

# RAPPORT D'EVALUATION

Système de classification de commandes vocales  
pour drone

*Basé sur le corpus VoiceStick (Henry et al., 2025)*

L'évaluation des modèles SVM et MLP repose sur une double analyse : une validation croisée avec des locuteurs indépendants à cinq plis (5-fold), ainsi qu'une évaluation finale sur un jeu de test indépendant composé de 1408 segments issus de neuf locuteurs non vus pendant l'entraînement. Cette méthodologie permet d'estimer à la fois la robustesse des modèles face à la variabilité interlocuteur et leur capacité réelle de généralisation.

Les métriques utilisées sont l'exactitude, le F1-score macro, le F1-score pondéré ainsi que les F1-scores par classe. Le F1-macro constitue la métrique principale, car il accorde le même poids à chaque classe indépendamment de leur fréquence. Dans le cadre d'un système de commandes vocales pour drone, il est essentiel que l'ensemble des commandes fonctionne correctement, y compris les classes minoritaires.

## Validation croisée

En validation croisée, le SVM présente des performances stables. Les résultats agrégés sur les cinq folds sont présents dans la *Table 2*.

L'écart-type faible du F1-macro indique une bonne robustesse du modèle SVM face aux variations de composition des locuteurs entre les folds. L'analyse par classe en validation croisée met en évidence les performances suivantes :

| Classe          | Précision | Rappel | F1-score | Support |
|-----------------|-----------|--------|----------|---------|
| <i>backward</i> | 0.65      | 0.82   | 0.72     | 152     |
| <i>down</i>     | 0.73      | 0.85   | 0.78     | 384     |
| <i>forward</i>  | 0.86      | 0.89   | 0.87     | 1229    |
| <i>left</i>     | 0.80      | 0.69   | 0.74     | 322     |
| <i>none</i>     | 0.95      | 0.88   | 0.91     | 4265    |
| <i>right</i>    | 0.63      | 0.73   | 0.68     | 251     |
| <i>up</i>       | 0.74      | 0.89   | 0.80     | 342     |
| <i>yawleft</i>  | 0.71      | 0.76   | 0.74     | 239     |
| <i>yawright</i> | 0.60      | 0.68   | 0.64     | 355     |

*Table 1. Performances par classe du SVM en validation croisée*

On observe que les classes « *none* » et « *forward* » sont les mieux reconnues, ce qui s'explique par leur support élevé et leur signature acoustique relativement distincte. Les principales difficultés concernent les commandes de translation et de rotation, notamment « *yawright* », qui présente le F1 le plus faible. Les confusions apparaissent principalement entre « *left* » et « *yawleft* », ainsi qu'entre « *right* » et « *yawright* » et peuvent s'expliquer par un biais contextuel lié au protocole de collecte des données. Dans le cadre de la tâche expérimentale, les pilotes n'ont pas toujours interprété ces commandes de manière strictement uniforme. Selon la situation, ils ont privilégié soit une translation, soit une rotation, afin d'optimiser la réalisation de l'objectif et de réduire la confusion dans l'interaction avec l'utilisateur dictant les commandes. Cette variabilité d'interprétation a

probablement introduit une ambiguïté supplémentaire dans les données, rendant la séparation acoustique de ces classes plus complexe pour les modèles, ainsi que pour l'annotation préliminaire.

Le MLP, en validation croisée, présente un comportement plus instable (*Table 2*). Le F1-macro de MLP est fortement dégradée par l'effondrement de trois folds sur cinq.

|            | <b>Fold 1</b> | <b>Fold 2</b> | <b>Fold 3</b> | <b>Fold 4</b> | <b>Fold 5</b> |
|------------|---------------|---------------|---------------|---------------|---------------|
| <b>SVM</b> | 0.773         | 0.725         | 0.786         | 0.747         | 0.765         |
| <b>MLP</b> | 0.709         | 0.650         | <b>0.020</b>  | <b>0.024</b>  | <b>0.024</b>  |

*Table 2. Comparaison des F1-macro par fold*

Ces résultats indiquent une instabilité d'entraînement sur certains sous-ensembles de locuteurs. Les métriques agrégées en validation croisée ne reflètent donc pas fidèlement la capacité intrinsèque du modèle MLP, mais plutôt une sensibilité à certaines configurations de données.

## Évaluation sur le jeu de test indépendant

| <b>Modèle</b> | <b>Exactitude</b> | <b>F1-macro</b> | <b>F1-pondéré</b> |
|---------------|-------------------|-----------------|-------------------|
| SVM           | 0.836             | 0.735           | 0.843             |
| MLP           | 0.866             | 0.774           | 0.868             |

*Table 3. Performances globales des modèles sur le jeu de test indépendant*

L'évaluation finale sur le jeu de test indépendant fournit une estimation plus fiable des performances réelles des modèles. Le MLP surpasse le SVM sur l'ensemble des métriques globales, avec un gain de 4 points en F1-macro par rapport au SVM.

| <b>Classe</b>   | <b>F1 SVM</b> | <b>F1 MLP</b> |
|-----------------|---------------|---------------|
| <i>backward</i> | 0.648         | <b>0.808</b>  |
| <i>down</i>     | <b>0.800</b>  | 0.738         |
| <i>forward</i>  | 0.850         | <b>0.889</b>  |
| <i>left</i>     | 0.637         | 0.634         |
| <i>none</i>     | 0.903         | <b>0.923</b>  |
| <i>right</i>    | 0.683         | <b>0.789</b>  |
| <i>up</i>       | 0.828         | <b>0.857</b>  |
| <i>yawleft</i>  | <b>0.661</b>  | 0.623         |
| <i>yawright</i> | 0.608         | <b>0.706</b>  |

*Table 4. Comparaison des F1-scores par classe par modèle sur le jeu de test*

## Observations clés

L'analyse détaillée des F1-scores par classe (*Table 4*) permet de mieux comprendre les forces et les limites respectives des deux modèles. Le MLP obtient de meilleures performances que le SVM sur 6 des 9 classes, à savoir « backward », « forward », « none », « right », « up » et « yawright ». L'amélioration la plus marquée concerne la classe « backward », avec un F1 de 0.81 pour le MLP contre 0.65 pour le SVM, soit un gain de 16 points. Des progrès notables sont également observés pour « right » (0.79 vs 0.68, +11 points) et « yawright » (0.71 vs 0.61, +10 points). Le SVM conserve toutefois un avantage sur certaines classes. Il est supérieur pour « down » (0.80 vs 0.74, +6 points) et pour « yawleft » (0.66 vs 0.62, +4 points), et présente une légère avance sur « left » (0.64 contre 0.63).

Les classes les plus difficiles pour les deux modèles restent « left », « yawleft », « right » et « yawright », avec des F1 compris entre 0.61 et 0.66 environ. Ces difficultés s'expliquent probablement par leur proximité phonétique et par un support plus limité dans le jeu de test. À l'inverse, les classes les mieux reconnues sont « none » (F1 supérieur à 0.90) et « forward » (F1 supérieur à 0.85). Elles bénéficient à la fois d'un nombre d'exemples plus important et d'un profil acoustique plus distinctif, ce qui facilite leur séparation dans l'espace des embeddings.

Les matrices de confusion (*Figure 1*) confirment que les erreurs ne sont pas distribuées de manière aléatoire. On observe une diagonale dominante pour les deux modèles, ce qui indique une reconnaissance globalement correcte des classes. Les erreurs se concentrent principalement entre classes sémantiquement proches, en particulier entre « left » contre « yawleft », ainsi que « right » contre « yawright ». On constate également des confusions ponctuelles entre certaines commandes directionnelles et la classe « none », mais les dernières restent limitées. Cette organisation structurée des erreurs suggère que les limites du système sont liées au biais contextuel et directionnel des commandes plutôt qu'à une incapacité générale du modèle à apprendre la tâche.

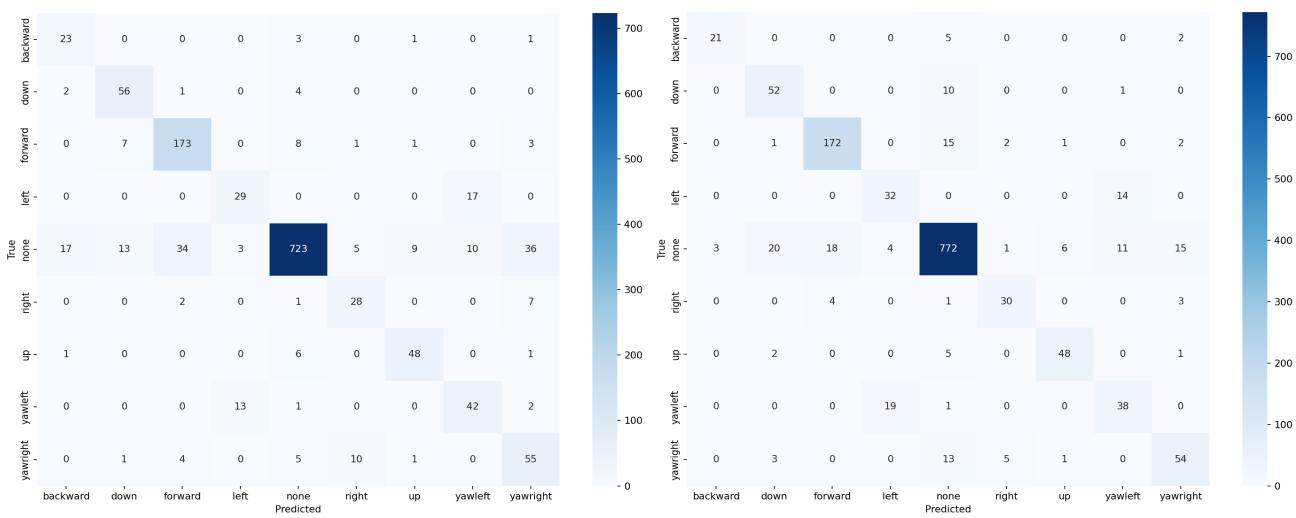


Figure 1. Matrices de confusion de la partition test : (gauche) modèle SVM, (droite) modèle MLP.

## Conclusion

Le SVM se distingue par sa stabilité en validation croisée et par une robustesse constante d'après les folds. Le MLP, bien que plus instable en validation croisée, obtient de meilleures performances absolues sur le jeu de test indépendant. Il dépasse l'objectif indicatif fixé pour le MVP en F1-macro et performe mieux sur plusieurs classes.

Ainsi, le SVM apparaît comme le modèle le plus fiable en matière de stabilité, tandis que le MLP constitue la solution la plus performante en termes de capacité prédictive globale pour le système final de classification des commandes.