

UQAC

Université du Québec
à Chicoutimi

**CREATION D'UN DATASET POUR L'ANALYSE COMPARATIVE DU NIVEAU DE
DIFFICULTE DANS LES JEUX VIDEO DE TYPE PLATFORMER**

PAR YAO JEAN-EUDES ADJANOHOUN

**MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI COMME
EXIGENCE PARTIELLE EN VUE DE L'OBTENTION DU GRADE DE MAÎTRE ÈS
SCIENCES EN INFORMATIQUE**

QUÉBEC, CANADA

© YAO JEAN-EUDES ADJANOHOUN, 2025

RÉSUMÉ

Les ensembles de données normalisés (*datasets*) sont fondamentaux pour la recherche scientifique. Alors que des domaines comme le traitement du langage naturel et la vision par ordinateur disposent de jeux de données largement acceptés qui favorisent les progrès, la recherche sur les jeux vidéo manque encore de telles ressources, en particulier pour l'étude de la difficulté des niveaux. Les ressources existantes reposent sur des ensembles de données isolés et personnalisés, ce qui limite les comparaisons entre études et freine le développement de modèles généralisables. Aussi, il apparaît que la majorité des *datasets* disponibles portent sur des données joueurs (position de la tête, réactions, etc.) et non sur des données objectives de performance (nombre de morts, nombre de sauts, etc.).

Pour combler cette lacune, nous présentons dans ce mémoire LADDER, un nouvel ensemble de données spécialement conçu pour analyser et évaluer la difficulté des niveaux dans les jeux vidéo. Contrairement aux ensembles de données précédents, qui se concentrent principalement sur les données physiologiques et comportementales des joueurs, LADDER intègre des mesures de performance objectives (par exemple, la santé perdue, le nombre de tentatives avant la réussite, le nombre de morts), des caractéristiques de niveau (par exemple, le nombre de zones dangereuses, la disposition des objets), ainsi que des évaluations de la difficulté perçue à travers plusieurs jeux de type *platformer*.

Cet ensemble de données permet aux chercheurs d'établir des références, de renforcer la collaboration interdisciplinaire et d'améliorer la reproductibilité des études. LADDER fournit une base normalisée pour explorer la relation entre les éléments de conception du jeu et l'expérience des joueurs. En facilitant l'évaluation de la difficulté et l'équilibrage des niveaux, il soutient les avancées en conception de jeux, en modélisation des joueurs et en systèmes de jeu adaptatifs. Ce mémoire, présenté sous forme d'article scientifique, propose une introduction à la problématique, une revue de la littérature, une présentation de la contribution scientifique sous forme d'un article publié dans le cadre des actes de la conférence *IEEE Gaming, Entertainment, and Media conference*, ainsi qu'un bilan des réalisations et une présentation des travaux futurs.

TABLE DES MATIÈRES

Table des matières	ii
I Introduction	1
II Revue de la littérature scientifique	6
2.1 Le Platformer Experience Dataset (PED)	8
2.2 Le GamingVideoSET dataset	11
2.3 Le Toadstool dataset	13
2.4 Le PlayMyData dataset	15
2.5 STARDATA : un dataset pour l'IA dans le jeu Starcraft	17
2.6 Le dataset Again	19
2.7 La dataset AMuCS	21
2.8 Dataset sur Counter-Strike : Global Offensive	24
2.9 Conclusion	26
III Ladder : un nouveau dataset pour l'évaluation de la difficulté dans les jeux de plateforme	28
3.1 Article scientifique présenté à IEEE-GEM 2025	30
IV Bilan sur les travaux et sur la construction de LADDER	37
4.1 Sélection des sources de données	39

4.2	Protocole expérimental	40
4.3	Particularités technique	40
V	Conclusion	42
	Bibliographie	45

CHAPITRE I

INTRODUCTION

Dans le domaine des sciences, les ensembles de données standardisés sont devenus un pilier essentiel de la recherche moderne, favorisant la reproductibilité, la comparabilité et l'innovation dans un large éventail de disciplines [24]. Ils offrent un langage commun et une structure partagée qui permettent aux chercheurs de valider des résultats, de tester de nouvelles hypothèses et de développer des méthodologies de référence sur des bases cohérentes et transparentes. En garantissant que les expériences puissent être reproduites dans différents contextes et par des équipes indépendantes, ces ensembles de données constituent une pierre angulaire de la science ouverte et cumulative.

Dans des domaines comme le traitement automatique du langage naturel ou la vision artificielle, la disponibilité de jeux de données de référence largement reconnus, tels qu'ImageNet¹ ou COCO², a permis des avancées sans précédent en établissant un terrain d'expérimentation commun pour l'évaluation et la comparaison des modèles [4]. Ces ressources partagées ont non seulement accéléré les progrès techniques, mais elles ont aussi favorisé l'émergence de véritables communautés de recherche unifiées autour d'objectifs méthodologiques communs.

1. <https://image-net.org/>

2. <https://cocodataset.org/>

Elles ont démontré qu'un corpus bien structuré peut transformer un domaine entier, en rendant les résultats comparables, mesurables et reproductibles.

Dans le contexte de la recherche sur le jeu vidéo [22], cependant, l'absence d'ensembles de données standardisés demeure un obstacle critique à la progression du domaine [25]. Les études portant sur la difficulté des jeux, le comportement des joueurs ou la conception des niveaux reposent encore trop souvent sur des ensembles de données isolés, conçus ad hoc pour des expériences spécifiques. Cette fragmentation méthodologique rend difficile la comparaison entre études, empêche la généralisation des résultats et limite la portée des modèles analytiques [16]. En l'absence d'un socle commun, chaque chercheur doit repartir de zéro, ce qui freine considérablement la reproductibilité et la cumulativité des connaissances.

Un ensemble de données standardisé portant sur la difficulté des niveaux, par exemple, permettrait aux chercheurs de :

- **Établir des points de référence communs** : en fournissant un banc d'essai partagé, un tel ensemble de données permettrait la comparaison rigoureuse d'algorithmes et de modèles d'évaluation de la difficulté, facilitant l'identification des approches les plus efficaces et robustes.
- **Favoriser la collaboration interdisciplinaire** : un corpus ouvert et bien documenté rassemblerait des chercheurs issus de domaines variés — intelligence artificielle, psychologie cognitive, ergonomie, design de jeu — pour croiser les perspectives et enrichir les analyses.
- **Promouvoir la transparence scientifique** : l'accès libre aux données rend les études plus faciles à reproduire, renforce la fiabilité des résultats et permet de mieux identifier les biais méthodologiques ou contextuels des expériences.

La création d'un tel ensemble de données dédié à la difficulté des niveaux comblerait donc un vide important dans la recherche sur les jeux vidéo. Elle ouvrirait la voie à la standardisation de méthodologies comparables à celles qu'ImageNet a instaurées pour la vision par ordinateur [4]. Cette initiative représente une étape décisive vers une compréhension plus fine de la manière dont les joueurs interagissent avec les environnements virtuels et de la façon dont la conception des niveaux influence l'engagement, la performance et la satisfaction des joueurs.

Dans ce mémoire, présenté sous forme d'article scientifique (voir la référence à notre contribution originale [2]), notre objectif est de réduire cette lacune en fournissant à la communauté scientifique un outil permettant d'explorer et d'affiner les concepts de difficulté de manière reproductible, évolutive et scientifiquement rigoureuse. Plus précisément, nous concentrons notre analyse sur la difficulté des niveaux individuels au sein des jeux vidéo. Le niveau constitue en effet l'unité fondamentale de l'expérience ludique : il structure la progression du joueur, module la tension dramatique et détermine en grande partie le plaisir et la motivation à poursuivre le jeu [16].

Malgré cette importance centrale, la notion de *difficulté d'un niveau* demeure largement subjective, dépendant de l'interaction complexe entre la conception du niveau et les compétences individuelles du joueur [26]. Quantifier cette difficulté est pourtant essentiel, non seulement pour concevoir des jeux équilibrés et engageants, mais aussi pour faire progresser la recherche en conception de jeux, en modélisation des comportements de joueurs et en systèmes de jeu adaptatifs [10]. Cependant, l'absence d'ensembles de données standardisés spécifiquement conçus pour évaluer la difficulté des niveaux représente un défi majeur, tant pour les chercheurs que pour les praticiens du domaine.

Les ensembles de données existants [13, 27, 31, 5, 30, 10] se concentrent principalement sur des mesures physiologiques ou comportementales recueillies pendant le jeu, telles que les mouvements oculaires, la pression sanguine, ou les niveaux de stress. Ces approches apportent des informations précieuses sur la réaction du joueur, mais elles négligent souvent les dimensions structurelles et objectives du niveau lui-même.

Un ensemble de données véritablement complet devrait donc intégrer, de manière systématique et normalisée :

- Des **indicateurs de performance objectifs** (par exemple, le pourcentage de santé perdu, le nombre de tentatives avant la réussite, le nombre de projectiles utilisés);
- Des **caractéristiques structurelles du niveau** (par exemple, la densité des zones dangereuses, la disposition des obstacles et des objets interactifs);
- Des **évaluations de la difficulté perçue** par les joueurs, permettant de relier les mesures objectives à l'expérience subjective vécue.

Pour répondre à ce besoin, nous présentons dans ce mémoire le nouveau *dataset* intitulé **LADDER**, qui correspond à un nouvel ensemble de données spécifiquement conçu pour l'analyse et l'évaluation de la difficulté des niveaux dans les jeux de plateformes. Cet ensemble de données offre une base solide pour l'étude des relations entre les éléments de conception du jeu et les comportements des joueurs, ouvrant la voie à de nouveaux outils pour l'évaluation de la difficulté et l'équilibrage des niveaux.

LADDER est librement accessible en ligne³ et destiné à l'ensemble de la communauté scientifique. Dans ce mémoire, nous commençons par passer en revue les ensembles de jeux de données existants dans la littérature. Ensuite nous décrivons, par la présentation de notre article

3. <https://github.com/Iannyck/LADDER>

scientifique publié dans le cadre des actes de la conférence *IEEE Gaming, Entertainment, and Media conference*, la méthodologie utilisée pour la construction de **LADDER** et pour l'extraction systématique des données en provenances de sessions de jeux réels. Cet article présente la contribution scientifique en soulignant les principales caractéristiques et en démontrant son potentiel à travers une série d'analyses. En offrant une ressource ouverte et standardisée, nous visons à soutenir la recherche et l'innovation dans le domaine du jeu vidéo, tout en contribuant à la création d'expériences ludiques plus engageantes, équilibrées et accessibles pour les joueurs de tous niveaux de compétence.

Le reste de ce mémoire est organisé de la manière suivante. Le chapitre 2 présentera une revue des jeux de données existants dans la littérature, en décrivant leurs forces et leurs faiblesses en lien avec notre objectif d'analyse de la difficulté dans les jeux. Par la suite, le chapitre 3 présentera la principale contribution du mémoire, notre article scientifique, qui décrit la méthodologie de construction du *dataset*, le processus d'extraction des données à partir de divers sites publics et de sessions de jeu, les jeux traités, ainsi que l'organisation de notre *dataset*. Le chapitre 4 présentera un bilan en français portant sur cette contribution et présentera également les travaux supplémentaires ayant été effectués après la parution de l'article. Enfin, le chapitre 5 présentera la conclusion du mémoire ainsi que les perspectives de travaux futurs.

CHAPITRE II

REVUE DE LA LITTÉRATURE SCIENTIFIQUE

Dans la littérature scientifique, très peu de jeux de données standardisés ont été proposés pour étudier la difficulté dans les jeux vidéo [13, 27, 31, 5, 30, 10]. La majorité des jeux de données existants se concentrent principalement sur des éléments centrés sur le joueur (*player centric*) collectés lors de sessions de jeu [13, 26, 30], tels que la pression sanguine, la position du regard, des mesures physiologiques variées ou encore des indicateurs comportementaux. Ces approches, bien que riches et pertinentes pour comprendre la réaction du joueur face aux stimuli du jeu, laissent de côté une dimension tout aussi fondamentale : la difficulté intrinsèque du jeu. Cette orientation met en lumière un besoin clair et encore peu comblé dans la recherche actuelle, soit la création de jeux de données mettant l'accent sur les éléments centrés sur le jeu, capables de mesurer de manière systématique et reproductible la complexité intrinsèque des défis proposés par le jeu lui-même.

La difficulté intrinsèque, ou difficulté centrée sur le jeu (*game centric*), renvoie aux caractéristiques intentionnellement intégrées par les concepteurs afin de créer des défis de complexité variable. Ces caractéristiques englobent, par exemple, la topologie d'un niveau, la quantité, la nature et le positionnement spatial des ennemis, mais aussi l'utilisation de minuteriers, de mécaniques d'équilibrage ou d'éléments dynamiques tels que les plateformes mobiles ou

les obstacles temporisés [8]. Contrairement aux mesures centrées sur le joueur, qui reflètent une réaction individuelle influencée par l'expertise, l'état émotionnel ou les préférences personnelles, la difficulté centrée sur le jeu permet d'évaluer la structure même du défi, indépendamment du profil du joueur.

Cette distinction est cruciale, car la difficulté perçue d'un jeu résulte d'une interaction entre ces deux dimensions. Sans données fiables sur les propriétés internes des niveaux ou des mécaniques du jeu, il devient difficile d'établir des relations causales solides entre les réponses du joueur et les conditions objectives du jeu. De plus, l'absence de jeux de données standardisés axés sur la difficulté intrinsèque nuit à la reproductibilité des travaux scientifiques : chaque laboratoire se voit contraint de créer ses propres niveaux, métriques ou environnements d'évaluation, rendant les comparaisons entre études ardues et limitant l'émergence de modèles généraux.

Ainsi, le développement de jeux de données centrés sur le jeu constitue un enjeu méthodologique majeur pour l'étude de la difficulté vidéoludique. De tels jeux de données permettraient non seulement de mieux comprendre comment la structure d'un niveau influence la performance et l'expérience du joueur, mais aussi de créer des modèles plus robustes et généralisables pour l'analyse, la génération procédurale de contenu ou l'équilibrage automatique des jeux. Cette lacune dans l'état de l'art représente une opportunité importante pour la communauté scientifique et justifie pleinement l'importance de travaux cherchant à combler ce manque.

Dans les prochaines sections, nous tenterons de décrire les principaux jeux de données ayant été proposés par la communauté scientifique en mettant de l'avant leurs forces et leurs limitations de manière à positionner notre contribution.

2.1 LE PLATFORMER EXPERIENCE DATASET (PED)

L'un des premiers et des plus connus des jeux de données est celui de Yannakakis et de son équipe, nommé Platformer Experience Dataset (PED) [13], publié en 2015. PED est un corpus ouvert d'expérience de jeu offrant un jeu de données multimodal composé d'informations recueillies auprès de joueurs d'*Infinite Mario Bros.*, un jeu de plateforme inspiré par le jeu *Super Mario Bros.*. Conçu pour capturer l'expérience des joueurs, ce jeu de données intègre des informations basées sur le contexte (c.-à-d. le contenu du jeu) ainsi que des enregistrements comportementaux et visuels de joueurs de jeux de plateformes. Il inclut également des données démographiques et des annotations auto-rapportées de l'expérience des joueurs sous forme d'évaluations et de classements.

La modélisation de l'expérience des joueurs constitue un domaine de recherche central en intelligence artificielle appliquée au jeu vidéo, en interaction homme-machine et en informatique affective. Deux approches dominent historiquement ce champ : les modèles top-down, fondés sur des théories cognitives et émotionnelles du jeu, et les approches bottom-up, qui s'appuient sur des données empiriques recueillies pendant le gameplay. Si les premières décrivent conceptuellement les dimensions de l'expérience (immersion, flow, engagement), elles peinent à être validées empiriquement dans des contextes interactifs. Les secondes, de plus en plus utilisées, exploitent les données comportementales ou physiologiques afin de construire des modèles prédictifs plus robustes et plus sensibles aux variations du joueur.

Cependant, la progression des approches bottom-up a longtemps été limitée par l'absence de corpus publics et multimodaux permettant une comparaison systématique entre méthodes. Les travaux en modélisation du joueur se basaient fréquemment sur des jeux propriétaires, des

données incomplètes ou des expériences difficiles à reproduire. Les grands corpus ouverts existants (par ex. StarCraft datasets) ne contenaient que des actions en jeu et n'intégraient pas d'informations affectives ou visuelles, empêchant l'étude de la relation entre comportement observable, émotions et perception de la difficulté.

L'article de Karpouzis, Yannakakis, Shaker et Asteriadis [13] introduit le Platformer Experience Dataset (PED), premier corpus multimodal et ouvert consacré à l'expérience de jeu dans un contexte de platformer 2D. Le PED rassemble des données synchronisées provenant de trois sources principales : (1) l'analyse vidéo (expressions faciales, posture, mouvements de la tête et du regard), (2) le comportement en jeu (mouvements, collisions, temps, réussite, trajectoires) et (3) les caractéristiques du niveau (structure, ennemis, obstacles). Ces données sont complétées par des annotations auto-rapportées (échelle d'engagement, frustration, challenge, préférences) et des informations démographiques. Cette richesse multimodale répond directement aux limites des corpus antérieurs et permet d'étudier les interactions entre contrôle moteur, difficulté perçue, réactions visuelles et conception des niveaux

La littérature sur l'estimation de l'expérience de jeu met en lumière plusieurs limites des méthodes fondées uniquement sur les questionnaires : elles sont intrusives, sensibles à la mémoire épisodique, à la subjectivité et aux biais linguistiques. En réponse, diverses approches ont exploré l'utilisation de signaux physiologiques (fréquence cardiaque, EDA), d'analyses comportementales ou de données de gameplay massives. Toutefois, l'intégration simultanée de données visuelles et comportementales était jusqu'ici rare. Les travaux initiaux menés sur le PED montrent justement que la fusion multimodale permet d'atteindre de meilleures performances prédictives que l'utilisation d'une seule modalité, en particulier pour la frustration

et le challenge. Ces résultats confirment l'intérêt des modèles de PEM intégrant à la fois la dynamique visuelle du joueur et les événements clés du jeu.

Le PED est une source centrale dans le domaine du jeu qui permis de combler un manque méthodologique : le besoin d'un environnement reproductible pour l'étude de l'expérience utilisateur en jeu vidéo. En utilisant une version ouverte d'Infinite Mario Bros⁴, l'ensemble des paramètres de génération des niveaux peut être enregistré et contrôlé. Ceci permet non seulement la reproduction exacte des expériences, mais aussi l'utilisation du corpus dans des démarches de génération procédurale de contenu adaptées au profil des joueurs (Experience-Driven Procedural Content Generation).

En somme, l'introduction du PED a marqué une évolution importante pour la recherche sur l'expérience de jeu. Elle fournit un socle empirique standardisé permettant d'explorer les liens complexes entre design de niveaux, réactions comportementales, expressivité visuelle et expérience vécue. Ce corpus multimodal ouvre la voie à des modèles plus complets et reproductibles du joueur, et constitue une ressource essentielle pour les chercheurs travaillant sur la personnalisation adaptative, la détection d'émotions en contexte interactif ou l'analyse de la difficulté dans les jeux vidéo. Cependant, comme mentionnée en introduction, le PED se concentre principalement sur des éléments liés à la difficulté centrée sur le joueur et n'utilise pas les niveaux originaux de *Super Mario Bros.*, mais plutôt des niveaux personnalisés générés de manière procédurale. Les données parties sont limités et ne sont pas centré sur le jeu en tant que tel. Cela limite son utilité pour analyser la difficulté centrée sur le jeu, et l'utilisation de niveaux personnalisés plutôt que de niveaux standard réduit la reproductibilité.

4. <https://www.smbgames.be/infinite-mario.php>

2.2 LE GAMINGVIDEOSET DATASET

Dans le domaine de l'évaluation de la qualité des vidéos de jeux, la recherche s'est longtemps appuyée sur des bases de données vidéographiques classiques, principalement conçues pour des contenus non interactifs. C'est dans cette perspective que Barman et ses collègues ont introduit *GamingVideoSET*, l'un des premiers jeux de données complets dédiés à la recherche sur la qualité des vidéos de jeux [3]. Le corpus, disponible publiquement, comprend 24 vidéos de référence non compressées (30 secondes, 1080p, 30 fps) issues de 12 jeux couvrant un large éventail de genres : MOBA, FPS, RPG, jeux de cartes, simulations, sport et battle royale. Deux scènes par jeu ont été enregistrées afin de représenter des comportements typiques de joueurs et de capturer une grande diversité de complexité visuelle et temporelle. Une analyse basée sur les métriques *Spatial Information (SI)* et *Temporal Information (TI)* confirme que ces vidéos couvrent un spectre étendu de complexité, en cohérence avec les besoins de l'évaluation de la qualité vidéo.

En plus de ces vidéos de référence, *GamingVideoSET* inclut 576 versions compressées obtenues par combinaison de multiples paires résolution-débit binaire, produites à l'aide du codec H.264/MPEG-AVC. Ces vidéos dégradées permettent de simuler les conditions réelles de diffusion sur des plateformes telles que Twitch ou YouTube Gaming, où la compression en temps réel et les fluctuations de bande passante influencent directement la qualité perçue. Le corpus fournit aussi une étude subjective complète conforme aux recommandations ITU-R BT.500, incluant des *Mean Opinion Scores (MOS)* et des taux d'acceptation pour différentes conditions de test.

Parallèlement, plusieurs métriques objectives de qualité vidéo, telles que PSNR, SSIM et VMAF, sont fournies pour l'ensemble du corpus. Ce jeu de données permet ainsi de comparer rigoureusement les performances de différents algorithmes d'évaluation de qualité vidéo, de tester des stratégies d'encodage ou encore de développer des modèles prédictifs de *Quality of Experience* (QoE) pour des contenus vidéoludiques.

Toutefois, malgré sa richesse et sa pertinence méthodologique, *GamingVideoSET* présente une limite majeure lorsque l'objectif de recherche porte sur la difficulté intrinsèque des jeux. Le corpus se concentre exclusivement sur les aspects visuels et sur l'évaluation de la qualité vidéo, sans inclure d'informations centrées sur le jeu lui-même, telles que :

- la topologie des niveaux,
- la disposition des ennemis ou obstacles,
- la nature des objectifs,
- les mécaniques de jeu et leurs contraintes,
- les éléments interactifs complexes (plateformes mobiles, pièges, minuterries, etc.).

Ces informations sont pourtant essentielles pour l'étude de la *difficulté centrée jeu*, c'est-à-dire la difficulté découlant des propriétés intrinsèques du niveau, indépendante de la perception visuelle ou des conditions de diffusion. Par conséquent, bien que *GamingVideoSET* constitue un outil précieux pour les travaux portant sur la qualité vidéo des contenus ludiques, il demeure peu adapté aux recherches visant à analyser ou modéliser la difficulté structurelle ou topologique des jeux vidéo.

2.3 LE TOADSTOOL DATASET

La recherche sur l'intelligence artificielle émotionnelle appliquée aux jeux vidéo s'appuie largement sur des jeux de données multimodaux permettant de capturer l'état émotionnel des joueurs pendant une session de jeu. Historiquement, plusieurs travaux se sont concentrés sur la détection ou la modélisation des émotions à partir de sources variées telles que les images [19], les signaux physiologiques [12], ou encore les expressions faciales [6]. Dans ce contexte, les jeux vidéo constituent un terrain particulièrement intéressant, puisqu'ils génèrent des réactions émotionnelles riches, dynamiques et directement liées à l'interactivité du joueur. Cependant, malgré cet intérêt croissant, peu de jeux de données publics associent simultanément les signaux physiologiques, les vidéos du joueur et les informations issues du jeu lui-même.

C'est dans cette optique que Svoren et al. ont proposé *Toadstool*, un jeu de données multimodal conçu spécifiquement pour étudier les interactions entre l'état émotionnel des joueurs et leur activité dans le jeu [27]. Le corpus comprend des vidéos des joueurs (face camera), des données issues du jeu *Super Mario Bros.*, ainsi que des mesures physiologiques provenant du bracelet médical Empatica E4, incluant la fréquence cardiaque (HR), l'activité électrodermale (EDA), l'accélérométrie, l'interbeat interval (IBI) et le signal photopléthysmographique (BVP). Les données ont été collectées auprès de dix participants présentant des niveaux d'expérience variés, ce qui permet d'obtenir une diversité de réponses émotionnelles et comportementales. L'ensemble du protocole, incluant le questionnaire, la méthodologie d'enregistrement et les scripts de synchronisation, est fourni afin d'assurer une forte reproductibilité.

L'un des principaux intérêts de *Toadstool* réside dans sa polyvalence méthodologique. Les auteurs démontrent par des expériences préliminaires que les signaux physiologiques peuvent

être prédits à partir d'images du jeu et des expressions faciales du joueur, illustrant ainsi le potentiel du dataset pour la modélisation multimodale et l'apprentissage profond. Ce type d'approche s'inscrit dans une tendance de plus en plus répandue visant à enrichir les modèles d'IA avec des signaux physiologiques ou émotionnels pour améliorer la prise de décision ou l'adaptation au joueur, notamment dans les travaux sur les agents intelligents sensibles à l'état émotionnel [18]. Par ailleurs, les applications envisagées par les auteurs incluent l'analyse de l'engagement, la prédiction d'états émotionnels, l'étude de la relation entre performance, stress et difficulté perçue, ainsi que des usages en *reinforcement learning* émotionnellement informé.

Le jeu de données présente également un intérêt pour l'étude des corrélations entre progression dans le jeu, état émotionnel et stratégie du joueur. Grâce au *gym-super-mario-bros*⁵, il est possible de rejouer précisément les sessions de jeu afin d'extraire les frames, les événements du jeu, ou d'analyser les comportements de navigation du joueur à travers les niveaux. Cette capacité de relecture ouvre la voie à des analyses plus fines des interactions entre structure du jeu et réactions physiologiques.

Cependant, malgré sa richesse multimodale, *Toadstool* présente une limite importante lorsqu'il s'agit d'étudier la *difficulté intrinsèque centrée jeu*. En effet, le dataset est fortement orienté sur le joueur : signaux physiologiques, expressions faciales, données démographiques, et actions du joueur. Les informations relatives à la structure même des niveaux de *Super Mario Bros.* demeurent limitées. Le corpus ne fournit pas, par exemple, la topologie détaillée des niveaux, la position, le type des ennemis et obstacles, etc. Ces informations sont pourtant essentielles pour analyser la *difficulté centrée sur le jeu* que nous avons défini. Sans une description complète des éléments structurels de chaque niveau, *Toadstool* ne permet pas d'étudier de manière

5. <https://pypi.org/project/gym-super-mario-bros/>

systématique la relation entre difficulté intrinsèque d'un niveau et réactions émotionnelles du joueur. Le dataset constitue donc une ressource de premier plan pour l'étude de la difficulté *centrée joueur*, mais il reste insuffisant pour des analyses orientées sur la complexité structurelle ou topologique du jeu.

2.4 LE PLAYMYDATA DATASET

Dans la même lignée d'analyse comportementale, D'Angelo et al. ont récemment introduit *PlayMyData* [5], un jeu de données de grande ampleur rassemblant 99 864 jeux issus des principales plateformes vidéoludiques : PlayStation, Xbox, Nintendo et PC. Le dataset est collecté à partir du site IGDB grâce à une API dédiée, ce qui permet d'extraire automatiquement des métadonnées structurées telles que le résumé du jeu, le genre, la date de sortie, les évaluations utilisateurs, les captures d'écran (443 630 images) et les liens vers des vidéos de gameplay (43 812 URL). L'usage d'une API officielle constitue un avantage majeur en termes de reproductibilité et de mise à jour potentielle du corpus, un problème touchant de nombreux jeux de données antérieurs dans le domaine.

Une particularité notable de *PlayMyData* est l'enrichissement du corpus par les temps de complétion issus de la base *HowLongToBeat* (HLTB). Ces données incluent le temps nécessaire pour compléter l'histoire principale, les quêtes secondaires (*main+extra*) et le jeu dans son intégralité (*completionist*), permettant ainsi d'ouvrir la voie à des analyses temporelles du gameplay. La fusion entre les jeux IGDB et HLTB est effectuée via une distance de Levenshtein [20], limitant les erreurs de correspondance entre les titres. Au total, 35 815 jeux disposent d'un temps de complétion associé.

Grâce à sa couverture multi-plateforme et à sa diversité de données multimodales (texte, images, vidéos, métadonnées structurées), *PlayMyData* soutient un large éventail d'applications, allant de la classification automatique des jeux [11] aux approches multimodales en vision par ordinateur [14]. Les auteurs identifient également des usages potentiels en analyse de sentiments, systèmes de recommandation ou encore en recherche sur l'évolution de l'industrie du jeu vidéo au fil des décennies. Le corpus constitue ainsi une base méthodologiquement solide pour les travaux combinant fouille de données, IA et ingénierie logicielle.

Toutefois, malgré son ampleur et sa richesse, *PlayMyData* présente une limite importante pour les recherches portant sur la *difficulté centrée jeu*. En effet, l'accent est fortement mis sur :

- les métadonnées descriptives (résumé, genres),
- les supports multimédias (captures d'écran, vidéos),
- les données comportementales globales (temps pour terminer le jeu),
- la disponibilité multi-plateforme.

Le dataset ne fournit aucune information détaillée sur la structure interne des jeux, ou les éléments plus orientés sur la structure d'un niveau ou sur les points de difficulté. Ces éléments sont pourtant fondamentaux pour étudier la *difficulté intrinsèque* ou *game-centric difficulty*, c'est-à-dire la difficulté directement liée à la conception du niveau ou aux mécaniques du jeu. Ainsi, malgré sa grande valeur pour les études en ingénierie logicielle, classification, recommandation ou analyse multimodale, *PlayMyData* ne permet pas d'évaluer correctement la complexité structurelle des défis proposés par les jeux. Le corpus reste orienté vers des analyses globales et descriptives des jeux plutôt que vers l'étude de leur architecture interactive, limitant son utilité pour les travaux visant à modéliser ou quantifier la difficulté intrinsèque.

2.5 STARDATA : UN DATASET POUR L'IA DANS LE JEU STARCRAFT

La recherche en intelligence artificielle appliquée aux jeux de stratégie en temps réel (RTS) bénéficie depuis plusieurs années de l'existence de jeux de données fondés sur des replays de parties humaines. *StarCraft : Brood War*⁶, en particulier, constitue un environnement privilégié en raison de sa complexité stratégique, de sa dynamique multi-échelle, de sa nature partiellement observable et de la disponibilité publique de nombreuses parties expertes. Ces caractéristiques en font un terrain d'expérimentation idéal pour des approches variées, notamment en classification stratégique [21], en apprentissage par renforcement profond [29, 23], en planification tactique ou encore en modélisation du comportement des joueurs [28]. Cependant, la majorité des jeux de données existants souffrent de limitations importantes : formats hétérogènes, fréquence d'échantillonnage trop faible, incompatibilités entre versions du jeu ou encore absence de validation systématique des replays.

C'est pour combler ses lacunes que Lin et al. ont créé le *STARDATA* [15], un jeu de données de très grande envergure contenant 65 646 replays de *StarCraft*, soit environ dix fois plus que les corpus précédemment disponibles. Le dataset fournit à la fois les replays originaux et un enregistrement complet de l'état du jeu, capturé toutes les trois images (environ 8 fps) grâce à la bibliothèque TorchCraft. Ce choix garantit une résolution temporelle fine et exploitable pour un large éventail de tâches, allant de la micromanipulation d'unités (*micro*) à la planification économique et stratégique (*macro*). *STARDATA* répond explicitement à plusieurs exigences méthodologiques fondamentales : universalité des données, diversité des stratégies, validité des parties, portabilité entre systèmes et standardisation de l'interface d'accès au corpus.

6. <https://starcraft.blizzard.com/fr-fr/>

L'une des forces majeures du dataset réside dans sa diversité. Les replays proviennent principalement de plateformes compétitives telles que ICCup⁷, TeamLiquid⁸ et GosuGamers⁹, et couvrent un large éventail de cartes (524 cartes uniques), de confrontations (PvZ, PvT, ZvZ, etc.) et de niveaux de compétence. Les auteurs ont appliqué plusieurs heuristiques de filtrage afin d'éliminer les parties corrompues, notamment celles produites par des versions incompatibles du jeu ou présentant des comportements atypiques (faible utilisation des ressources, situations bloquées). STARDATA inclut également des métadonnées structurées utiles pour des tâches spécifiques, telles que la détection automatique de batailles ou l'analyse de styles d'ouverture (clustering d'armées dans les premières minutes de jeu).

Ce corpus permet ainsi d'aborder des problèmes variés :

- **classification stratégique** (analyse des styles, prédiction des *build orders*);
- **apprentissage par renforcement inversé** (approximation de fonctions de récompense à partir de traces expertes);
- **apprentissage par imitation** (cartographie état–action à partir de démonstrations humaines);
- **modélisation prédictive** (forward modelling, vidéo ou état futur du jeu);
- **gestion de l'information partielle** (fog of war, estimation des unités ennemies).

La granularité fine des états enregistrés permet aux chercheurs de construire des modèles précis de la dynamique du jeu, tandis que le volume massif du dataset répond aux besoins des algorithmes d'apprentissage profond, traditionnellement très gourmands en données. STARDATA représente donc une ressource clé pour les travaux visant à dépasser les environ-

7. <https://iccup.com/>

8. <https://teamliquid.com/>

9. <https://www.gosugamers.net/>

nements simplifiés et à étudier l'intelligence artificielle dans des contextes riches, complexes et adversariaux.

Cependant, malgré ses nombreuses qualités, *STARDATA* présente une limite importante dans le cadre spécifique de l'étude de la *difficulté centrée jeu*. Le dataset est conçu pour supporter des tâches de contrôle, de stratégie ou de modélisation comportementale, mais il n'inclut pas de caractérisation explicite et structurée des éléments intrinsèques des cartes ou des défis posés par l'environnement. Autrement dit, *STARDATA* fournit l'état complet du jeu à chaque instant, mais ne propose aucune abstraction ou annotation permettant d'évaluer la difficulté intrinsèque du terrain de jeu ou des situations rencontrées. Cette absence de données centrées sur la structure de l'environnement limite l'utilisation du dataset pour des analyses cherchant à corréler difficulté objective et performance du joueur. Par conséquent, bien que *STARDATA* constitue une ressource incontournable pour la recherche en IA sur les RTS, il reste peu adapté à l'étude de la difficulté intrinsèque telle que définie dans notre cadre, et ne permet pas d'extraire directement des mesures structurelles comparables entre parties ou entre cartes.

2.6 LE DATASET AGAIN

Dans le domaine de l'informatique affective appliquée aux jeux vidéo, la capacité de modéliser les émotions de manière générale, c'est-à-dire au-delà d'un seul jeu, d'une seule tâche ou d'un seul contexte, demeure un défi méthodologique majeur. La majorité des corpus de données existants reposent sur des stimuli non interactifs (images, vidéos, musique), ou sur un nombre limité de jeux offrant des dynamiques émotionnelles restreintes. Pour répondre à ce manque, Melhart, Liapis et Yannakakis ont introduit *AGAIN* (Arousal Game AnnotatIoN Dataset) [7], un corpus à grande échelle explicitement conçu pour étudier la dynamique de l'*arousal* dans

un large éventail de situations ludiques interactives. Dans le contexte des jeux vidéo, l'*arousal* désigne le niveau d'activation émotionnelle ou physiologique ressenti par le joueur pendant l'interaction. Il reflète l'intensité de la réaction du joueur face aux événements du jeu, allant d'un état de calme ou d'ennui (faible *arousal*) à un état d'excitation, de stress ou de tension (fort *arousal*). L'*arousal* permet ainsi de mesurer la dynamique émotionnelle du joueur en fonction des défis, de la vitesse, des dangers ou de l'incertitude présents dans le gameplay.

Le dataset AGAIN se distingue par son ampleur et sa diversité : plus de 1 100 sessions de jeu annotées, 37 heures de vidéos de gameplay, et des données issues de neuf jeux originaux construits spécialement pour l'expérimentation. Ces jeux couvrent trois genres différents, course, tir et plateforme, et ont été développés pour proposer des mécaniques simples, immédiatement compréhensibles et adaptées à la collecte rapide via crowdsourcing. Cette approche permet de capturer des réactions émotionnelles spontanées, situées dans des contextes interactifs variés, tout en maintenant une structure de données comparable d'un jeu à l'autre.

L'un des aspects méthodologiques les plus innovants du dataset réside dans l'utilisation du protocole *RankTrace* [17], qui permet l'annotation continue de l'*arousal*. Les participants, après avoir joué chaque jeu pendant deux minutes, réannotent leur propre vidéo via une interface dédiée. Cette annotation en continu, non bornée, permet de capturer les nuances temporelles et relatives des fluctuations émotionnelles, en évitant les artefacts typiques des échelles absolues (saturation, difficulté d'étalonnage, biais d'ancrage). L'approche met l'accent sur les relations ordinales entre les segments, ce qui facilite l'application de méthodes d'apprentissage basées sur la comparaison ou la préférence.

En complément des vidéos, AGAIN inclut un ensemble riche de télémétries comportementales : actions du joueur, déplacements, objets visibles, événements contextuels, interactions avec les ennemis, etc. Ces données, normalisées entre les différents jeux, permettent d'étudier les corrélations entre l'évolution du contexte ludique et l'*arousal* rapporté par les joueurs. Les auteurs proposent également un pipeline de nettoyage rigoureux, combinant resampling, détection d'irrégularités via *Dynamic Time Warping* [1], et élimination des sessions comportant des comportements aberrants ou des annotations incohérentes. Le résultat est un corpus propre, robuste et particulièrement adapté à l'analyse automatique.

AGAIN offre ainsi un terrain d'expérimentation favorable à de nombreuses approches : modélisation de l'*arousal* à partir de la vidéo seule (vision par ordinateur), apprentissage multimodal combinant télémétrie et pixels, étude du transfert de modèles affectifs entre jeux distincts, création de modèles de prédiction de l'engagement, du stress ou du flow. L'article propose d'ailleurs des modèles de base (*baseline models*) fondés sur des forêts aléatoires (*random forests*), qui démontrent la faisabilité de la prédiction de l'*arousal* et la pertinence des caractéristiques comportementales générales extraites. Bien que ce dataset contiennent en partie des éléments orienté jeu, tels que le nombre d'interactions avec les ennemis, il n'est que partiellement complet en ce sens et ne peut répondre entièrement à nos besoins en terme d'évaluation automatique du niveau de difficulté centré gameplay.

2.7 LA DATASET AMUCS

Toujours dans la lignée de l'analyse multimodale des émotions et des comportements en contexte vidéoludique, Fanourakis et Chanel ont cette année introduit AMuCS [7], un jeu de données affectif multimodal enregistré lors d'événements *LAN party*, offrant ainsi des

conditions plus naturelles et écologiquement valides pour l'étude des émotions en jeu vidéo. AMuCS se distingue principalement par sa taille et sa diversité. Le corpus rassemble les données de 256 participants jouant à *Counter-Strike : Global Offensive*¹⁰, un jeu de tir compétitif exigeant dont l'intensité émotionnelle constitue un terrain particulièrement propice à l'observation du stress, de la charge cognitive et des réactions physiologiques. L'un des apports majeurs du dataset est son enregistrement dans un contexte réaliste, c'est-à-dire lors de compétitions LAN, où les joueurs évoluent dans un environnement social, sonore et compétitif difficilement reproductible en laboratoire. Cette authenticité renforce la validité écologique du corpus et permet de capturer des comportements plus proches des situations réelles de jeu compétitif.

Le jeu de données intègre un ensemble impressionnant de modalités complémentaires : signaux physiologiques (ECG, activité électrodermale, respiration), signaux comportementaux (expressions faciales, mouvements oculaires, images de profondeur, pression du siège), interactions clavier-souris, vidéos de gameplay et journaux d'événements internes au jeu. Cette richesse multimodale permet d'explorer les relations entre charge physiologique, performance, tension émotionnelle, coordination perceptivo-motrice et fluctuations d'état internes. L'article montre d'ailleurs que le volume de données et la variété des modalités sont particulièrement adaptés à l'entraînement de modèles d'apprentissage automatique, notamment pour la prédiction d'états émotionnels ou la classification de comportements complexes.

AMuCS ouvre ainsi la porte à un large éventail d'applications : modélisation du stress en temps réel, détection des moments critiques d'une partie, analyse de la dynamique émotionnelle selon les rôles ou les situations de jeu (attaque, défense, confrontation), ou encore création de

10. <https://www.counter-strike.net/>

systèmes d'adaptation intelligents capables d'ajuster la difficulté ou le rythme du jeu en fonction de l'état de l'utilisateur. La combinaison de données physiologiques fines et d'annotations comportementales précises en fait un corpus rare et précieux pour la recherche en informatique affective et en interaction humain-machine.

Bien que AMuCS constitue une ressource exceptionnelle pour l'étude des émotions, du stress et des comportements humains dans un contexte compétitif, il présente des limites importantes lorsqu'il est examiné sous l'angle de la difficulté intrinsèque centrée-jeu, qui est au cœur de notre cadre de recherche. Plus spécifiquement, AMuCS ne fournit aucune description structurelle des niveaux ou des cartes de Counter-Strike : Global Offensive. Les éléments essentiels à l'analyse de la difficulté centrée-jeu, topologie des environnements, zones de risque, chokepoints, lignes de vue, répartition des objectifs, etc., ne sont pas inclus.

Le dataset est profondément orienté joueur : la quasi-totalité des modalités décrivent l'état du joueur (physiologie, comportements, actions), mais pas la structure formelle du défi posé par l'environnement. Les données de gameplay accessibles (vidéos, logs, actions) permettent de déduire des comportements ou des stratégies, mais pas de reconstruire systématiquement la complexité intrinsèque des situations, faute d'annotations topologiques ou d'un modèle explicite des cartes.

Ainsi, même s'il s'agit d'un corpus clé pour l'analyse affective et comportementale, AMuCS ne peut pas être utilisé pour étudier la difficulté comme propriété du jeu lui-même, ni pour explorer la relation entre les caractéristiques structurelles d'un niveau et la performance des joueurs. Il demeure un outil complémentaire, plutôt qu'un dataset adapté à l'analyse de la difficulté centrée-jeu.

2.8 DATASET SUR COUNTER-STRIKE : GLOBAL OFFENSIVE

La recherche sur les jeux de tir à la première personne (FPS) repose eux aussi de plus en plus sur des approches empiriques faisant intervenir des corpus massifs, annotés et reproductibles. Dans ce contexte, l'article récent de Halbhuber et al. [9] apporte une contribution importante en proposant un dataset inédit dédié à l'étude des joueurs de *Counter-Strike : Global Offensive* (CS :GO), le même jeu utilisé pour le dataset AMuCS, un des FPS les plus populaires au monde, avec plusieurs centaines de milliers de joueurs quotidiens. Alors que plusieurs corpus préexistants s'intéressaient déjà à certains aspects du jeu, comme les statistiques de matchs, les comportements oculaires, ou les données de haut niveau concernant les armes ou les performances, aucun ne fournissait un accès direct au gameplay visuel synchronisé avec des données d'entrée utilisateur, un élément crucial pour les analyses fines du comportement et de la performance dans un environnement FPS compétitif. Le dataset de Halbhuber et al. comble précisément cette lacune en offrant un matériel riche, complet et hautement annoté, aligné au niveau de la frame.

La démarche expérimentale est centrée sur la collecte de replays POV (*Point of View*), un format interne de CS :GO permettant l'enregistrement d'une partie depuis la perspective du joueur tout en conservant un accès détaillé à l'état du jeu. Afin de garantir la validité écologique et d'éviter les biais liés à des environnements contrôlés, les auteurs ont demandé à douze joueurs expérimentés (temps moyen de jeu : 1864 heures) d'enregistrer leurs parties directement depuis leur propre matériel, dans des conditions naturelles de jeu. Cette décision vise à maximiser la représentativité des comportements, tout en permettant d'obtenir un corpus homogène en termes de compétence grâce à un filtrage préalable par rangs compétitifs.

La génération du dataset repose ensuite sur une pipeline technique avancée combinant extraction visuelle et extraction d'état. À l'aide de l'outil *H LAE*¹¹, les auteurs ont extrait plus de 2,7 millions d'images à partir des replays, représentant plus de 2 764 800 frames capturées à 64 fps. Ce matériel visuel est proposé en version brute et en deux versions réduites (60×34 pixels et 30×17 pixels), facilitant son utilisation dans des modèles de deep learning. Parallèlement, l'outil demofile a été employé pour extraire des informations complètes sur l'état du jeu à chaque frame : santé, armure, localisation, arme équipée, situation tactique (joueur repéré ou non), possession d'objets clés (C4, kit de désamorçage), ainsi que toutes les entrées clavier et souris (touches pressées, déplacements relatifs du curseur). L'ensemble constitue un dataset multimodal parfaitement synchronisé, permettant d'explorer conjointement l'environnement visuel, les actions du joueur et l'évolution de la situation tactique.

L'intérêt de ce dataset réside également dans ses possibilités d'utilisation. Les auteurs en proposent deux démonstrations : la prédiction du mouvement de la souris à l'aide d'un réseau neuronal *feed-forward*, et la prédiction de frames futures à partir des données visuelles. Ces exemples illustrent la richesse du corpus pour des tâches comme la modélisation du comportement, la compensation de latence, l'analyse de style de jeu, ou encore la prédiction en temps réel d'éléments visuels dans des environnements dynamiques. Plus largement, ce dataset peut soutenir un large éventail de travaux : analyse du stress ou de la pression en compétition, identification de patterns de décision dans des situations de *clutch*, apprentissage supervisé d'agents capables de détecter ou anticiper des menaces, ou encore développement d'outils d'entraînement pour joueurs novices.

11. <https://github.com/f0e/hlae-configs>

Ainsi, la contribution de Halbhuber et al. représente un pas essentiel vers la constitution d'un écosystème de données riche pour l'analyse des FPS. En rendant accessibles et reproductibles des données de gameplay visuel et comportemental jusque-là difficiles à obtenir, leur travail ouvre la voie à des avancées méthodologiques et expérimentales dans un champ où l'accès aux données constitue un verrou majeur. Leur démarche s'inscrit dans notre optique et dans cette tendance plus large visant à doter le domaine vidéoludique de ressources structurantes, comparables à celles qui ont transformé d'autres secteurs de l'intelligence artificielle et des sciences des données. Cependant, comme dans l'ensemble des jeux de données évalués dans cette section, celui-ci ne cible pas les données centrées jeu et ne permet pas d'évaluer correctement le type de difficulté envisagée par nos travaux.

2.9 CONCLUSION

En conclusion, l'absence de jeux de données standardisés spécifiquement dédiés à l'évaluation de la difficulté centrée sur le jeu représente un obstacle majeur pour la recherche comme pour les applications pratiques [8]. Les rares jeux de données actuellement disponibles [13, 27, 31, 5, 30, 10] se concentrent principalement sur des éléments centrés sur le joueur, comme les signaux physiologiques, les actions en jeu, les comportements capturés ou encore les réponses émotionnelles. Bien que ces corpus soient utiles pour comprendre l'expérience subjective du joueur, ils ne permettent pas d'étudier de manière systématique la structure interne du jeu, ses mécaniques, ni la complexité intrinsèque des défis proposés.

Cette situation crée une zone d'ombre importante dans la littérature : il manque un pont entre la modélisation de la performance du joueur, l'analyse objective des caractéristiques des niveaux, et la mesure de la difficulté perçue. L'absence de données standardisées limite la comparabilité

entre études, empêche la reproductibilité expérimentale et freine le développement de modèles généraux capables d'expliquer comment la conception d'un niveau influence concrètement la difficulté rencontrée par différents profils de joueurs.

Ainsi, un besoin pressant se fait sentir pour un jeu de données complet et transversal, qui intégrerait simultanément : (i) des mesures de performance du joueur (réussites, erreurs, trajectoires, temps d'exécution), (ii) des caractéristiques détaillées des niveaux (topologie, ennemis, obstacles, éléments dynamiques, structure spatiale), (iii) des évaluations subjectives de la difficulté perçue, (iv) et ce, à travers une diversité de jeux, de styles et de genres vidéoludiques.

Un tel corpus constituerait un outil fondamental pour l'étude rigoureuse de la difficulté vidéoludique, ouvrirait la voie à des méthodologies comparatives robustes et permettrait d'élaborer des modèles prédictifs plus précis et plus généralisables. Il représenterait également un apport significatif pour les concepteurs de jeux, les chercheurs en IA, les spécialistes de l'adaptive gameplay et les développeurs de systèmes de génération procédurale de contenu. Ainsi, notre contribution dans ce mémoire [2] vise à apporter des solutions à cette lacune. Elle constitue une étape clé pour faire progresser l'étude scientifique de la difficulté centrée sur le jeu et pour permettre une compréhension plus fine du lien entre design, performance et expérience du joueur.

CHAPITRE III

LADDER : UN NOUVEAU DATASET POUR L'ÉVALUATION DE LA DIFFICULTÉ DANS LES JEUX DE PLATEFORME

La chapitre précédent sur la revue littérature à montrer l'absence de jeux de données standardisés, qui constitue un frein majeur à la reproductibilité des études, à la comparaison rigoureuse des méthodes et au développement de modèles généralisables. Alors que des disciplines telles que la vision par ordinateur ou le traitement automatique du langage bénéficient depuis longtemps de bases de données de référence, favorisant l'émergence de benchmarks, de méthodologies partagées et d'avancées rapides, la recherche sur la difficulté dans les jeux vidéo repose encore très largement sur des jeux de données disparates, incomplets ou centrés uniquement sur le joueur.

Notre contribution originale dans ce mémoire, qui prend la forme d'un article scientifique intitulé : **LADDER : Level Analysis Dataset for Difficulty Evaluation and Ranking** [2] propose de combler en partie cette lacune en introduisant LADDER, un nouveau jeu de données standardisé spécifiquement conçu pour l'analyse de la difficulté des niveaux dans les jeux de plateforme. La contribution de ce travail réside dans la création d'un dataset structuré, reproductible et ouvert, intégrant simultanément trois dimensions essentielles : (i) des métriques objectives de performance du joueur (ex. : nombre de morts, nombre d'essais,

temps de complétion, nombre de sauts), (ii) des caractéristiques intrinsèques des niveaux (ex. : nombre d'ennemis, types et positions des obstacles, nombre de trous, briques, danger zones), (iii) des évaluations subjectives de la difficulté perçue, recueillies à travers différentes sources de données publiques.


Le dataset se distingue par son ampleur et sa diversité. Il inclut notamment les 32 niveaux originaux de Super Mario Bros. accompagnés de 386 947 sessions de jeu, ainsi que des données issues de Super Mario 64, Super Mario World, Rockman/Megaman et Super Meat Boy. Cette richesse permet d'associer finement les comportements observés des joueurs aux éléments structurels des niveaux, et offre un terrain d'expérimentation commun pour tester, comparer et valider des approches d'évaluation de la difficulté, qu'elles soient centrées sur le joueur (ML) ou centrées sur le jeu (analyse structurelle, modélisation formelle).



L'article montre également le potentiel analytique de LADDER à travers plusieurs exemples d'exploitation : génération de heatmaps de difficulté basées sur les échecs des joueurs, comparaison entre une approche d'apprentissage automatique et un modèle formel non-ML de difficulté motrice, ou encore corrélation avec la courbe de progression classique décrite par Nishikado. Ces résultats permettent de visualiser comment un même corpus standardisé permet, pour la première fois, d'évaluer la cohérence, la stabilité et la validité scientifique de méthodes hétérogènes.

En final, LADDER constitue une contribution structurante pour la communauté scientifique en offrant une base commune, ouverte et reproductible pour l'étude de la difficulté dans les jeux vidéo. Il ouvre la voie à de nouvelles méthodes de modélisation, facilite les comparaisons entre approches, et favorise la conception de jeux plus équilibrés, accessibles et engageants.

3.1 ARTICLE SCIENTIFIQUE PRÉSENTÉ À IEEE-GEM 2025

LADDER: Level Analysis Dataset for Difficulty Evaluation and Ranking

Yao Jean-Eudes Adjanohoun
dept. of computer science and mathematics
Université du Québec à Chicoutimi
Chicoutimi, Canada
yjadjanoho@etu.uqac.ca
Hugo Tremblay 
dept. of computer science and mathematics
Laboratory of Research in Modelling, Graphs
and Optimization
Université du Québec à Chicoutimi
Chicoutimi, Canada
hugo_tremblay2@uqac.ca

Yannick Francillette 
dept. of computer science and mathematics
Université du Québec à Chicoutimi
Chicoutimi, Canada
yannick_francillette@uqac.ca
Bruno Bouchard 
dept. of computer science and mathematics
Université du Québec à Chicoutimi
Chicoutimi, Canada
bruno.bouchard@uqac.ca

Abstract—Standardized datasets are fundamental to scientific research. While fields like natural language processing and computer vision have widely accepted datasets that drive progress, video game research still lacks such resources, particularly for studying level difficulty. Existing studies rely on isolated, custom datasets, limiting cross-study comparisons and hindering the development of generalizable models. To bridge this gap, we introduce LADDER, a novel dataset specifically designed to analyze and evaluate level difficulty in video games. Unlike previous datasets that primarily focus on physiological and behavioral player data, LADDER integrates objective performance metrics (e.g., health lost, number of attempts before success), level characteristics (e.g., number of danger zones, object placement), and perceived difficulty ratings across multiple platformer games. This dataset enables researchers to establish benchmarks, enhance collaboration across disciplines, and improve study reproducibility. LADDER provides a standardized foundation for investigating the relationship between game design elements and player experience. By facilitating difficulty assessment and level balancing, it supports advancements in game design, player modeling, and adaptive gameplay systems. We present an overview of existing datasets, describe the methodology behind LADDER’s construction, and showcase its potential through preliminary analyses. The dataset is freely available online, offering a valuable resource for the scientific community to develop more engaging and accessible gaming experiences.

Index Terms—Video Game, Level of Difficulty, Standardized Dataset, Platformer Games

I. INTRODUCTION

Standardized datasets have become an essential pillar of scientific research, fostering reproducibility, comparability, and innovation across disciplines [1]. They enable researchers to validate findings, test new hypotheses, and develop benchmark methodologies, ensuring that progress is built on a consistent foundation. In fields such as natural language processing or computer vision, the availability of widely accepted datasets (e.g., ImageNet¹, COCO²) has driven unprecedented advances

¹<https://image-net.org/>

²<https://cocodataset.org/>

by creating a common ground for evaluation [2].

In the context of video game research [3], however, the lack of standardized datasets remains a critical barrier [4]. Studies on game difficulty, player behavior, and level design often rely on custom, isolated datasets tailored to specific experiments. This fragmentation hampers cross-study comparisons and slows the development of generalizable models or insights [5]. A standardized dataset addressing level difficulty, for example, would allow researchers to:

- Establish baselines: By providing a shared benchmark, such datasets can facilitate the comparison of algorithms or models for assessing difficulty.
- Encourage collaboration: A shared resource can unite researchers from different domains (e.g., AI, psychology, design) and foster interdisciplinary studies.
- Promote transparency: Open datasets make it easier to replicate studies, assess the robustness of findings, and identify potential biases in experimental designs.

The creation of such a dataset tailored to level difficulty would not only fill this gap but also pave the way for standardized methodologies in video game research, much like ImageNet has done for computer vision [2]. It represents a necessary step toward advancing our understanding of how players interact with games and how design influences engagement, performance, and satisfaction.

In this paper, we aim to bridge this gap by providing researchers with a tool to explore and refine concepts of difficulty in a reproducible, scalable, and insightful manner. More specifically, we focus on analyzing the difficulty of individual video game levels. The challenge posed by each level plays a critical role in shaping the player experience, directly influencing engagement, satisfaction, and retention [5]. Despite its importance, “level difficulty” remains a subjective concept, highly dependent on the interplay between level design and player skill [6]. Quantifying difficulty is essential

not only for designing balanced and engaging games but also for advancing research in game design, player modeling, and adaptive gameplay systems [7]. However, the lack of standardized datasets dedicated to evaluating level difficulty presents a major challenge for both researchers and practitioners.

Existing datasets [7]–[12] primarily focus on player data collected during gameplay, such as eye movements, blood pressure or stress levels. A comprehensive dataset that integrates objective player performance metrics (e.g., health lost, number of attempts before success, bullets used), level characteristics (e.g., number of danger zones, object placement), and perceived difficulty across various games and genres is therefore crucial.

To address this issue, we introduce LADDER, a novel dataset specifically designed to analyze and evaluate level difficulty in platformer video games. This dataset provides a robust foundation for studying the relationship between game design elements and player behavior, facilitating the development of more advanced tools for difficulty assessment and level balancing. Ladder is freely available online and can be used by all the scientific community³. We first review existing datasets, then describe the methodology used to construct LADDER, highlight its key features, and demonstrate its potential through a series of preliminary analyses. By offering a standardized resource, we aim to support future research and innovation in game design, ultimately contributing to more engaging and accessible gaming experiences for players of all skill levels.

II. RELATED WORK

In the scientific literature, very few standardized datasets have been proposed to investigate difficulty in video games [7]–[12]. Most existing datasets focus on player-centric elements collected during gameplay sessions [6], [8], [12], such as blood pressure, eye position, and other physiological or behavioral metrics. This highlights a clear need for datasets emphasizing game-centric elements that measure the intrinsic difficulty of the challenges posed by the game itself.

Intrinsic (or game-centric) difficulty refers to the features intentionally designed to create challenges of varying complexity. These include, for example, the topology of a level, the number, type, and positioning of enemies, the use of timers, balancing mechanics, and dynamic elements such as moving platforms [13].

One of the first and most well-known dataset is the one of Yannakakis and his team named Platformer Experience Dataset (PED) [8], which as been published in 2015. PED is an open-access game experience corpus providing a multimodal dataset of user data collected from *Infinite Mario Bros.* players. Designed to capture player experiences, the database integrates context-based information (i.e., game content) with behavioral and visual recordings of platform game players. Additionally, it includes demographic data and self-reported annotations of player experience in the form of ratings and rankings.

³<https://github.com/lannnyck/LADDER>

However, it primarily focuses on elements related to player-centric difficulty and did not use the original levels of *Super Mario Bros.*, but rather custom levels procedurally generated. This limits its usefulness for analyzing game-centric difficulty, and the use of custom levels instead of standard ones reduces reproducibility.

Later, Barman et al. [14] introduced GamingVideoSET, a dataset comprising twenty-four uncompressed raw gaming videos, each 30 seconds long, with 1080p resolution and a frame rate of 30 fps. While this dataset was primarily designed for research on gaming video quality assessment and the effects of video encoding techniques, it was not explicitly created for difficulty assessment. Although it could potentially be repurposed for evaluating game difficulty, its utility is significantly limited by the lack of annotations, absence of player performance data, and missing descriptions of the levels featured in the videos.

Svoren et al. [9] introduced Toadstool: A Dataset for Training Emotionally Intelligent Machines Playing *Super Mario Bros.* The dataset comprises video, sensor, and demographic data collected from ten participants during gameplay. Sensor data was gathered using the Empatica E4 wristband, a device that provides high-quality measurements and is classified as a medical-grade tool. While this dataset can be used to analyze the difficulty of a level, it primarily focuses on player-oriented data and lacks information about game-centric elements, limiting its applicability for evaluating intrinsic game difficulty.

Recently, in 2024, D’Angelo et al. [11] introduced PlayMyData, a curated dataset comprising 99,864 multi-platform games gathered from the IGDB website. Using a dedicated API, they collected relevant metadata for each game, including descriptions, genres, ratings, gameplay video URLs, and screenshots. Additionally, they enriched PlayMyData with completion time estimates for each game by mining data from the HowLongToBeat⁴ website. This dataset is one of the most comprehensive currently available. However, it was not specifically designed to analyze game difficulty, and it lacks detailed data for individual game levels, limiting its utility for this particular purpose.

In summary, the lack of standardized datasets dedicated to evaluating game-centric difficulty presents a significant challenge for both researchers and practitioners [13]. The few datasets currently available [7]–[12] predominantly focus on player-centric elements. There is a pressing need for a comprehensive dataset that integrates player performance, level characteristics, and perceived difficulty across a variety of games and genres.

III. LADDER: A NEW LEVEL ANALYSIS DATASET FOR DIFFICULTY EVALUATION AND RANKING

Our project aims to develop a comprehensive and structured dataset containing player performance data for a series of platformer video games, and the information concerning each levels. The primary games selected for data collection are

⁴<https://howlongtobeat.com/>

*Super Mario Bros.*⁵, *Super Meat Boy*⁶, and *Rockman*⁷, Figure 1, and 2 show screenshot of some of these games.

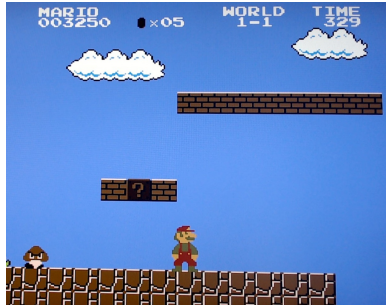


Fig. 1: *Super Mario Bros.*: in game screenshot

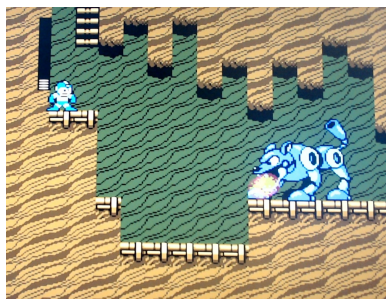


Fig. 2: *Rockman 2*: in game screenshot

The goal is to gather precise and detailed information on the levels of each games and on playing sessions containing player actions, for instance the time spent on each level, the number of attempts required to complete an objective, the number of death, the number of jumps, the areas where players encounter the most difficulty, and the strategies employed to overcome obstacles.

Our study primarily rely on freely available online data. It also incorporate data extracted from in-game performance capture tools and post-session questionnaires to gather player feedback on perceived difficulty. In the following sections, we present methodology form acquiring data, the structure of the dataset and its contain of the dataset. Thereafter, we show how the dataset can be used to process and analyzed game difficulty using statistical and machine learning techniques to uncover correlations between player performance and game design elements.

A. Data Sources for Dataset Creation

We used different freely available sources to get the necessary data that allowed us to construct the dataset. Here a list

⁵<https://www.nintendo.com/jp/famicom/software/smb1/index.html>

⁶<http://www.supermeatboy.com/>

⁷<https://www.capcom-games.com/ja-jp/series/megaman/>

of the data source:

- 1) **Speedrun.com** is a community-driven platform dedicated to speedrunning, where players from around the world share their performances and records across various video games. The dedicated *Super Mario Bros.* (SMB1) page features leaderboards, speedrun categories, guides, and run videos, serving as a key reference point for enthusiasts looking to optimize their times and compete with others. Speedrun.com offers data from several popular games, including *Super Mario Bros.*, *Super Mario 64*⁸, *Rockman*, and *Super Meat Boy*.
- 2) **Kaggle.com** is an online platform that serves as a hub for data science and machine learning communities. It hosts datasets related to player performance in *Super Mario Maker*⁹. Notably, the "SMMnet - Super Mario Maker" dataset¹⁰ offers temporal data on player interactions within the game [15]. Accompanying this dataset are analyses such as "Data Analysis - Super Mario Maker," which explores player performance metrics, and "Super Mario Maker Top 10 Players," focusing on top player statistics. These resources provide insights into player behaviors, level completion times, and other performance indicators within *Super Mario Maker*.
- 3) **Archive.org**, also known as *the Internet Archive*, is a non-profit digital library that provides free access to a vast collection of digitized content, including websites, books, audio recordings, videos, and software. On archive, we found datasets containing detailed information on 100 levels of *Super Mario Maker 2*, focusing on difficulty metrics, player performance, and engagement patterns.
- 4) **Super Mario Bros. Gameplay Dataset**: This dataset was created by Raphael C.P. and is freely available on GitHub¹¹. It contains valuable information about each *Super Mario Bros.* level, including a complete image of every level and 280 recorded play sessions. These sessions can be analyzed frame by frame at 60 frames per second.
- 5) **Platformer Experience Dataset (PED)** [8]: As cited in the Related Work section, The Platformer Experience Dataset (PED) is the first open-access game experience corpus, containing multiple modalities of user data from *Infinite Mario Bros.* players. It includes demographic information, gameplay data, and self-reported player experiences, such as engagement, frustration, and challenge levels. Additionally, the dataset offers visual recordings of players during gameplay, capturing facial expressions and head movements. While the dataset is player-oriented, it also contains valuable game-centric data that can be compiled and merged with data from other sources.

⁸https://www.nintendo.co.jp/n01/n64/software/nus_p_nsmj/index.html (in japanese) last accessed february 28th 2025

⁹<https://supermariomaker.nintendo.com/>

¹⁰<https://www.kaggle.com/datasets/leomauro/smmnet/data>

¹¹<https://github.com/rafaelcp/smbdataset>

B. Data Acquisition Protocol

The process of creating a standardized dataset for studying difficulty in platform game levels involves several key steps:

Data Manipulation with Python The first stage utilizes Python for data analysis. The *pandas* library¹² is employed for data manipulation and analysis, enabling efficient handling of structured data. *NumPy*¹³ is used for numerical computation, facilitating mathematical operations on large datasets. Data is processed from various formats, including CSV and JSON, allowing for flexibility in working with diverse file types. At the end, all the compiled data of the Ladder dataset is given in the form of a database and each level data could be downloaded in the form of a .csv file.

Web Scraping and Data Collection To gather relevant data, web scraping techniques are applied. HTTP requests are made using the Requests library to fetch data from websites. For parsing HTML content, we used the library *BeautifulSoup*¹⁴, providing a simple way to extract structured data from web pages. In cases where dynamic content needs to be captured, *Selenium*¹⁵ is employed to interact with websites and retrieve the necessary data. Additionally, web APIs are managed to access structured data from online sources. For instance, in the case of *Speedrun.com*, we used the Streamlit applications¹⁶.

Databases The Ladder dataset is organized and stored in databases for efficient querying and management. Fundamental SQL operations such as SELECT, INSERT, UPDATE, and DELETE are used to interact with relational databases. The SQLite database¹⁷, in combination with Python, offers a lightweight solution for storing and retrieving the dataset.

Data Cleaning and Preprocessing Before using the dataset for difficulty analysis, the dataset undergoes a rigorous cleaning and preprocessing stage. *pandas*¹⁸ is used for cleaning, ensuring that the data is accurate and usable. Missing values and outliers are addressed to improve data quality. Finally, the data is normalized and standardized to ensure consistency and comparability, which is crucial for further analysis and modeling.

This systematic approach ensures that the dataset is reliable, consistent, and ready for detailed analysis, providing valuable insights into the difficulty levels of platform game levels.

C. Structure of the LADDER Dataset

The dataset comprises a large number of elements that have been extracted and compiled from various sources in accordance with the acquisition protocol. For the *Super Mario Bros* game, the dataset includes the following elements:

- An image (PNG) of the map for each of the 32 levels.

¹²<https://pandas.pydata.org/>

¹³<https://numpy.org/>

¹⁴<https://speedrundata.streamlit.app/>

¹⁵<https://pypi.org/project/beautifulsoup4/>

¹⁶<https://speedrundata.streamlit.app/>

¹⁷<https://www.sqlite.org/>

¹⁸<https://pandas.pydata.org/>

- Data from 386,947 game episodes covering all levels of the original *Super Mario Bros*.
- Number of jumps and run actions (run, right, left).
- Total number of frames.
- User scores for each game episode.
- For each of the 32 levels: minimum score, maximum score, mean score, mean number of jumps, etc.
- For each of the 32 levels: number of enemies, types of enemies, number of bricks, number of holes, etc.
- For 280 specific game episodes, additional data is available:
 - A detailed log of each action performed in the episode with a timestamp (jump, click, start, select, direction).
 - Frame-by-frame recording of the episode at 60 frames per second (a total of 737,134 images).
 - A tag indicating whether the episode resulted in a failure or a success.

The data related to *Super Mario Bros* is the most complete and extensive in the dataset. Naturally, it is easier to find gaming session data for this game compared to the others. Additionally, we provide data for 50 *Super Mario 64* game sessions and 18 *Super Mario World*¹⁹ sessions.

For *Rockan*, we have data from 26 game sessions, capturing the same types of information as for the 280 Mario episodes (jump, click, start, select, direction, frame-by-frame recording, etc.). Similarly, for *Super Meat Boy*, the dataset currently includes 23 game sessions. All data for these two games were collected from speedrun.com.

To extract the data, we developed a script that analyzes video recordings of gameplay sessions. While the number of sessions for these two games is still limited, this is just the beginning. We plan to expand the dataset significantly to improve our analysis.

IV. ANALYZING DIFFICULTY USING THE LADDER DATASET

As stated in the introduction, our goal is to provide researchers with a tool that offers standardized and comparable data, enabling them to explore and refine concepts of difficulty in a reproducible, scalable, and insightful manner. Our work, like that of many researchers in the field [16]–[22], focuses on analyzing the difficulty of individual video game levels to develop tools that assist game designers. The difficulty of each level is a key factor in shaping the player experience, directly influencing engagement, satisfaction, and retention [5]. Accurately assessing difficulty is crucial not only for crafting well-balanced and compelling games but also for driving advancements in game design, player modeling, and adaptive gameplay systems [7].

In this section, we provide few examples of how the dataset can be used to assess and compare methods for analyzing game difficulty. As a first example, consider a team developing

¹⁹<https://www.nintendo.co.jp/n02/shvc/mw/index.html> (in japanese) last accessed february, 28th 2025

a machine learning algorithm to estimate the difficulty of platformer game levels based on players performance. This team could easily leverage our dataset to test their approach and determine the effective difficulty of each level using the player data recorded across 386,947 game episodes. For instance, they could generate a heat map for each level, visualizing difficulty intensity based on player performance metrics. Figure 3 illustrates such an analysis, presenting a basic heat map computation for the original *Super Mario Bros.* levels. This heat map is derived solely from the number of failures and retries recorded in each game episode, highlighting areas of higher difficulty.

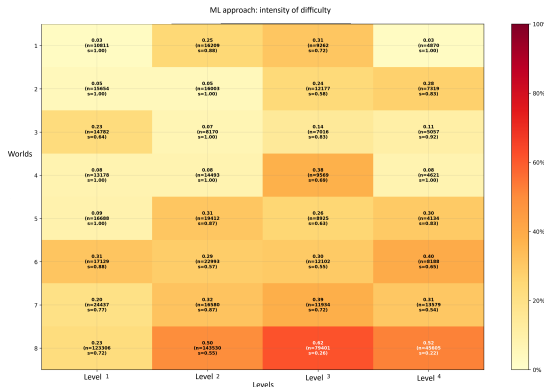


Fig. 3: Calculated intensity of difficulty for each Super Mario Bros. level using ML technique with linear regression

As a second example, we recently introduced a novel formal model designed to assist game designers in creating platformer-type games [13]. Our model is entirely game-centric, enabling the automated assessment of executive and motor difficulty within a level by analyzing its interactive elements—specifically, those that produce collisions. It identifies static hazards (e.g., pits, spikes) and simulates dynamic threats posed by enemy movements using a pheromone-based algorithm [23]. This approach eliminates the need for extensive playtesting by providing an objective measure of motor skill difficulty. It does not take into account the players performance. To illustrate the usefulness of our dataset, we also computed a comparative heat maps for each original *Super Mario Bros.* level using this non-ML approach, leveraging the same dataset. The results, shown in Figure 4, show how our model assesses difficulty without relying on machine learning.

We observe that the results of both approaches differ. However, the advantage of using a standardized dataset like LADDER is that it allows for an objective and consistent comparison of both results on a common factual basis.

As a comparative example, Figure 5 presents a comparison of the computed difficulty for each level using the ML-based approach based on player performance, our non-ML game-centric approach, and the designer-intended difficulty progres-

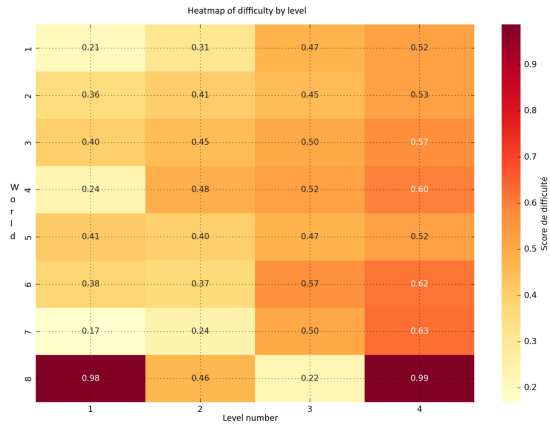


Fig. 4: Calculated intensity of difficulty for each Super Mario Bros level using the non-ML game-centric approach presented in [13]

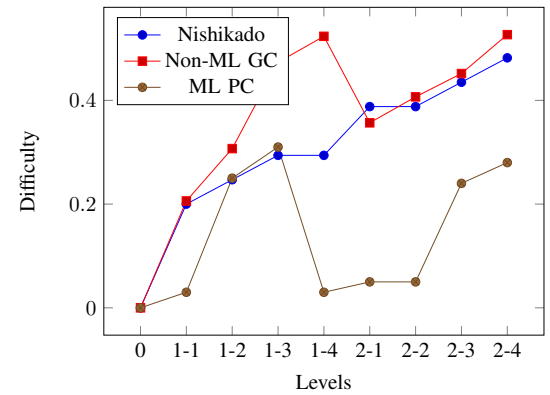


Fig. 5: Nishikado curve and computed difficulty curves. Nishikado: Nishikado progression curve; Non-ML GC: Non-ML game-centric approach; ML PC: ML player centric approach

sion of the game, following the Nishikado curve [24]. As shown, for the first eight levels of the game, our non-ML approach exhibits a good correlation of 0.63 with the Nishikado curve. For the second part of the game, our approach give less coherent results because the elements of late games (advanced enemies like Lakitu, moving platforms, etc.) as not yet been included in our model. In contrast, the ML-based difficulty evaluation, derived from player performance, appears significantly less correlated, with a correlation factor of 0.27. A possible explanation is that players attempting speedrun challenges tend to be highly skilled, making them unrepresentative of mainstream players. Additionally, the difficulty in the

specific context of speedrunning does not always align with the difficulty of simply completing a level. Ultimately, these results provide a solid scientific foundation for assessing the effectiveness of the proposed difficulty measures based on a standardized dataset.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced LADDER, a standardized dataset designed to analyze and evaluate level difficulty in video games. By providing a structured and reproducible framework for studying the relationship between game design elements and player behavior, LADDER addresses a critical gap in video game research [4]. Its provide a solid foundation for studying game difficulty and, more specifically, the difficulty of levels in a game. Our preliminary analyses demonstrate its potential for fostering new insights into difficulty assessment, level balancing, and player experience modeling. Ladder is freely available online and can be used by all the scientific community.

Despite these contributions, several avenues remain open for future research. First, while LADDER integrates objective performance metrics and level characteristics, expanding it to include richer physiological and behavioral data (e.g., real-time stress responses, cognitive load measures) could further enhance difficulty modeling. Additionally, increasing the dataset's diversity by incorporating more game genres and player demographics would improve its generalizability and applicability across different gaming contexts.

Another promising direction involves leveraging LADDER to develop models for automated difficulty estimation and adaptive game design. These tools could be incorporated as plug-in inside common game engines such as Unreal Engine²⁰. Game designers would then be able to automatically assess the difficulty of a level and understand the impact on difficulty, in real time, of a design choice.

Finally, fostering collaboration within the research community is essential to refining and expanding LADDER. We encourage researchers and developers to contribute additional data, refine methodologies, and explore novel applications, ultimately pushing the boundaries of standardized game difficulty analysis. By establishing a shared foundation for difficulty assessment, we hope LADDER will inspire further innovation in game design, leading to more engaging, balanced, and inclusive gaming experiences.

REFERENCES

- [1] N. Pessanha Santos, "The expansion of data science: Dataset standardization," *Standards*, vol. 3, no. 4, pp. 400–410, 2023.
- [2] H. F. Castro, J. S. Cardoso, and M. T. Andrade, "A Systematic Survey of ML Datasets for Prime CV Research Areas—Media and Metadata," *Data*, vol. 6, no. 2, pp. 1–85, January 2021.
- [3] P. D. Paraschos and D. E. Koulouriotis, "Game difficulty adaptation and experience personalization: A literature review," *International Journal of Human–Computer Interaction*, vol. 39, no. 1, pp. 1–22, 2023.
- [4] E. J. Pretty, R. Guarese, H. M. Fayek, and F. Zambetta, "Replicability and transparency for the creation of public human user video game datasets," in *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2023, pp. 74–81.
- [5] Y. F. H. T. B. B. S. L. M. R. J. Linard, "Automated difficulty assessment model for platformer games: A comprehensive approach," in *2023 IEEE CTSoc Gaming, Entertainment and Media conf.*, 2023, pp. 1–6.
- [6] A. Sarkar and S. Cooper, "Inferring and comparing game difficulty curves using player-vs-level match data," in *2019 IEEE Conference on Games (CoG)*, 2019, pp. 1–4.
- [7] P. B. Jeppe Theiss Kristensen, "Difficulty modelling in mobile puzzle games: An empirical study on different methods to combine player analytics and simulated data," *International Journal of Computer Games Technology*, p. 14 pages, 2024.
- [8] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis, "The platformer experience dataset," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 712–718.
- [9] H. Svoren, V. Thambawita, P. Halvorsen, P. Jakobsen, E. Garcia-Ceja, F. M. Noori, H. L. Hammer, M. Lux, M. A. Riegler, and S. A. Hicks, "Toadstool: a dataset for training emotional intelligent machines playing super mario bros," in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 309–314.
- [10] X. Zhao, S. Liu, X. Li, G. Li, and X. Xu, "Video coding tool analysis and dataset for gaming content," in *2021 Picture Coding Symposium (PCS)*, 2021, pp. 1–5.
- [11] A. D'Angelo, C. Di Sipio, C. Politowski, and R. Rubci, "Playmydata: a curated dataset of multi-platform video games," in *Proceedings of the 21st International Conference on Mining Software Repositories*, ser. MSR '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 525–529.
- [12] H. H. R. Yin, L.; Zhang, "Game difficulty prediction based on facial cues and game performance," *Applied Sciences*, vol. 14, p. 8778, 2024.
- [13] Y. Francillette, H. Tremblay, B. Bouchard, and B.-A. Menelas, "A comprehensive model of automated evaluation of difficulty in platformer games," *ACM Games*, vol. 3, no. 1, Jan. 2025. [Online]. Available: <https://doi.org/10.1145/3705013>
- [14] N. Barman, S. Zadootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "Gamingvideoseet: A dataset for gaming video streaming applications," in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2018, pp. 1–6.
- [15] L. M. P. Moraes and R. L. F. Cordeiro, "Smmnet: A social network of games dataset," in *Brazilian Symposium on Databases (SBBD) 2019 - Dataset Showcase Workshop (DSW)*. Ceará, Brazil: Brazilian Computer Society (SBC), oct 2019. [Online]. Available: <https://www.kaggle.com/leomauro/smmnet>
- [16] F. Mourato and M. Santos, "Measuring difficulty in platform videogames," 01 2010, pp. 173–180.
- [17] S. J. Alexander, J.t. and A. Oikonomou, "An investigation of the effects of game difficulty on player enjoyment," *Entertainment Computing*, vol. 4(1), pp. 53–62, 2013.
- [18] V. Palban, "Managing difficulty in games," *Game Design Development*, pp. 1–4, 2021.
- [19] J. T. Kristensen, A. Valdivia, and P. Burelli, "Statistical modelling of level difficulty in puzzle games," in *2021 IEEE Conference on Games (CoG)*. IEEE Press, 2021, p. 1–8.
- [20] J. Knorr and C. Vaz de Carvalho, "Using dynamic difficulty adjustment to improve the experience and train fps gamers," in *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, ser. TEEM'21. New York, NY, USA: Association for Computing Machinery, 2021, p. 195–200.
- [21] J. Zhang, "Directly controlling the perceived difficulty of a shooting game by the addition of fake enemy bullets," in *CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '21. New York, NY, USA: Association for Computing Machinery, 2021.
- [22] N. Fisher and A. K. Kulshreshth, "Exploring dynamic difficulty adjustment methods for video games," *Virtual Worlds*, vol. 3, no. 2, pp. 230–255, 2024.
- [23] B.-F. Florea, O. Grigore, and M. Datu, "Pheromone averaging exploration algorithm," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 617–622.
- [24] P. Holleman, *Reverse Design: Super Mario World*. CRC Press, 2018.

²⁰<https://www.unrealengine.com/en-US/unreal-engine-5>

CHAPITRE IV

BILAN SUR LES TRAVAUX ET SUR LA CONSTRUCTION DE LADDER

L'article **LADDER : Level Analysis Dataset for Difficulty Evaluation and Ranking** [2] constitue une contribution structurante dans le domaine de l'analyse de la difficulté vidéoludique. Contrairement aux approches antérieures qui s'appuient presque exclusivement sur des données centrées sur le joueur, ce travail propose un changement de perspective en intégrant de manière systématique les caractéristiques intrinsèques des niveaux et leur lien avec les performances observées. Ce positionnement méthodologique offre une base plus solide pour étudier la difficulté comme propriété du jeu lui-même, plutôt que comme simple reflet de la variabilité entre joueurs.

Un premier apport majeur de l'article est la construction d'un corpus standardisé, volumineux et diversifié, réunissant près de 387 000 sessions de jeu ainsi que des données structurelles détaillées pour les niveaux de plusieurs jeux emblématiques. La rigueur du protocole de collecte, la variété des sources mobilisées et la normalisation des données permettent de surmonter un obstacle récurrent dans la littérature : l'hétérogénéité et la fragmentation des données disponibles. En rendant l'ensemble librement accessible, l'article contribue également à améliorer la transparence et la reproductibilité des recherches en jeu vidéo, deux aspects souvent soulignés comme insuffisants dans les travaux antérieurs.

Le second apport réside dans la démonstration empirique de l'utilité du dataset pour comparer différentes approches d'évaluation de la difficulté. L'article illustre notamment comment LADDER peut servir à confronter une approche d'apprentissage automatique basée sur la performance des joueurs à un modèle formel entièrement centré jeu. Cette mise en perspective, rarement réalisable en l'absence de données communes, permet d'identifier les forces et limites de chaque méthode, de documenter leurs divergences, et, surtout, de promouvoir une évaluation plus stricte des outils d'analyse.

Enfin, l'article ouvre des perspectives concrètes pour le développement futur de modèles d'assistance à la conception de niveaux. La disponibilité d'un jeu de données structuré facilite l'émergence de pipelines d'analyse automatisée, que ce soit par apprentissage, par modélisation déterministe ou par hybridation des deux. Le potentiel d'intégration dans des moteurs de jeu ou des outils destinés aux designers témoigne de l'impact appliqué de la contribution : LADDER n'est pas seulement un corpus de recherche, mais un socle pour des outils de conception et de validation plus intelligents, plus cohérents et mieux ancrés dans la réalité des pratiques professionnelles.

Dans le contexte du présent mémoire, cet article occupe une place centrale puisqu'il fournit le référentiel méthodologique et empirique nécessaire pour étudier de manière comparative les différentes approches d'évaluation de la difficulté. Il sert également de fondation au développement de modèles complémentaires. Ainsi, la contribution de LADDER dépasse la simple mise à disposition de données : elle structure le terrain de recherche et ouvre la voie à une étude plus fine, plus cumulative et mieux outillée de la difficulté vidéoludique.

4.1 SÉLECTION DES SOURCES DE DONNÉES

Le processus de sélection des sources de données ayant mené à la construction du dataset LADDER, tel que décrit dans l'article, repose sur une approche méthodique visant à rassembler des données à la fois riches, variées et reproductibles. Nous avons privilégié des sources publiques et accessibles, permettant de combiner des informations objectives sur la performance des joueurs, des caractéristiques structurelles des niveaux, ainsi que des traces comportementales issues de sessions réelles. Parmi ces sources, *Speedrun.com*¹² occupe une place centrale, puisqu'elle offre un accès à des milliers d'enregistrements de parties jouées par des joueurs experts, accompagnés de leurs temps, de leurs scores et de nombreuses métadonnées associées. Cette plateforme constitue un vecteur incontournable pour obtenir des données réelles, détaillées et exploitables. D'autres sources communautaires ont également été retenues, notamment *Kaggle*¹³, qui propose plusieurs corpus structurés sur *Super Mario Maker*¹⁴, et *Archive.org*¹⁵, où ont été trouvées des archives de niveaux annotés pour *Super Mario Maker 2*¹⁶. Ces ressources ont été complétées par des datasets académiques déjà établis, comme le Platformer Experience Dataset (PED) [13], ainsi que par un corpus spécifique pour *Super Mario Bros* publié sur *GitHub*, contenant des cartes complètes des niveaux et des sessions enregistrées à 60 images par seconde. Ensemble, ces sources permettent de combiner données centrées jeu et données centrées usagé et de bâtir un ensemble de données complet, cohérent et diversifié.

12. <https://www.speedrun.com/>

13. <https://www.kaggle.com/>

14. https://en.wikipedia.org/wiki/Super_Mario_Maker

15. <https://www.archive.org/>

16. <https://www.nintendo.com/en-ca/store/products/super-mario-maker-2-switch/?srsltid=AfmB0oqR6fXsS6n9YNiPT0T7NZ37AwL9hrH0aGY0XmFrCLYhzI64MCB9>

4.2 PROTOCOLE EXPÉRIMENTAL

Le protocole expérimental adopté pour construire LADDER suit une chaîne de traitement, conçue pour assurer cohérence, qualité et reproductibilité. La collecte débute par un ensemble de techniques de *web scraping*, utilisant notamment les bibliothèques *Requests* et *BeautifulSoup* pour extraire des données statiques, et *Selenium* pour capturer des informations présentes dans des pages dynamiques ou nécessitant une interaction. Lorsque les données proviennent de vidéos, notamment pour *Rockman* ou *Super Meat Boy*¹⁷, un processus d'analyse image-par-image permet d'extraire automatiquement les actions du joueur, les séquences temporelles, les tentatives, les réussites ou les échecs. Une fois les données brutes capturées, elles sont structurées au sein d'une base *SQLite*, afin de faciliter les requêtes, la gestion des jeux et des niveaux, et l'organisation des métriques. Une étape essentielle de nettoyage suit cette intégration : elle consiste à normaliser les formats, à harmoniser les noms de niveaux, à éliminer les valeurs inconsistantes et à standardiser l'ensemble des unités temporelles ou statistiques. La dernière étape consiste à générer un fichier indépendant pour chaque niveau, généralement au format CSV, contenant à la fois les données brutes et les agrégats statistiques pertinents (moyennes, extrêmes, distributions de performances, etc.).

4.3 PARTICULARITÉS RECHNIQUE

Sur le plan technique, la construction de LADDER repose sur une boîte à outils entièrement basée sur *Python*. Les bibliothèques *pandas* et *NumPy* sont utilisées pour la manipulation de données, les matrices numériques et les opérations statistiques. Les outils de *scraping* assurent

17. https://en.wikipedia.org/wiki/Super_Meat_Boy

la récupération automatisée de l'information, tandis que *SQLite* constitue une solution légère, mais robuste, pour l'organisation interne du dataset. Les données produites prennent différentes formes : fichiers CSV pour les sessions, JSON pour certaines métadonnées, images PNG des niveaux, ainsi que des extraits vidéo lorsqu'ils sont nécessaires à l'annotation. Pour chaque niveau, LADDER regroupe un ensemble de caractéristiques structurelles (nombre d'ennemis, types d'obstacles, briques, trous, topologie du terrain), des mesures comportementales issues des joueurs (tentatives, sauts, morts, durée des sessions, logs temporels précis) et, lorsque disponible, des évaluations subjectives de la difficulté perçue. Cette organisation technique permet ensuite de réaliser différentes analyses : génération de cartes thermiques de difficulté, comparaison de modèles centrés joueur et centrés jeu, ou encore étude de la progression de difficulté des jeux à l'aide de courbes telles que la progression de Nishikado. Dans l'ensemble, le protocole et les choix techniques adoptés assurent la construction d'un dataset standardisé, complet et hautement exploitable pour les travaux de recherche sur la difficulté dans les jeux vidéo.

CHAPITRE V

CONCLUSION

Dans ce mémoire, nous avons présenté LADDER [2], un jeu de données standardisé conçu pour analyser et évaluer la difficulté des niveaux dans les jeux vidéo. En offrant un cadre structuré, reproductible et transparent pour l'étude des relations entre les éléments du design d'un niveau et le comportement des joueurs, LADDER répond à un besoin important identifié dans les travaux récents en recherche vidéoludique [13]. Il propose une base méthodologique solide pour l'étude de la difficulté vidéoludique et, plus précisément, pour l'analyse fine de la difficulté intrinsèque des niveaux. Grâce à la combinaison de métriques de performance objectives, de caractéristiques structurelles et d'annotations subjectives, le dataset permet de documenter de manière systématique les mécanismes qui rendent un niveau accessible, exigeant ou frustrant. Nos analyses préliminaires illustrent clairement le potentiel de LADDER pour soutenir de nouvelles avancées dans l'évaluation de la difficulté, l'équilibrage des niveaux et la modélisation de l'expérience joueur. LADDER est librement accessible en ligne, ce qui facilite son adoption par l'ensemble de la communauté scientifique et encourage la reproductibilité des travaux futurs.

Malgré cette contribution prometteuse au domaine, plusieurs pistes de recherche demeurent ouvertes pour la suite de ce projet. D'une part, bien que LADDER rassemble déjà des données

objectives et des caractéristiques structurelles des niveaux, son extension vers des données plus riches, notamment physiologiques (réponses au stress en temps réel, variations cardiaques) ou comportementales fines (mouvements oculaires, mesures de charge cognitive, analyse gestuelle), permettrait de raffiner encore davantage les modèles de difficulté et d'ouvrir vers un compréhension plus large de la difficulté perçues. Une telle intégration favoriserait l'émergence d'approches multimodales capables de mieux expliquer la perception de la difficulté, les variations individuelles entre joueurs ou l'impact de certains éléments du design sur la performance. D'autre part, accroître la diversité du corpus en y intégrant un éventail plus large de genres vidéoludiques (ex. : jeux de tir, puzzles complexes, rogue-lites, jeux de course) ainsi que des profils de joueurs plus variés améliorerait la généralisabilité des analyses et renforcerait la pertinence de LADDER dans des contextes ludiques très différents. Ces élargissements permettraient également de tester la robustesse des modèles d'analyse sur des topologies, des mécaniques et des dynamiques de jeu contrastées.

Un autre axe de développement prometteur concerne la création de modèles automatisés d'estimation de la difficulté et de systèmes d'adaptation dynamique du contenu. Grâce à l'architecture normalisée de LADDER, il devient envisageable de concevoir des outils capables d'estimer automatiquement la difficulté d'un niveau avant même qu'il ne soit testé par un joueur humain. Ces modèles pourraient ensuite être intégrés sous forme de plug-ins dans des moteurs de jeu populaires, tels qu'Unreal Engine¹⁸. Une telle intégration offrirait aux concepteurs de jeux des outils permettant d'évaluer, en temps réel, l'impact de chaque décision de design sur la difficulté, et ainsi d'ajuster leurs créations de manière plus fine et plus efficace. Cela ouvrirait la voie à des environnements de développement intelligents, où le moteur lui-même assiste le concepteur dans la recherche d'un équilibre optimal entre défi et accessibilité.

18. <https://www.unrealengine.com/en-US/unreal-engine-5>

En conclusion, la collaboration active de la communauté scientifique constitue un élément essentiel pour affiner, valider et enrichir LADDER. Nous encourageons les chercheurs, les développeurs et les concepteurs de jeux à contribuer à l'enrichissement du dataset, à proposer de nouvelles méthodes d'annotation, à partager leurs analyses comparatives ou à développer des applications inédites. La mise en commun de travaux réalisés à partir d'un référentiel standardisé permettra d'accélérer la maturité des approches d'analyse de la difficulté et de favoriser l'émergence de méthodologies communes. En établissant un socle partagé pour l'étude systématique de la difficulté, nous espérons que LADDER stimulera l'innovation en conception de jeux et encouragera la création d'expériences plus équilibrées, plus engageantes et plus inclusives pour tous les types de joueurs.

BIBLIOGRAPHIE

- [1] *Dynamic Time Warping*. Berlin, Heidelberg : Springer Berlin Heidelberg, 2007, p. 69–84. [En ligne]. Disponible : https://doi.org/10.1007/978-3-540-74048-3_4
- [2] Y. J.-E. Adjanohoun, Y. Francillette, H. Tremblay, et B. Bouchard, “Ladder : Level analysis dataset for difficulty evaluation and ranking,” dans *2025 IEEE Gaming, Entertainment, and Media Conference (GEM)*, 2025, p. 1–6.
- [3] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, et S. Möller, “Gamingvideaset : A dataset for gaming video streaming applications,” dans *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, 2018, p. 1–6.
- [4] H. F. Castro, J. S. Cardoso, et M. T. Andrade, “A Systematic Survey of ML Datasets for Prime CV Research Areas—Media and Metadata,” *Data*, vol. 6, n^o. 2, p. 1–85, January 2021.
- [5] A. D’Angelo, C. Di Sipio, C. Politowski, et R. Rubei, “Playmydata : a curated dataset of multi-platform video games,” dans *Proceedings of the 21st International Conference on Mining Software Repositories*, ser. MSR ’24. New York, NY, USA : Association for Computing Machinery, 2024, p. 525–529.
- [6] L. De Silva, T. Miyasato, et R. Nakatsu, “Facial emotion recognition using multi-modal information,” dans *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme : Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat., vol. 1, 1997, p. 397–401 vol.1.*
- [7] M. Fanourakis et G. Chanel, “Amucs : Affective multimodal counter-strike video game dataset,” *Scientific Data*, vol. 12, p. 1325, 2025, published 30 July 2025. [En ligne]. Disponible : <https://www.nature.com/articles/s41597-025-1325>
- [8] Y. Francillette, H. Tremblay, B. Bouchard, et B.-A. Menelas, “A comprehensive model of automated evaluation of difficulty in platformer games,” *ACM Games*, vol. 3, n^o. 1, Jan. 2025. [En ligne]. Disponible : <https://doi.org/10.1145/3705013>

- [9] D. Halbhuber, J. Höpfinger, V. Schwind, et N. Henze, “A dataset to investigate first-person shooter players,” dans *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY ’22. New York, NY, USA : Association for Computing Machinery, 2022, p. 51–56. [En ligne]. Disponible : <https://doi.org/10.1145/3505270.3558331>
- [10] P. B. Jeppe Theiss Kristensen, “Difficulty modelling in mobile puzzle games : An empirical study on different methods to combine player analytics and simulated data,” *International Journal of Computer Games Technology*, p. 14 pages, 2024.
- [11] Y. Jiang et L. Zheng, “Deep learning for video game genre classification,” *Multimedia Tools Appl.*, vol. 82, n^o. 14, p. 21085–21099, Feb. 2023. [En ligne]. Disponible : <https://doi.org/10.1007/s11042-023-14560-5>
- [12] E. Kanjo, L. Al-Husain, et A. Chamberlain, “Emotions in context : examining pervasive affective sensing systems, applications, and analyses,” *Personal and Ubiquitous Computing*, vol. 19, n^o. 7, p. 1197–1212, 2015. [En ligne]. Disponible : <https://doi.org/10.1007/s00779-015-0842-3>
- [13] K. Karpouzis, G. N. Yannakakis, N. Shaker, et S. Asteriadis, “The platformer experience dataset,” dans *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, p. 712–718.
- [14] A. Krizhevsky, I. Sutskever, et G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” dans *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS’12. Red Hook, NY, USA : Curran Associates Inc., 2012, p. 1097–1105.
- [15] Z. Lin, J. Gehring, V. Khalidov, et G. Synnaeve, “Stardata : A starcraft ai research dataset,” dans *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-17)*. AAAI Press, 2017, p. 50–57. [En ligne]. Disponible : <https://github.com/TorchCraft/StarData>
- [16] Y. F. H. T. B. B. S. L. M. R. J. Linard, “Automated difficulty assessment model for platformer games : A comprehensive approach,” dans *2023 IEEE CTSoc Gaming, Entertainment and Media conf.*, 2023, p. 1–6.
- [17] P. Lopes, G. N. Yannakakis, et A. Liapis, “Ranktrace : Relative and unbounded affect annotation,” dans *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, p. 158–163.
- [18] D. McDuff, E. Agtas, et R. E. Kaliouby, “A multimodal emotion sensing platform for building emotionally aware applications,” *IEEE Transactions on Affective Computing*, vol. 10, n^o. 4, p. 553–565, 2019.

- [19] W. Mellouk et W. Handouzi, “Facial emotion recognition using deep learning : review and insights,” *Procedia Computer Science*, vol. 175, p. 689–694, 2020, the 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC), The 15th International Conference on Future Networks and Communications (FNC), The 10th International Conference on Sustainable Energy Information Technology. [En ligne]. Disponible : <https://www.sciencedirect.com/science/article/pii/S1877050920318019>
- [20] F. P. Miller, A. F. Vandome, et J. McBrewhster, *Levenshtein Distance : Information Theory, Computer Science, String Metric, Damerau–Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press, 2009.
- [21] S. Ontañón, G. Synnaeve, A. Uriarte, F. Richoux, D. Churchill, et M. Preuss, “A survey of real-time strategy game ai research and competition in starcraft,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, n^o. 4, p. 293–311, 2013.
- [22] P. D. Paraschos et D. E. Koulouriotis, “Game difficulty adaptation and experience personalization : A literature review,” *International Journal of Human–Computer Interaction*, vol. 39, n^o. 1, p. 1–22, 2023.
- [23] P. Peng, Q. Tian, Y. Zhang, et et al., “Multiagent bidirectionally-coordinated networks for learning to play starcraft combat games,” dans *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, p. 3617–3623.
- [24] N. Pessanha Santos, “The expansion of data science : Dataset standardization,” *Standards*, vol. 3, n^o. 4, p. 400–410, 2023.
- [25] E. J. Pretty, R. Guarese, H. M. Fayek, et F. Zambetta, “Replicability and transparency for the creation of public human user video game datasets,” dans *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2023, p. 74–81.
- [26] A. Sarkar et S. Cooper, “Inferring and comparing game difficulty curves using player-vs-level match data,” dans *2019 IEEE Conference on Games (CoG)*, 2019, p. 1–4.
- [27] H. Svoren, V. Thambawita, P. Halvorsen, P. Jakobsen, E. Garcia-Ceja, F. M. Noori, H. L. Hammer, M. Lux, M. A. Riegler, et S. A. Hicks, “Toadstool : a dataset for training emotional intelligent machines playing super mario bros,” dans *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys ’20. New York, NY, USA : Association for Computing Machinery, 2020, p. 309–314.
- [28] G. Synnaeve et P. Bessiere, “A bayesian model for opening prediction in rts games with application to starcraft,” dans *Proceedings of IEEE Conference on Computational Intelligence and Games (CIG)*, 2011, p. 1–8.
- [29] N. Usunier, G. Synnaeve, Z. Lin, et S. Chintala, “Episodic exploration for deep deterministic policies : An application to starcraft micromangement tasks,” dans

Proceedings of the International Conference on Learning Representations (ICLR), 2017.
[En ligne]. Disponible : <https://openreview.net/forum?id=SJ0kZsKxg>

- [30] H. H. R. Yin, L. ; Zhang, “Game difficulty prediction based on facial cues and game performance,” *Applied Sciences*, vol. 14, p. 8778, 2024.
- [31] X. Zhao, S. Liu, X. Li, G. Li, et X. Xu, “Video coding tool analysis and dataset for gaming content,” dans *2021 Picture Coding Symposium (PCS)*, 2021, p. 1–5.