caractéristiques :** Employez les prédictions des deux modèles comme nouvelles caractéristiques pour entrainer ce nouveau classificateur.
- **Décision final :** la décision de ce modèle.

B. Méthode précoces:

La **méthode précoce** consiste à fusionner des données provenant de différentes modalités (l'image et l'audio) dès le début du processus, avant de les soumettre aux algorithmes de classification.

1. **Extraction des caractéristiques** pertinentes d'après les caractéristiques MFCC par un modèle LSTM

- Les caractéristiques (features) des données audio et visuelles sont extraites séparément.
- Ces caractéristiques sont ensuite combinées.
- Cette fusion est effectuée avant le passage des données au modèle de classification.

	Label	SubClass	IMAGE	feature_0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	...	feature_54	feature_55	featu
0	FOREST	forest	forest0.png	0.0	0.149667	0.000000	0.225900	0.000000	0.0	0.0	...	0.0	0.0	0.5
1	FOREST	forest	forest1.png	0.0	0.000000	0.000000	0.035934	0.000000	0.0	0.0	...	0.0	0.0	0.5
2	FOREST	forest	forest2.png	0.0	0.510642	0.020100	0.360731	0.004015	0.0	0.0	...	0.0	0.0	0.3
3	FOREST	forest	forest3.png	0.0	0.143005	0.025325	0.093704	0.000000	0.0	0.0	...	0.0	0.0	1.2
4	FOREST	forest	forest4.png	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	...	0.0	0.0	0.3

Les méthodes utilisées :

a. Réseaux de neurones:

- Un réseau de neurones est entraîné pour apprendre une représentation commune des données fusionnées.
- Les données combinées sont traitées à travers plusieurs couches pour réaliser la classification.
- L'apprentissage adaptatif des poids permet de capter des relations complexes entre les modalités.

b. SVM (Support Vector Machines):

- L'algorithme cherche à séparer les classes dans l'espace des caractéristiques fusionnées en maximisant la marge.
- Un noyau approprié (rbf) est utilisé pour traiter les données fusionnées et capturer les relations complexes.

c. Forêts aléatoires:

- Chaque arbre de la forêt est entraîné sur des échantillons des données fusionnées.
- Le modèle utilise des arbres de décision pour voter et déterminer la classe majoritaire.
- Cette méthode est robuste face aux données bruitées et gère efficacement les ensembles fusionnés.

C. Méthode hybride:

La **fusion hybride** combine les avantages de la fusion précoce et tardive. Elle intègre les de différentes modalités dans une étape intermédiaire, permettant d'exploiter à la fois les interactions croisées entre modalités (comme le fusion précoce) et les prédictions indépendantes des modèles (comme le fusion tardive).

a. Division des Données:

Les données (caractéristiques des images et caractéristiques extraites par le LSTM à partir de MFCC) sont initialement partagées en deux ensembles :

- 80 % pour l'entraînement
- 20 % pour le test

L'ensemble d'entraînement est ensuite divisé en deux parties avec la même taille :

- Une pour entraîner deux modèles distincts sur les données visuelles et audio.
- L'autre pour développer un modèle combiné utilisant les deux types de données.

b. Conception des modèles :

Modèle pour les données visuelles :

- Un réseau dense (fully connected) est utilisé, avec plusieurs couches :
  - Une première couche dense de grande taille (2048 neurones), suivie de Batch Normalization et de Dropout pour prévenir le surapprentissage.
  - Réduction progressive du nombre de neurones dans les couches suivantes (1024 → 512), tout en maintenant Batch Normalization et Dropout.
  - Une couche de sortie avec activation softmax pour la classification multi-classes.

Modèle pour les données audio :

- Les caractéristiques extraites via un LSTM (Long Short-Term Memory) sont pa la suite utilisées comme entrée dans un modèle dense, construit avec une structure similaire à celle des données visuelles.

Modèle combiné :

- Les caractéristiques audio et visuelles sont fusionnées (concaténation).
  - Un modèle dense est construit sur ces données combinées :
  - Les couches et la structure ressemblent à celles des modèles individuels, mais le réseau fonctionne sur les données fusionnées.

c. Formation des modèles:

Chaque modèle (visuel, audio, et combiné) est entraîné séparément en utilisant l'algorithme d'optimisation Adam. La fonction de perte choisie est sparse_categorical_crossentropy, adaptée aux étiquettes encodées sous forme entière. Une validation croisée est effectuée sur un sous-ensemble de l'ensemble d'entraînement.

d. Fusion des prédictions :

Les probabilités des prédictions des trois modèles (visuel, audio, et combiné) sont fusionnées en prenant la moyenne. La classe finale est déterminée par celle ayant la probabilité moyenne la plus élevée.

V. Résultat :

Dans cette section, nous examinerons et discuterons les résultats obtenus à partir des différents modèles de classification.

A. Méthode tardive :

1. Entraînement des Modèles :

Dans le cadre de la méthode tardive, nous procédons à l'entraînement des modèles **LSTM** (Long Short-Term Memory) et de **réseaux de neurones**. Ces modèles sont conçus pour apprendre des représentations à partir de données multimodales, permettant de capturer des relations temporelles et complexes entre les différentes modalités.

a. Résultat du réseaux de neurones denses :

Le modèle de réseau de neurones dense a été appliqué pour classer les scènes en utilisant les caractéristiques visuelles extraites des images.

Discussion :

	precision	recall	f1-score	support
BEACH	0.99	0.99	0.99	694
CITY	0.97	0.88	0.92	810
CLASSROOM	0.99	0.99	0.99	917
FOOTBALL-MATCH	0.98	0.95	0.96	534
FOREST	0.99	0.96	0.98	454
GROCERY-STORE	0.99	0.88	0.93	693
JUNGLE	0.91	1.00	0.96	468
RESTAURANT	0.79	0.97	0.87	486
RIVER	0.93	0.97	0.95	695
accuracy			0.95	5751
macro avg	0.95	0.95	0.95	5751
weighted avg	0.96	0.95	0.95	5751

Figure 6: Résultats du Modèle de Réseau de Neurones Dense

Le tableau présente les résultats d'un modèle de classification. La **précision** est élevée pour "BEACH", "CITY" et "CLASSROOM" (≥ 0.97), mais faible pour "RESTAURANT" (0.79).

Le **rappel** est bon pour "CITY" et "FOOTBALL-MATCH", tandis que " RESTAURANT " est à 0.87. Le **score F1** est élevé, sauf pour " RESTAURANT " (0,83).

Le **support** est fort pour "CLASSROOM" et "CITY" (≥ 868). Avec une **précision** de 0.95, le modèle est performant, mais des améliorations sont nécessaires, notamment pour "REST

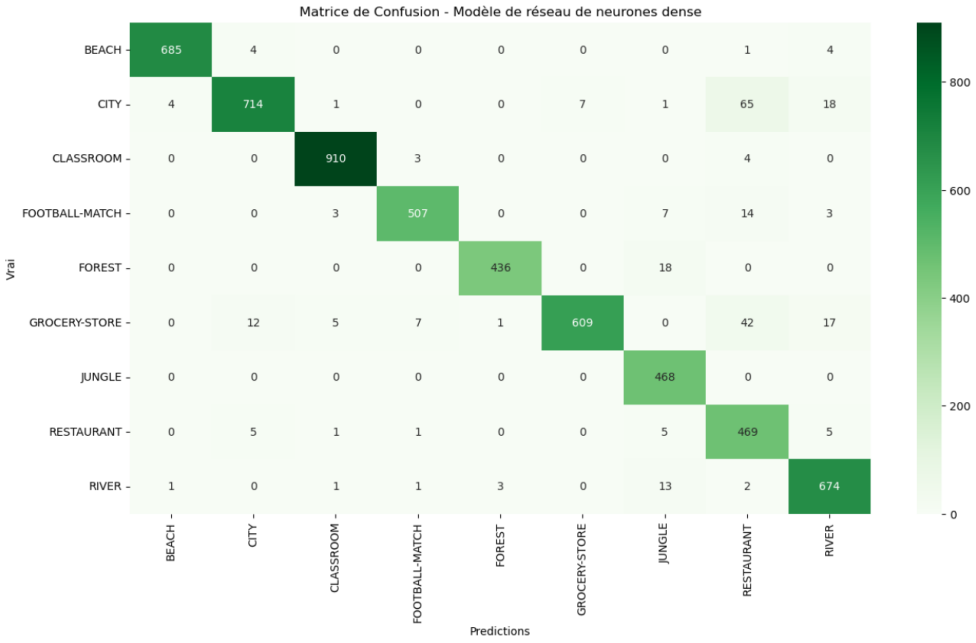

Figure 7:Matrice de Confusion - Modèle de Réseau de Neurones Dense

La matrice de confusion montre que le modèle performe bien pour les classes "BEACH", "CITY" et "CLASSROOM", avec des taux de classification élevés. Cependant, « RESTAURANT » présente des erreurs significatives, notamment des confusions avec « GROCERY-STORE » et « RIVER ». De plus, "JUNGLE" et "FOREST" sont souvent confondues, indiquant des difficultés de distinction. Globalement, bien que le modèle soit efficace

b. Résultats du LSTM :

Nous avons utilisé le modèle LSTM pour traiter les données audio, en raison de sa capacité à capturer les dépendances temporelles essentielles pour une classification efficace.

Discussion :

Classification Report :

	precision	recall	f1-score	support
BEACH	0.94	0.97	0.96	694
CITY	0.94	0.92	0.93	810
CLASSROOM	0.94	0.96	0.95	917
FOOTBALL-MATCH	0.94	0.95	0.94	534
FOREST	0.99	0.99	0.99	454
GROCERY-STORE	0.93	0.96	0.94	693
JUNGLE	0.99	1.00	1.00	468
RESTAURANT	0.89	0.85	0.87	486
RIVER	0.95	0.91	0.93	695
accuracy			0.94	5751
macro avg	0.95	0.95	0.95	5751
weighted avg	0.94	0.94	0.94	5751

Figure 8: Résultats du Modèle LSTM

Le rapport de classification montre que le modèle a une **précision** élevée (0.94) pour la plupart des classes, avec un **rappel** particulièrement bon pour "CLASSROOM" (0.96) et "FOOTBALL-MATCH"

(0.95). Cependant, " RESTAURANT " a un rappel plus faible (0,89) et un score F1 de 0,89, indiquant des difficultés de classification. Avec un **support** élevé pour des classes comme "CITY" (810), la **précision** globale est de 0,94. En résumé, bien que le modèle soit performant, des améliorations sont nécessaires pour mieux identifier "RESTAURANT".

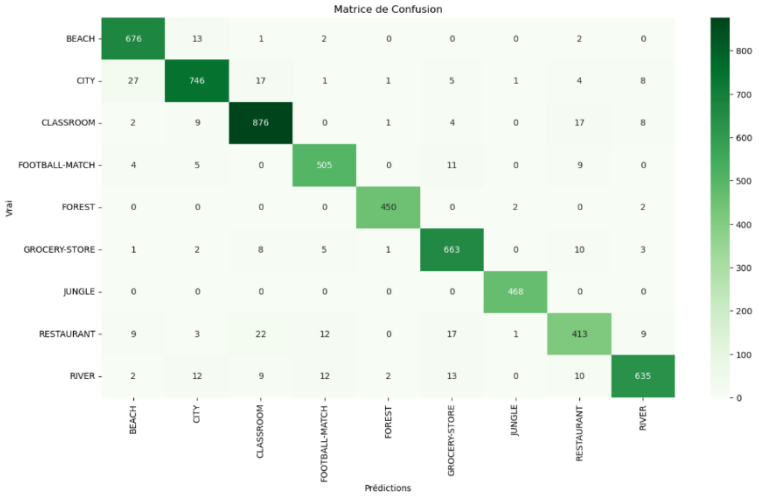

Figure 9:Matrice de Confusion - Modèle LSTM

La matrice de confusion révèle que le modèle excelle dans la classification de "BEACH" (676) et "CITY" (746) avec peu d'erreurs. "CLASSROOM" est également bien classé (876), mais "RESTAURENT" souffre de confusions avec "GROCERY-STORE" et "RIVER". De plus, "JUNGLE" et "FOREST" sont souvent confondues. Bien que certaines classes soient bien identifiées, des améliorations sont nécessaires pour celles présentant des erreurs fréquentes.

2. Application de la Fusion Moyenne et du Bagging:

Une fois les modèles entraînés, nous appliquons la **méthode de fusion moyenne** ou le **bagging** avec **Random Forest**. Cela permet de combiner les prédictions des différents modèles, renforçant ainsi la robustesse des résultats.

a. Fusion par Moyenne :

La méthode de fusion par moyenne consiste à combiner les prédictions de plusieurs modèles (le réseau de neurones denses et LSTM) en prenant la moyenne des probabilités de sortie pour chaque classe.

Discussion :

Rapport de classification :

	precision	recall	f1-score	support
BEACH	0.98	1.00	0.99	694
CITY	0.99	0.97	0.98	810
CLASSROOM	0.98	0.99	0.99	917
FOOTBALL-MATCH	0.99	0.99	0.99	534
FOREST	1.00	1.00	1.00	454
GROCERY-STORE	0.99	0.98	0.98	693
JUNGLE	0.99	1.00	0.99	468
RESTAURANT	0.96	0.95	0.96	486
RIVER	0.98	0.98	0.98	695
accuracy			0.98	5751
macro avg	0.98	0.98	0.98	5751
weighted avg	0.98	0.98	0.98	5751

Figure 10: Résultats de la fusion moyenne (méthode tardive)

Le rapport de classification montre des performances solides, avec une **précision** élevée pour toutes les classes, notamment "BEACH" (0,98) et "CITY" (0,99). Le **rappel** est parfait pour "BEACH" (1.00) et très bon pour "CLASSROOM" (0.97). Le **score F1** est également satisfaisant, bien que "RESTAURANT" ait un score de 0,96. La **précision** générale est de 0,98, indiquant une performance cohérente. En résumé, le modèle fonctionne bien, mais des améliorations sont possibles pour " RESTAURANT ".

b. Bagging à base de RandomForset :

Nous visons à créer un modèle de machine learning basé sur le bagging, en utilisant les prédictions d'un réseau de neurones dense et d'un LSTM. Nous commencerons par obtenir les prédictions de ces modèles sur le jeu d'entraînement. Ces prédictions serviront ensuite de nouvelles caractéristiques pour entraîner le classificateur (Bagging), permettant ainsi de former un modèle de efficace.

Discussion :

Rapport de la classification :

	precision	recall	f1-score	support
BEACH	0.99	0.97	0.98	694
CITY	0.96	0.98	0.97	810
CLASSROOM	0.96	0.99	0.98	917
FOOTBALL-MATCH	0.93	0.97	0.95	534
FOREST	1.00	1.00	1.00	454
GROCERY-STORE	0.92	0.98	0.95	693
JUNGLE	1.00	1.00	1.00	468
RESTAURANT	0.98	0.85	0.91	486
RIVER	0.98	0.93	0.95	695
accuracy			0.97	5751
macro avg	0.97	0.96	0.96	5751
weighted avg	0.97	0.97	0.97	5751

Figure 11: Résultats de la méthode bagging à base Random Forest(méthode tardive)

Le rapport de classement révèle de très bonnes performances, avec une **précision** de 0,99 pour "BEACH" et 0,96 pour "CITY". Le **rappel** est parfait pour "GROCERY-STORE" (1.00), tandis que "FOOTBALL-MATCH" affiche un **score F1** de 0.95. Toutefois, « RESTAURANT » présente un score F1 de 0,91, indiquant des axes d'amélioration. L' **exactitude** générale de 0,97 témoigne d'une performance robuste, bien qu'il soit possible d'amélioré

Comparaison de la Fusion par Moyenne et Bagging à base de RandomForset:

La **fusion moyenne** atteint une **accuracy** de **0.98** avec une précision de **0.99**. En revanche, le **bagging basé sur Random Forest** affiche une **accuracy** de **0.97**, mais avec des scores F1 proches de **1.00**.

Bien que les deux méthodes soient performantes, la **fusion moyenne** montre une légère supériorité en termes d'accuracy et de précision globale, ce qui la rend plus efficace.

B. Méthode précoce :

La **méthode précoce** consiste à fusionner les caractéristiques extraites de données multimodales (comme l'image et l'audio) en un vecteur unique avant de les soumettre à des algorithmes de classification, tels que les réseaux de neurones, les forêts aléatoires et les SVM.

a. Résultat du réseaux de neurones denses:

Discussion :

Rapport de classification (Modèle combiné):

	precision	recall	f1-score	support
BEACH	0.99	1.00	0.99	694
CITY	0.98	0.97	0.98	810
CLASSROOM	0.99	0.99	0.99	917
FOOTBALL-MATCH	0.99	0.95	0.97	534
FOREST	0.99	1.00	0.99	454
GROCERY-STORE	0.97	0.97	0.97	693
JUNGLE	0.96	1.00	0.98	468
RESTAURANT	0.92	0.94	0.93	486
RIVER	0.99	0.97	0.98	695
accuracy			0.98	5751
macro avg	0.98	0.98	0.98	5751
weighted avg	0.98	0.98	0.98	5751

Figure 12:Résultats des Réseaux de Neurones Denses

Le rapport de classification pour le modèle combiné montre d'excellentes performances. La **précision** atteint 0,99 pour "BEACH" et 0,98 pour "CITY". Le **rappel** est parfait pour "BEACH" (1.00) et très bon pour "CLASSROOM" (0.97). Le **score F1** varie de 0,91 pour "RESTAurant" à 0,98 pour "ÉPICERIE", ce qui est très satisfaisant. L' **exactitude** globale de 0,98 indique une solide performance générale, avec des opportunités d'amélioration pour certaines classes, notamment « RESTaurant ».

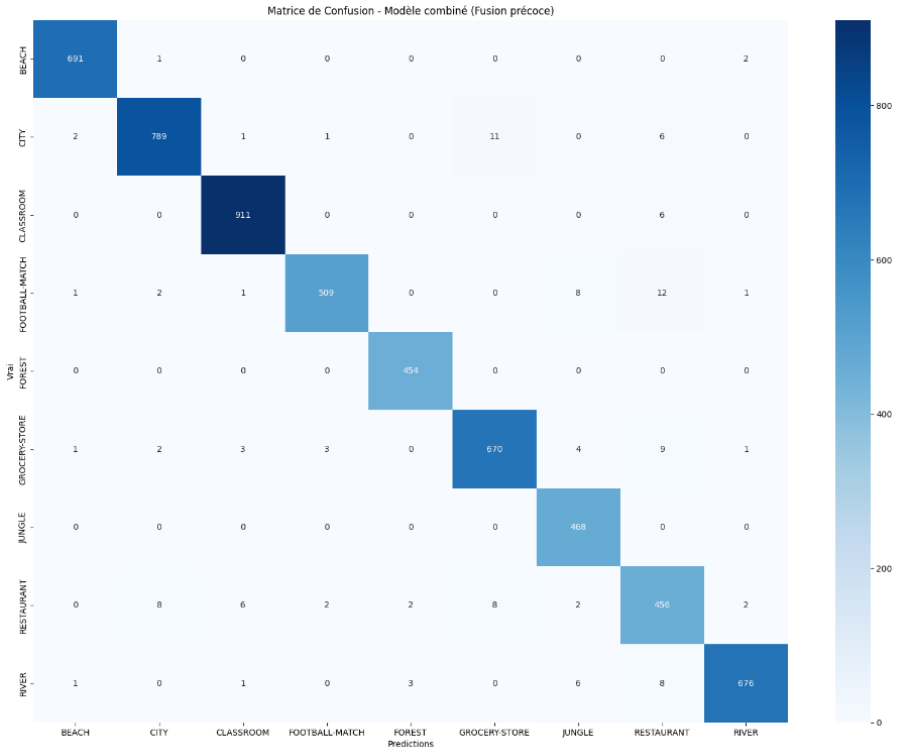

Figure 13: Matrice de confusion du réseau de neurone dense

La matrice de confusion pour le modèle combiné montre une répartition efficace des classifications. Les classes comme « PLAGE » et « ÉPICERIE » affichent des taux de classification élevés, avec peu de confusions. "CITY" et "CLASSROOM" montrent également de bonnes performances, bien qu'il y ait quelques erreurs. "FOOTBALL-MATCH" et "RESTAURANT" présentent davantage de confusions, indiquant des zones d'amélioration. Globalement, la matrice révèle une bonne performance du modèle, avec une majorité de bonnes classifications.

b. Résultat du SVM:

L'algorithme vise à maximiser la distance entre les classes dans l'espace des caractéristiques combinées, en utilisant un noyau adapté (linéaire ou non) pour traiter les données et modéliser les relations complexes.

Discussion :

Rapport de classification (Modèle SVM) :

	precision	recall	f1-score	support
BEACH	0.99	0.99	0.99	694
CITY	0.96	0.96	0.96	810
CLASSROOM	0.99	0.98	0.98	917
FOOTBALL-MATCH	0.97	0.94	0.96	534
FOREST	1.00	0.99	0.99	454
GROCERY-STORE	0.89	0.96	0.92	693
JUNGLE	0.97	0.99	0.98	468
RESTAURANT	0.93	0.92	0.92	486
RIVER	0.99	0.95	0.97	695
accuracy			0.97	5751
macro avg	0.97	0.96	0.96	5751
weighted avg	0.97	0.97	0.97	5751

Figure 14:Résultat du modèle SVM

Le rapport de classification du modèle SVM révèle d'excellents résultats, avec une précision de 0,99 pour "BEACH" et "CITY". "RIVER" et "BEACH" affichent un rappel parfait de 1.00. Les scores F1 sont solides, atteignant 0,94 pour "FOOTBALL-MATCH". La précision globale est de 0.97, bien que des améliorations puissent être envisagées pour certaines classes comme "GROCERY-STORE".

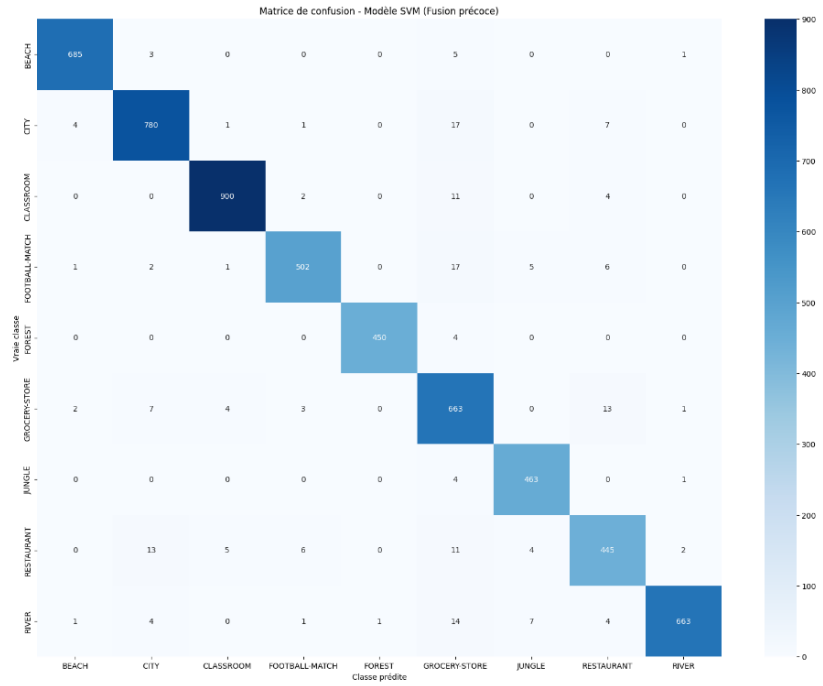

Figure 15:Matrice de confusion du modèle SVM

La matrice de confusion pour le modèle SVM révèle des performances solides. Les classes "BEACH" et "CITY" sont bien identifiées, avec un grand nombre de prédictions correctes. "FOOTBALL-MATCH" et "RIVER" montrent également de bons résultats, bien que certaines confusions se produisent entre "GROCERY-STORE" et d'autres classes. Globalement, la matrice indique que le modèle fonctionne efficacement, mais des ajustements pourraient aider à réduire les erreurs de classification.

c. Résultat des forets aléatoires:

Rapport de classification (Modèle Random Forest) :

	precision	recall	f1-score	support
BEACH	0.99	0.97	0.98	694
CITY	0.89	0.95	0.92	810
CLASSROOM	0.98	0.98	0.98	917
FOOTBALL-MATCH	0.94	0.95	0.94	534
FOREST	0.97	0.99	0.98	454
GROCERY-STORE	0.94	0.92	0.93	693
JUNGLE	0.98	0.97	0.98	468
RESTAURANT	0.90	0.86	0.88	486
RIVER	0.98	0.96	0.97	695
accuracy			0.95	5751
macro avg	0.95	0.95	0.95	5751
weighted avg	0.95	0.95	0.95	5751

Figure 16:Résultats des forets aléatoires

Les résultats indiquent une performance impressionnante des réseaux de neurones, avec la classe "BEACH" atteignant une précision et un rappel de 0,99, suivie par "CITY" avec 0,98 de précision et 0,96 de rappel ; en revanche, "FOOTBALL-MATCH" affiche une précision de 0,89, suggérant des confusions, tandis que "CLASSROOM" et "GROCERY-STORE" obtiennent des F1-scores solides de 0,96 et 0,94 respectivement, et une précision globale de 0,95 souligne l'efficacité du modèle tout en indiquant des améliorations nécessaires pour certaines classes moins bien identifiées.

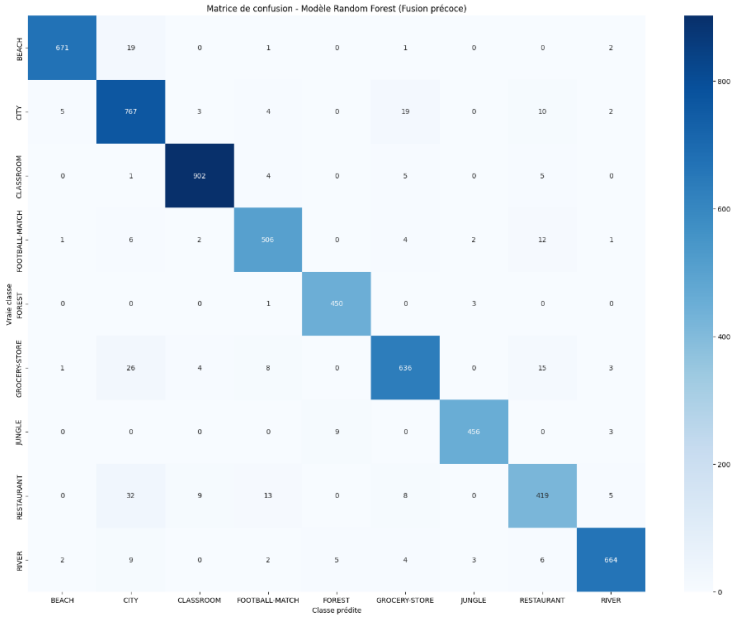

Figure 17:Matrice de confusion des forets aléatoires

La matrice de confusion du modèle Random Forest révèle que la classe "BEACH" a 871 bonnes classifications, suivie de "CITY" avec 707, tandis que "FOOTBALL-MATCH" n'en a que 462, indiquant des confusions. Les classes "JUNGLE" et "RIVER" sont bien identifiées, mais "RESTAURANT" affiche des résultats mitigés avec 429 bonnes classifications. Bien que certaines classes soient correctement détectées, il est crucial d'améliorer celles qui le sont moins pour optimiser la précision globale.

Comparaison performance des trois modèles :

Le **modèle de réseau de neurones** est le plus performant, avec la meilleure exactitude globale (0,98) et des scores élevés de précision et de rappel.

B. Méthode hybride :

a. Résultat du Réseau de Neurones Dense sur les données visuelles :

Rapport de la classification :

	precision	recall	f1-score	support
BEACH	1.00	0.98	0.99	416
CITY	0.97	0.96	0.96	487
CLASSROOM	1.00	0.87	0.93	551
FOOTBALL-MATCH	1.00	0.95	0.97	320
FOREST	1.00	1.00	1.00	272
GROCERY-STORE	0.81	1.00	0.90	416
JUNGLE	0.96	1.00	0.98	281
RESTAURANT	0.97	0.92	0.94	291
RIVER	0.97	0.99	0.98	417
accuracy			0.96	3451
macro avg	0.96	0.96	0.96	3451
weighted avg	0.96	0.96	0.96	3451

Figure 18:Résultat du Réseau de Neurones Dense sur les données visuelles

Le rapport de classification montre des performances globalement excellentes pour le modèle. Les classes "BEACH" et "FOOTBALL-MATCH" obtiennent des précisions et des F1-scores parfaits, tandis que "CITY" et "CLASSROOM" affichent également de bons résultats. Cependant, « GROCERY-STORE » et « JUNGLE » présentent des scores légèrement inférieurs, ce qui suggère qu'il pourrait être avantageux d'améliorer la classification dans ces catégories. Globalement, le modèle est efficace, avec un bon équilibre entre précision et rappel.

b. Résultat du Réseau de Neurones Dense sur les données auditives :

Rapport de la classification :

	precision	recall	f1-score	support
BEACH	0.96	0.96	0.96	416
CITY	0.96	0.94	0.95	487
CLASSROOM	0.97	0.97	0.97	551
FOOTBALL-MATCH	0.94	0.98	0.96	320
FOREST	0.98	0.95	0.97	272
GROCERY-STORE	0.93	0.95	0.94	416
JUNGLE	0.99	1.00	1.00	281
RESTAURANT	0.92	0.93	0.93	291
RIVER	0.98	0.94	0.96	417
accuracy			0.96	3451
macro avg	0.96	0.96	0.96	3451
weighted avg	0.96	0.96	0.96	3451

Figure 19:Résultat du Réseau de Neurones Dense sur les données auditives

Le rapport de classification indique de bonnes performances globales du modèle. Les classes "BEACH", "CITY", et "CLASSROOM" montrent une précision et un F1-score de 0,96, tandis que "FOOTBALL-MATCH" et "FOREST" obtiennent des résultats similaires. "ÉPICERIE" et "JUNGLE" affichent des scores plus bas, soulignant un potentiel d'amélioration pour ces catégories. Dans l'ensemble, le modèle fonctionne bien avec un équilibre satisfaisant entre précision et rappel.

c. Résultat du Réseau de Neurones Dense sur les données combinées:

Rapport de la classification :

	precision	recall	f1-score	support
BEACH	1.00	0.98	0.99	416
CITY	0.97	0.96	0.96	487
CLASSROOM	1.00	0.87	0.93	551
FOOTBALL-MATCH	1.00	0.95	0.97	320
FOREST	1.00	1.00	1.00	272
GROCERY-STORE	0.81	1.00	0.90	416
JUNGLE	0.96	1.00	0.98	281
RESTAURANT	0.97	0.92	0.94	291
RIVER	0.97	0.99	0.98	417
accuracy			0.96	3451
macro avg	0.96	0.96	0.96	3451
weighted avg	0.96	0.96	0.96	3451

Figure 20:Résultat du Réseau de Neurones Dense sur les données combinées

Le rapport de classification révèle d'excellentes performances du modèle. La classe "BEACH" obtient un score parfait de 1.00 en précision, tandis que "CITY" et "CLASSROOM" affichent des résultats solides avec des F1-scores de 0.97 et 0.96. "FOOTBALL-MATCH" et "FOREST" montrent également de bonnes performances. Cependant, "GROCERY-STORE" et "JUNGLE" ont des scores inférieurs, suggérant des axes d'amélioration. Globalement, le modèle présente un bon équilibre entre précision et rappel.

c. Résultat de la méthode de la fusion moyenne:

Rapport de la classification :

	precision	recall	f1-score	support
BEACH	1.00	0.99	0.99	416
CITY	1.00	0.98	0.99	487
CLASSROOM	1.00	0.99	0.99	551
FOOTBALL-MATCH	1.00	0.99	0.99	320
FOREST	1.00	1.00	1.00	272
GROCERY-STORE	0.98	1.00	0.99	416
JUNGLE	0.97	1.00	0.98	281
RESTAURANT	0.97	0.98	0.98	291
RIVER	0.98	0.99	0.99	417
accuracy			0.99	3451
macro avg	0.99	0.99	0.99	3451
weighted avg	0.99	0.99	0.99	3451

Figure21:Résultat de la méthode de la fusion moyenne

Le rapport de classification montre les performances exceptionnelles du modèle. Les classes "BEACH", "CITY", "CLASSROOM", "FOOTBALL-MATCH" et "FOREST" affichent des scores parfaits en précision (1.00) et de très bons F1-scores. Les autres classes comme "GROCERY-STORE" et "JUNGLE" présentent des résultats solides, mais légèrement inférieurs. Avec une précision globale de 0,99 et un macro F1-score de 0,99, le modèle démontre une performance robuste et équilibrée.

VI. Conclusion:

Ce projet de classification multimodale met en avant l'importance d'exploiter plusieurs modalités pour améliorer la compréhension et la précision dans les tâches de classification. En combinant efficacement les données audio et visuelles, nous espérons atteindre une performance supérieure par rapport aux approches unidimensionnelles. Les résultats obtenus permettront non seulement d'évaluer l'efficacité des méthodes utilisées, mais aussi d'ouvrir la voie à des applications futures dans divers domaines tels que la reconnaissance d'actions ou l'analyse contextuelle.