SPELLS — A Representation Learning Approach to Latent Power Prediction for Magic: The Gathering

MSc Research Project
MSc in Artificial Intelligence

Sam Conneely
Student IDs: X24157651

School of Computing
National College of Ireland

Supervisor: Dr. Muslim Jameel Syed

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sam Conneely
Student IDs:	X24157651
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr. Muslim Jameel Syed
Submission Due Date:	11/12/2025
Project Title:	SPELLS — A Representation Learning Approach to Latent Power Prediction for Magic: The Gathering
Word Count:	15650
Page Count:	41

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sam Conneely
Date:	11th December 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgment Supplement

MSc Research Project

SPELLS — A Representation Learning Approach to Latent Power Prediction for Magic: The Gathering

Student Name:	Sam Conneely
Student ID:	X24157651
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI ACKNOWLEDGMENT

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to Tool
NA	NA	

DESCRIPTION OF AI USAGE

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used.**

Tool Name
Description of Use
Sample Prompt

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

Abstract

This thesis introduces SPELLS (Semi-Supervised Power Estimation and Label Learning System), a multimodal learning framework for estimating the latent power level of Magic: The Gathering cards. With no standardised power metrics and manual annotation of 26,795 cards being impractical, SPELLS uses a small human labelled subset ($< 0.4\%$ of cards) with multimodal embeddings derived from the oracle text, artwork, and structured metadata encoded on the cards. Anchor-graph label propagation generates pseudo-labels across the entire dataset, allowing for the training of regression models including a deep neural network, XGBoost and Support Vector Regressor.

Extensive preprocessing was done including the normalisation of oracle text as well as the generation of Sentence-BERT embeddings (384-dimensional) for keyword and type information, ResNet50 visual features (reduced to 256-dimensions via PCA), and engineered structural features. An active learning stage with human-in-the-loop correction was done iteratively to address outliers in the label propagation.

Evaluation on the manually labelled anchors shows that all models achieve Mean Absolute Error of < 1.0 on the 1-10 power scale. Further testing on cluster representative cards from the full manifold structure reveals that the DNN ($R^2 = 0.87$, Spearman=0.97) and XGBoost ($R^2 = 0.83$, MAE=0.19, Spearman=0.97) successfully learned the latent power, while Support Vector Regressor failed to maintain stable predictions on fused embeddings using all three modalities (MAE= 1.26 $R^2 = -0.28$).

This work proves that semi-supervised multimodal learning can effectively estimate subjective latent features in complex game systems with minimal expert annotation. SPELLS proves a foundation for future research on format specific modelling and transferability to other trading card games and domains requiring subjective quality assessment from heterogenous features.

Keywords: Semi-supervised learning, multimodal fusion, label propagation, trading card games, Magic: The Gathering power estimation

1 Introduction

Trading card games such as Magic: The Gathering (MTG) represent some of the most complex imperfect-information game systems in existence, with over 27,000 unique cards at the time of writing, each containing detailed and complex combinations of textual abilities, numerical statistics, and visual designs elements. Since its release in 1993, Magic: The Gathering has evolved into a multi-billions dollar franchise with millions of active players worldwide and a competitive esports scene and large prize pools. The card game's complexity comes not only from its enormous card pool, but from the number of possible interactions between cards - a single card's utility can vary dramatically depending on the rest of the cards within a player's deck, the current competitive metagame, and the format being played.

Central to both casual and competitive play is the concept of card power. This is an intrinsic measure of a card's strength relative to other cards within the game. It is critical that players understand a card's power for deck-building purposes. Game designers at Wizards of the Coast need to carefully balance new card design and releases to maintain competitive play, while avoiding power creep that often breaks game mechanics or causes some cards to become totally irrelevant. Players need to evaluate cards when building decks from their collections, trading with other players, or when purchasing new cards.

Tournament organisers must assess whether certain cards under specific conditions can cause infinite loops or break intended mechanics when played in a certain order and whether these warrant restrictions or banning. Despite the clear need, no objective, quantifiable measure of individual card power exists.

Unlike video games where character or item strength can be derived from win-rate statistics, trading card game power assessment faces a different challenge. First, cards are almost never played in isolation, their effectiveness depends on synergies with other cards, making individual contribution difficult to interpret and unlikely within a game scenario. Second, power varies greatly across different game formats: a card that is seeing lots of play, and is dominant in Standard (60-card decks, recent sets) may not be utilised as often in Commander (100-card singleton decks, nearly all cards are legal). Third, and most importantly, explicit power level labels are non-existent within the game itself. Each card displays its mana cost, card type, abilities, and combat statistics, as can be seen in figure 1, but nowhere on the card is an explicit power level stat encoded. This lack of a definitive label makes traditional supervised machine learning approaches impractical, as manually labelling 27,000+ cards would require a large amount of time and the costs to do so would far outweigh the benefits. Manual annotation may still reflect subjective human judgements over an objective truth based on the features of individual cards.

Figure 1: Informational breakdown of the sections on a Magic card: *The Ur-Dragon*

This research introduces SPELLS (Semi-Supervised Power Estimation and Label Learning System), a novel multimodal deep learning framework designed to predict the latent power level of Magic: The Gathering cards using minimal human annotation. SPELLS combines three modalities of card information into a unified representation: (1)

textual embeddings from card abilities and rules text using SBERT transformers, (2) visual features extracted from card artwork using ResNest50 convolutional neural networks, and (3) structured numerical and categorical features such as mana cost, power, toughness, and card type. Through the use of semi-supervised learning using graph-based label propagation, SPELLS generates power level labels from a small seed sample of manually annotated cards (fewer than 100 examples) across the entire dataset, then trains a deep neural network to predict the continuous power levels of unseen MTG cards.

The choice of Magic: The Gathering as a research domain is not an accident and is well-justified. MTG has been proven to be Turing-complete, meaning arbitrary computations can be embedded within the game states, establishing it as a system of genuine computational complexity [1]. The game serves as an ideal testbed for artificial intelligence research due to its combination of imperfect information (opponent's hands and deck contents are hidden), stochastic elements (randomised deck shuffling), and strategic depth requiring long-term planning. Previous AI research has successfully used imperfect-information games as benchmarks for advancing machine learning techniques, from poker [2] and Dota 2 [3], to StarCraft II [4]. Magic: The Gathering extends these domains by adding linguistic complexity (cards with verbose rules text requiring semantic understanding) and visual design complexity (artwork that correlates with card rarity and mechanical themes).

This work addresses the gap between the research into AI in games and multimodal machine learning. While substantial prior work has explored AI for playing Magic: The Gathering using Monte Carlo Tree Search [5, 6] and reinforcement learning, other research has investigated card drafting strategies [7, 8] and procedural card generation [9], no prior work has systematically tackled the problem of intrinsic power estimation from card features alone. Existing tools that claim to evaluate the power level of card typically assess entire decks, or a collection of cards, rather than individual cards and often rely on community-sourced data or competitive win rates over solely learning from card features and information encoded directly onto each card.

The contributions of this research extend beyond that of just the video game industry, and into other specific domains where machine learning technologies can be applied. SPELLS demonstrates that semi-supervised learning through the use of label propagation, a technique first proposed in 2002 by Zhu and Ghahramani [10], remains highly effective when combined with multimodal representations. The sparse-label approach addressed by SPELLS (less than 0.4% of examples labelled) is common across many real-world applications where expert annotation is expensive and time-consuming (medical diagnosis, content moderation, and document classification). By showing meaningful predications can be achieved with minimal supervision on a complex, subjective task, this work provides both methodological insights and builds a reusable framework for similar problems and future research.

Furthermore, this research contributes to the growing field of multimodal representation learning by demonstrating effective fusion of heterogeneous feature types - pretrained language model embeddings, convolutional neural network features and structured tabular data, within a semi-supervised learning pipeline. Recent surveys on multimodal learning [11, 12, 13] highlight the importance of finding effective fusions strategies that allow different modalities to complement each other. SPELLS employs early fusion through concatenation, allowing the downstream neural network to learn cross-modal interactions that may contribute to the prediction of latent card power.

The remaining sections of this document are structured as follows: Section 2 provides a comprehensive literature review covering related work in game AI, semi-supervised learning, label propagation, multimodal deep learning, and prior research focusing on Magic: The Gathering. Section 3 details the methodology, including data collection from the Scryfall API, feature engineering for each modality, the label propagation algorithm, and the neural network architecture. It also presents the experimental setup, including the manual labelling process and how the data has been stratified for label propagation and evaluation. Section 4 analyses the results, comparing model performance across different architectures including the feedforward neural network, XGBoost and Support Vector Regressor. Both pillars of evaluation are explained and discussed in detail. Section 5 discusses the implications of this work for game design, AI research, and transferable applications across other domains and industries. The limitations are also discussed throughout this section. Finally, section 6 ends the paper with the conclusions and focuses on directions for future research into this topic.

Figure 2: Examples of different card types and colours that can be played in Magic: The Gathering

2 Background and Related Work

Games have long been used as controlled, rules-based research environments for artificial intelligence, and many papers have been published on the topics of AI in perfect-information board games such as Chess [14], Go [15], and imperfect-information games like Poker [2]. Research into artificial intelligence through the medium of video games is also not a new concept [16], with a lot of research having been done with AI and games like Dota 2 and StarCraft II [3, 4].

Building on this tradition, SPELLS seeks to build a framework for the investigation of imperfect-information card games as rich multimodal datasets for deep representational learning. Magic: The Gathering offers an ideal environment for conducting research into this topic, with many thousands of detailed cards combining textual, visual and structured features [17] that can be used to determine the latent power of a specific card. Unlike previous research that has focused mainly on deckbuilding or drafting using complete and labelled data [18, 19, 20], SPELLS makes use of semi-supervised learning and deep multimodal fusion to infer latent power levels from a limited set of labelled examples. This project bridges prior research in AI for games, utilising graph-based label-propagation and multimodal representational learning.

2.1 Magic: The Gathering as a Research Domain

There are many reasons as to why Magic: The Gathering makes an excellent and challenging testbed for Artificial Intelligence research. The game itself is an extremely computational and mechanically complex one, with imperfect-information and a stochastic nature through randomly shuffled player decks [21, 5]. These attributes contribute to the game of Magic being a difficult to predict and equally as complex to solve as the Halting Problem, as explored by [1] in their paper *Magic: The Gathering is Turing Complete*.

There is a vast quantity of cards available to a player when constructing a deck and during a game. Within the *standard* format of Magic a player makes use of a deck with a minimum of 60 cards, for other formats, such as *Commander*, players must construct a singleton deck of 100 cards. This large search space leads to an enormous potential of different decks being built and used during play [17]. Throughout a game players make sequential decisions leading to choices that can be mapped structurally using a tree structure of nodes and edges. Complex games like Chess and Go have large branching factors, ranging from between 30 and 300 requiring advanced algorithms to model optimal moves, Magic has a much greater number than this, often growing upwards of one million [21]. With such a large branching factor, and the stochastic nature of the game, the mapping of possible decisions becomes unmanageable for search algorithms like Monte Carlo Tree Search [5].

Monte Carlo Tree Search algorithm is a technique often used with games such as Go, Chess, and more recently Hearthstone and Magic: The Gathering [22, 21] to simulate a games play space iteratively.

Unlike many other card games, MTG does not make use of a standard playing deck. These specifically designed cards contain unique and detailed rules about abilities and how they change the mechanics of the game. Many cards have effects on how cards interact with one another, dramatically steering the game in different ways. These rulings gives rise to a very rich set of ways that the game can unfold during a player's turn. Although not all card actions are considered strong, each card may have a different usefulness given a specific game state [21].

Although search-based approaches such as MCTS are very effective at simulating these game states, the astronomically large number of unique cards, hidden-information, and shuffling of decks, renders these strategies computationally impractical at scale. Throughout the papers *Ensemble Determinization in Monte Carlo Tree Search for the Imperfect Information Card Game Magic: The Gathering* [21] and *Monte Carlo search applied to card selection in Magic: The Gathering* [5], the authors of both papers, Cowling et al, and Ward et al, make use of advanced techniques and approaches to Monte Carlo Tree Search applied to Magic: The Gathering.

Although the papers saw an improvement in game AI playing strength when using MCTS, ultimately they both came to the conclusion that even when advanced techniques such as ensemble determinisation are applied, MCTS can only be used effectively within a simplified version of the game. As MCTS struggles to simulate the game space effectively, this reinforces the idea that systems such as SPELLS, which make use of representational learning methods may be useful for mapping the complex relationships between cards without the need to simulate every game state possibility.

2.2 Multimodal Representational Learning

Representational learning is an important element in machine learning methods, where the mapping of complex relationships between data are essential. The use of vector embeddings allows for the transforming of unstructured data such as text or images into numerical vectors that can then be processed by a neural network[23]. These vector embeddings then allow for the semantic meaning of data to be extracted by mapping similar elements closer to one another within the dimensional vector space. SPELLS makes use of multimodal representational learning as an approach to mapping complex relationships between cards based on their different features. This method is utilised as each card contains a number of different features, across different modalities, including numerical and categorical values, textual data with rich semantic meaning and images [17].

In order to have SPELLS work as a framework for research into artificial intelligence, the system must be able to consistently map the fused embedding representation space. Cards within local clusters of the manifold must receive similar scores, showing a smoothness of the learned feature space. The need to understand the dataset globally and locally with little variance across cards and sets is not a new idea, and is explored in detail throughout *Learning With Generalised Card Representations for "Magic: The Gathering"* [17]. The research done throughout this paper shows that machine learning models can learn generalised representations of cards through multiple sources of information. The learned embeddings generated were tested across multiple predictive tasks, with a hybrid representation of text, images, and structured data having the best performance on unseen cards. This result directly supports the approach taken by SPELLS and builds the multimodal architecture based on textual, visual, and structural features in order to best map the complex relationships between cards.

Multimodal representational learning focuses on the combination of different data from multiple sources such as text, images and structured numerical data to build a unified understanding of an object. The goal of this approach is to combine different views of data and create a more detailed representation containing all of the aspects of it. In the case of data used in SPELLS, information from images cannot be learned from categorical or numeric values. Through the use of different modalities deep learning has advanced the field by allowing neural networks to be trained on separate levels of data, giving them the ability to extract complex and context rich features [11].

Recent research into multimodal learning shows that advances made in areas such as computer vision, have helped with tasks such as image captioning and multimodal data retrieval. This research of multimodal systems has also contributed to the increasing ability of deep learning models to generalise and learn a broader knowledge across multiple topics [12]. Building on previous research, SPELLS extends multimodal learning into the area of semi-supervised learning by showing that even with limited labelled data detailed multimodal representations can be learned through structured fusion and graph-based label propagation.

The SPELLS architecture makes use of three key features, essential to high-quality multimodal representational learning:

- **Multimodal Feature Information**: Each of the unique modalities offers different but complementary information. In the context of Magic: The Gathering cards, textual descriptions (captured using SBERT embeddings) encode each card's abilities and gameplay mechanics. The visual data (extracted using ResNet50) contains colour identity and details correlated with rarity and power. Finally, structured numerical features provide measurable information such as damage output and defence statistics. When combined or fused together these modalities provide a complex, but rich representation of a card's power, when comparing this fused representation with that of a single modality, the context and information that can be extracted is far greater.

- **Sparse-Label Scalability**: SPELLS achieves predictive performance using fewer than 300 labelled examples across a dataset of more than 27,000 cards. This proves that the model has the ability to return effective results in sparse-label environments. The SPELLS architecture makes use of unlabelled examples through label propagation, allowing for the discovery of hidden relationships between card features even in situations where labelled data does not exist.

- **Graph-Based Structural Smoothness**: The SPELLS architecture makes use graph-based structural smoothness to stabilise learning with minimal labelled samples. Within the multimodal space cards with similar text, artwork and structural features form clusters, referred to as the fused embedding manifold. Smoothness is ensured by projecting the cards into a low-dimensional space and providing anchor points to regularise the graph structure.

Through the integration of structured knowledge, such as mana costs and card types, the ability of SPELLS is improved and aligns with recent research emphasising the need for explainability and fairness within the model decision making process [24]. SPELLS demonstrates this approach by fusing explicit domain knowledge with learned multimodal representations from text and images, improving the effectiveness of semi-supervised multimodal learning models.

2.3 Semi-Supervised Learning and Label Propagation

Semi-supervised learning is a technique that combines small samples of labelled data with a larger pool of unlabelled data in order to improve model performance. The approach is very important in areas where obtaining high-quality labelled datasets is difficult, expensive and time consuming. Traditional SSL methods usually focus on single-modality data like text or images. Semi-supervised learning can reduce the time taken and the cost of this manual labelling process, making use of pseudo labels through the similarity of features within the data [25]. However, many real-world problems are multimodal and involve more than one singular source of data across modalities such as images, textual descriptions and numerical features. SPELLS contributes to this field of research by demonstrating that semi-supervised learning can be extended to multimodal settings, while making use of these efficiency gains by utilising label propagation and deep neural networks.

One of the challenges of effective SSL is ensuring that similar clusters located close together within a feature space are labelled similarly. This principle is known as the manifold assumption [26]. Classic label propagation methods make use of this principle, but are applied to systems using transduction learning, where the model may only make predictions on the data seen during the training process. SPELLS uses a hybrid approach, utilising label propagation as a transductive pre-labelling step and then trains an inductive deep learning model which allows the system to make effective predictions on Magic: The Gathering cards.

By combining multiple embeddings from different modalities, such as SBERT for textual data, ResNet50 for visual features, and the structured numerical data, SPELLS builds a graph-based representation to capture complex relationships across all features. Having a single unified representation of the data allows label propagation to generate pseudo labels more effectively, improving the model's understanding of the entire dataset. SPELLS helps to bridge the gap between graph-based semi-supervised learning and multimodal deep learning, showing that label propagation is not only useful in neural network pipelines, but it can also improve results when labelled data is not widely available. This research shows that semi-supervised multimodal learning and the use of unlabelled data can be used to more efficiently train models with limited human annotation required.

2.4 Prior Work on Magic: The Gathering

Magic: The Gathering has been the subject of much research in recent years, with a variety of different areas of the game being explored, such as gameplay simulation, drafting strategies and even card generation - none have worked on the latent power estimation of card-based features.

Ward and Cowling applied the Monte Carlo Tree Search algorithm to Magic: The Gathering, demonstrating that it could be used to improve how well artificial intelligence played the game in a number of limited situation [5]. Later Cowling et al. extended the research into MCTS with ensemble learning. This work looked to handle imperfect information through the sampling of possible game states [21]. They succeeded in mapping in-game decisions in great detail, but despite these improvements, even with ensemble approaches, MCTS struggled with the vast complexity of Magic. This research although gameplay focused, differs from that of SPELLS. They looked into simulating the gameplay of Magic, and evaluated cards based on specific game state contexts. This research using MCTS was targeted towards whether or not certain cards should be played given the state of the game board, not latent power.

More recently, Bertram et al. in *Predicting Human Card Selection in Magic: The Gathering with Contextual Preference Ranking* have looked into the area of drafting within Magic: The Gathering, and how to use Siamese neural networks for the contextual ranking of cards. They were able to achieve an improvement in accuracy of around 56% over baseline models [7]. Their model was successful in learning the preferred picks of players, but the approach was limited to a narrow context of around 45 cards. Ward et al. explored reinforcement learning in *AI solutions for drafting in Magic: the Gathering*, again focusing on the drafting element of the game [19]. This aspect of the game alone, is a vast and detailed topic that has seen much research. This work saw the training of agents through self-play, with the goal of building competitive decks. While their work showed that AI could indeed learn how to correctly evaluate cards for the constructing of decks, the number of simulated drafts per card required was in the thousands, and this work did not focus on the estimation of card power. SPELLS does not evaluate cards within the context of drafting and learns their intrinsic power from features alone instead of gameplay outcomes.

Summerville and Mateas developed *Mystical Tutor* as a method of using procedural content generation to create new Magic cards from partial prompts (e.g., "3-cost blue instant")[9]. Their research made use of a sequence-to-sequence model to map partial card descriptions into complete card text. *Mystical Tutor* addresses the inverse problem to SPELLS: card generation, rather than evaluation. The model could learn what makes a valid card, but it cannot predict whether or not that card will be strong or weak during gameplay. Although they were successful in training the model to learn about the nuanced design cues of card art in MTG, they found that human oversight was still essential in the crafting of new card art. This research is an example of how such systems can be built as creative assistants for designers and artists alike.

Perhaps the most relevant work relating to SPELLS is the research done by Bertram et al. in *Learning With Generalised Card Representations for "Magic: The Gathering"* [17]. They showed that a hybrid approach using multimodal features outperformed single modality methods for supervised classification tasks like predicting a cards rarity using textual, visual and structural features. Their experiments with over 20,000 cards proved that textual embeddings encode mechanics, image features correlate with rarity, and structural features do provide explicit statistics of cards. However, where Bertram et al. focused solely on supervised learning using complete labels, with fully annotated samples for all training methods. SPELLS extends this work through the use multimodal representations and semi-supervised learning, with less than 0.4% of the data having ground-truth labels, the remaining percentage being generated through label propagation.

Work	Task	Method	Data Scale	Labels
Ward & Cowling (2009)	Gameplay decisions	MCTS	Game states	None
Cowling et al. (2012)	Imperfect info play	Ensemble MCTS	Game states	None
Bertram et al. (2021)	Draft prediction	Siamese network	45 cards/draft	Preferences
Ward et al. (2021)	Draft RL	Self-play	1000s drafts	Win/loss
Summerville (2021)	Card generation	Seq-to-seq	20K cards	Full text
Bertram et al. (2024)	Type/rarity predict	Multimodal fusion	20K cards	Labelled
SPELLS (2025)	**Power estimation**	**LP + DNN**	**27K cards**	**<0.4%**

Table 1: Comparison of SPELLS with Prior MTG AI Research

2.5 Research gap and Positioning

This section gives a breakdown of the areas that SPELLS could be applied to and other industries where the research done throughout this project could benefit from, particularly areas of multimodal research is new. Areas where the manual annotation of labels is time consuming and expensive are also mentioned within this section, as the research into semi-supervised learning is applicable to such industries.

Pre-Release Card Evaluation and Balance Testing

Trading card game companies like Wizards of the Coast often face problems with the creation and release of new content throughout the year. It is important for all new content, in the form of sets containing possibly hundreds of cards, to be well tested and balanced before being released to the public. Balancing in this case is a term used to describe the *"fairness"* of a game's mechanics. Often the term balancing is used when changes must be made to a certain element of a game if it is deemed too strong, or too weak, and often comes with unforeseen consequences, like a particular card causing powerful effects far too early into a game.

Artificial intelligence and game design have been closely linked since the early concepts of AI, but it has most recently found a home with things like: pathfinding for agents within video games or for the generation of procedural content. PCG or procedural content generation, has been used within the gaming industry for a long time, and is often used to create vast landscapes, interesting character stories or even the randomisation of weapon stats throughout a playthrough. AI is now expanding the concept of PCG and helping in the designing of games in other ways. AI can make use of data to generate many things and is now being used to elevate PCG to new levels by creating complex content tailored to specific players, adding a personalised depth that adapt to a player's style or skill level[1]. It is clear that AI is being used throughout many different industries, and game design is no different. This is where SPELLS fits in as a tool that could be leveraged to help with the generation and testing of new cards.

With Wizards of the Coast having seven new sets of Magic: The Gathering cards released in 2025, each containing roughly 200-300 new cards, the need for some level of automated balancing is essential[2]. Traditional testing methods are time consuming and cannot possibly cover all situations that may arise. Artificial intelligence can be used to simulate large volumes of play states and card interactions to efficiently identify issues and optimise content for a more enjoyable gameplay experience.

SPELLS helps to address the need for balancing by providing a power estimate based on the card's features - Oracle text, mechanics and card images - without the need for extensive playtesting. The SPELLS system could potentially assist designers at Wizards of the Coast, in the balancing of new cards by enabling:

- The balancing of mana costs for new cards against predicted power levels across card types.
- Keep track of power creep across sets by maintaining historical power assessments.
- Reduce the development costs by automating card evaluation.
- Identify outliers before the printing of any cards, reducing the need for post-release bans across many different MTG formats.

Although procedural content generation is not a new concept, and has existed in video games since the 1980s, with the addition of AI, PCG once again has the ability to reduce production costs, maximise the output of designers and developers, and continue to increase the replayability of games [27]. SPELLS works to extend this approach to trading card games, with a current focus on Magic: The Gathering by showing that artificial intelligence can assist in the creative process of game design by elevating human judgement instead of replacing it.

Beyond that of being of tool specific for designers, SPELLS could also be used in a more player oriented way to assist fans of Magic: The Gathering by allowing them access to:

- An intelligent deck-building assistant that will allow players to rank cards based on their power levels, helping to find potential synergies, and assist with the scaling of decks for competitive tournaments.
- An educational tool for new players learning about card abilities and how to better evaluate them.

[1] https://www.lenovo.com/us/en/gaming/ai-in-gaming/

[2] https://articles.starcitygames.com/magic-the-gathering/magic-the-gathering-full-set-release-schedule-for-2025/

- A collection management system that helps a player to better identify high-value cards for tournament play or for trading.

Recommendation Systems and Content Evaluation

Multimodal Recommender Systems have seen increased attention recently across both academia and industry as multimodal features help to address the issues of data sparsity while enabling systems to learn from more than one data source and type [28]. The methodology developed for SPELLS directly transfers to recommendation systems where e-commerce platforms require evaluation of product quality from images, textual information and specifications with limited user reviews. The SPELLS architecture also expands to content moderation systems that need to assess the severity of a user's textual comments or other information generated by a user in the form of images or videos. Educational platforms could also build upon the architecture for the estimation of difficulty for multimodal learning materials.

Cross Domain Relevance and Transferability

The issues of manual annotation and label sparsity addressed by SPELLS are not unique to MTG card power estimation. They are very relevant to the healthcare industry and across other expert lead industries where domain specific knowledge is key to understanding processes and operations.

- **Medical Imaging and Diagnostics**: Systems responsible for the prediction of certain diseases and their severity based on multimodal data like patient scans, doctor's notes or lab results. Processes where a medical professional is expensive and requires expert knowledge.
- **Document Classification**: Legal or financial documentation for detailed risk analysis through the use of multiple data sources.
- **Real Estate Valuation**: The valuation of properties from images, descriptions and structured features such as: size of the property, location data or number of rooms.

Deep neural networks repeatedly show improved performance in applications across many industries as they are scalable and can learn complex relationships from data. Their reliance on large volumes of high-quality labelled data limits their deployment in data-scarce and specialised domains [25]. SPELLS shows that domain specific predictions can be achieved with minimal expert labelling, reducing the need for costly manual annotation. This establishes a proof of concept for applying deep learning in data-scarce settings, creating a blueprint for scalable semi-supervised learning approaches in other multimodal industries. By combining graph-based propagation with multimodal fusion, SPELLS demonstrates that it is possible to build a DNN model for a domain where manual label costs have are high and restrictive to the adoption of artificial intelligence.

Benchmark for Semi-Supervised Multimodal Learning Methods

The development of reliable benchmarks is essential for the advancing of machine learning and artificial intelligence research. Benchmarks allow researchers to systematically compare different methods, identify strengths and weaknesses, and measure progress in a consistent and reproducible way. In relation to semi-supervised multimodal learning, there remains a shortage of publicly available, large-scale benchmarks that combine textual, visual and structured data under sparse label conditions. SPELLS addresses this gap by introducing a complex, real-world testbed for the evaluation of multimodal semi-supervised methods across multiple dimensions.

Graph-based semi-supervised learning often face issues with scaling as building the graph and label propagation can quickly become computationally expensive when applied to large datasets, although recent advances show that these methods can be made to be more efficient by building bipartite graphs. These graphs are structures that connect a smaller set of nodes called anchors to the raw data points while preserving the manifold of the data. This new approach captures the essential relationships between data points, while reducing the computational complexity [29]. SPELLS applies anchor points to reduce the computational load of label propagation in order to handle over 27,000 data samples efficiently, making this both practical and relevant as a research framework.

As a benchmark, SPELLS provides a platform for evaluating several components of multimodal semi-supervised learning systems, including:

- **Label Propagation Algorithms**: When comparing traditional methods for label propagation against newer, more advanced techniques such as adaptive anchor-based models and graph neural network (GNN) approaches that dynamically learn the appropriate propagation weights.
- **Multimodal Fusion**: The evaluation of different fusion methods including early fusion, where the features within the data are combined before modelling is done; late fusion, the combining of predictions from separate modality specific models into a single representation; and finally intermediate, fusion where by the learned embeddings are fused together.
- **Feature Extraction Techniques**: Assessment of a number of different encoding and feature learning approaches across modalities such as text embeddings (SBERT, RoBERTa), visual models (ResNet, ViT), and numerical feature engineering approaches.

The large Magic: The Gathering dataset from Scryfall used by SPELLS also possesses multiple qualities that make it a strong research benchmark. With the scale of the dataset being over 27,000 samples containing multiple modalities across textual, visual and structured data, this allows for detailed and robust evaluation. The complexity of the dataset adds another layer to the possible features that a downstream model can be trained on. The different game mechanics encoded within each card, complex semantics and wide variety of visual cues ensure that models require a wide view of the data to learn a deep understanding of the data rather than simply memorising a single element for predictions.

These characteristics allow SPELLS to be classified as a generalisable benchmark for semi-supervised multimodal research. The work completed within this thesis helps to close the gap between synthetic datasets and noisy real-world data, allowing for the comparison of multimodal learning methods and encourages further development of new algorithms that can make use of limited labels across multiple data modalities.

Handling Subjective Latent Attributes

A major contribution of SPELLS is its approach to learning of subjective, latent attributes. These qualities often depend heavily on human perception or expert judgement over a fixed objective truth. Unlike categorical classification tasks (identifying a card's colour or type), the estimation of a card's *power level* requires the model to infer in a way that is more nuanced that cannot be captured through discrete labelling. The use of a continuous labelling approach will allow the slight differences between cards without forcing cards into a single category.

Traditional machine learning models often struggle with subjectivity, especially when training data contains imbalanced data and class biases. Label propagation algorithms are often sensitive to noisy or ambiguous nodes. These nodes might include data points that overlap multiple classes or nodes that are not connected to any particular class. If these examples are not handled correctly this data can degrade the models performance by effecting data manifold causing over-smoothing - when distinct nodes become too similar to other surrounding nodes within the graph [30]. SPELLS addresses these issues through a combination of design steps and validation processes:

- **Continuous labelling**: Power levels are represented on a continuous power scale from 1.0 to 10.0. This allows the model to capture subtle differences between cards and grade them appropriately, without forcing discrete categories onto cards.
- **Manual Validation**: After the label propagation process is complete, human verification is applied to identify and correct propagated pseudo-labels that may be clearly misclassified. This application of expert understanding ensures that incorrect labels or labels with a low confidence can be updated before the next iteration.
- **Feature-Based Similarity**: The graph structure is defined using feature similarity across text, image, and numerical modalities, enabling the systems to group cards with related mechanics or effects even when their visuals or textual features differ significantly.

This combination of approaches allows SPELLS to build a consistent and smooth representation of the manifold geometry, to learn subjective qualities and attributes from multimodal card data across a large dataset. This framework has the potential to expand into many different applications in domains such as: product quality assessment, content moderation, educational skill evaluation and reputation scoring, where subjective human ratings are required but are expensive yet can be modelled through structured, multimodal signals.

3 Methodology

3.1 System Overview

The proposed architecture for SPELLS (Semi-Supervised Power Estimation and Label Learning System), is designed to estimate the latent power levels of Magic: The Gathering cards by integrating multiple data modalities - textual descriptions, visual imagery, and structured numerical features - into a unified predictive framework. The architecture follows a semi-supervised multimodal learning pipeline, combining graph-based label propagation with deep neural representational learning. Figure 3 shows an overview of the data flow and key components of the SPELLS system architecture.

SPELLS operates across three main stages: (1) feature extraction and representational learning, (2) semi-supervised label propagation, and (3) supervised model training. Each stage contributes to constructing an information-rich, interpretable, scalable representation of card power using minimal labelled data.

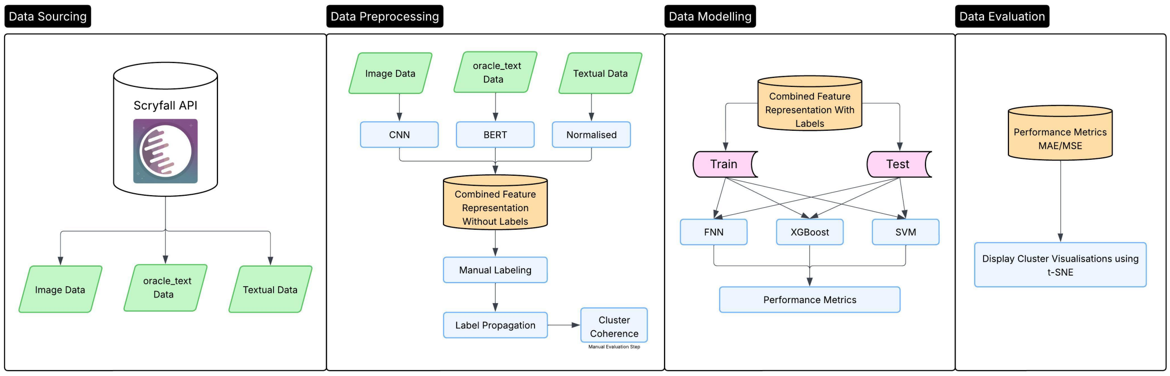

Figure 3: System architecture diagram, outlining each stage of the project

Stage 1: Multimodal Feature Extraction and Fusion

The first stage focuses on the converting of heterogeneous card data into comparable numerical representations.

- **Textual Features** are extracted from the *oracle text* of each card using a pre-trained SBERT transformer model. These embeddings capture semantic meaning from card abilities, keywords and rules text.

- **Visual Features** are obtained from full card images using a ResNet50 convolutional neural network pre-trained on ImageNet. The CNN learns high-level design and colour cues that often correlate with card rarity, creature type, or latent thematic power

- **Structured Features** include numeric and categorical attributes such as converted mana cost (CMC), power, toughness, and card type. These are normalised and encoded to ensure compatibility with the input layers of the neural network downstream.

Once extracted these modality-specific embeddings are concatenated into a single multimodal representation vector $x_i \in \mathbb{R}^F$ for each card i, where F is the total feature dimensional space after dimensionality reduction. Uniform Manifold Approximation and Projection, or UMAP was used initially on the fused embeddings before the label propagation step was complete. This reduction in dimensions helps to improve clustering and preserves local groups [31]. Principle Component Analysis was used to reduce the dimensions of the fused embeddings before training the deep neural network and other baseline models resulting in a feature matrix $X \in \mathbb{R}^{N \times F}$ serving as the foundation for downstream learning tasks.

Stage 2: Semi-Supervised Label Propagation

The fact that explicit card power labels do not exist for Magic, The Gathering, a small subset of cards (between 50 - 100) were manually annotated using expert domain knowledge. These labelled examples act as anchor nodes in a graph-based label propagation algorithm.

To propagate these labels out to the rest of the dataset, SPELLS makes use of a similarity graph generated over the entire dataset by condensing the multimodal representation down into a smaller set of anchor points using k-means clustering. Instead of using a full $N * N$ pairwise similarity graph, which with such a large dataset would be computationally expensive, SPELLS makes use of an anchor graph where each card is only connected to a number of neighbours. This approach reduces the computational cost of the propagation graph, improving the overall speed of the algorithm.

A bipartite similarity matrix Z is then generated using cosine distances between the cards embedding values and the anchors values. Label Propagation is then done in closed form using the anchor graph. This process enforces the manifold assumption, the idea that similar cards should receive similar predicted power estimations based on location within the representation space. As the manually labelled cards act as the ground-truth anchors for other cards propagated values, once the process is complete a subset of propagated labels are reviewed through a manual validation step in order to correct obvious outliers before being used for supervised learning [29].

Stage 3: Model Training and Prediction

The final stage of the process uses both the manually labelled and pseudo-labelled data to train predictive models. Three machine learning models are utilised:

- **Feedforwad Neural Network (FNN)**: Learns non-linear mappings from the multimodal vector space to continuous power predictions.
- **XGBoost Regressor**: Captures structured relationships in the data and provides interpretable comparison baseline results.
- **Support Vector Regressor (SVR)**: Serves as a lightweight baseline to assess separability in the feature space.

The trained models produce a continuous power estimation score between 1.0 - 10.0, providing interpretable estimates aligned with human-perceived scores.

3.2 Data Collection and Preprocessing

3.2.1 Data Sources

All of the data used for this research was obtained using the Scryfall REST API, a publicly available and well-documented database for Magic: The Gathering cards. The Scryfall API provides structured JSON data for all of the over 27,000 unique cards that have so far been released. Data retrieved through the API includes, textual descriptions, gameplay statistics such as card power and toughness and high-resolution card images. This dataset was selected as it is very well-documented, has a highly formatted structure and is publicly accessible, allowing for future research to be done while ensure reproducibility and transparency. As can be seen in figure 4 the total number of cards and their representing colour groupings.

Each card record retrieved from the API included several modalities of interest to this research project:

- **Textual Attributes**: Name, oracle text and type line
- **Numerical Attributes**: Converted mana cost (CMC), power and toughness
- **Categorical Attributes**: Card colours, rarity, set identity and layout
- **Visual Attributes**: Image URLs for all types of card art that might be present on a card

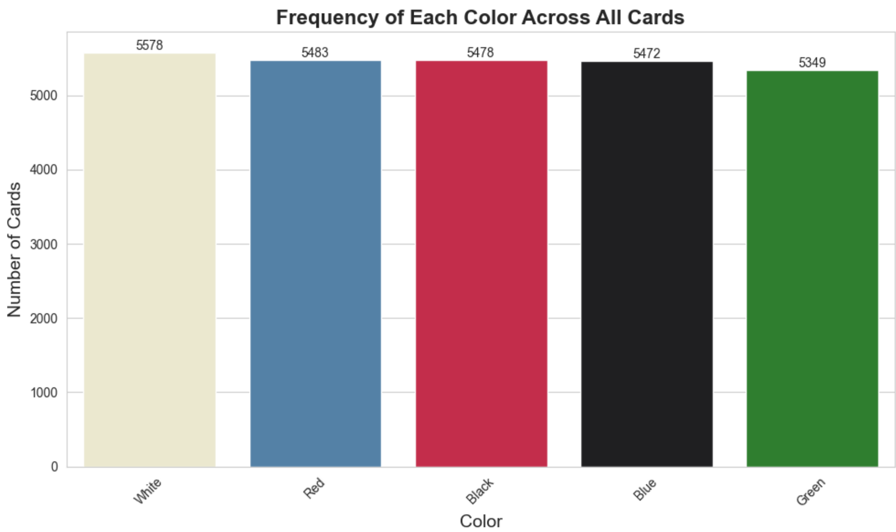

Figure 4: The breakdown of single card colours after data cleaning and processing.

Using Python's request library data was collected programmatically, adhering to the API's rate limits and usage policies. Each card entry was stored locally as a structured JSON object and in csv format, before being converted into a Pandas Dataframe for preprocessing and integration downstream.

The images used throughout the project were downloaded using a separate script to keep that code removed from the main pipeline for ease of use. A script called download_images.py making use of multithreading was implemented to download the image files. As each card within the dataset has an image file of around 100 KB, when downloading over 27,000 images, and while allowing for wait times in order to keep within the API rate limits, download times were exceeding an hour to pull all the images. To reduce the wait time for any new images pulled, a function was implemented to run concurrently using ThreadPoolExectuor, allowing for the image files to be downloaded much faster. An example of the cropped images used in the project can be seen in 5. This type of cropped image was used instead of the full art style in order to avoid data leakage from learned features such as rarity or colour identity through structural features. The images of this size were also closer to the 224x224 pixels ingested by the DNN and reduced the removal of important image information during resizing.

3.2.2 Data Cleaning

Although the raw data from the API was well structured and formatted it required cleaning before any feature extraction could take place. Extensive preprocessing was

(a) Wasteful Harvest (b) Geth's Grimoire (c) Meteor Golem

Figure 5: Example showing three of the cropped images used for the image modality.

done using Pandas and NumPy to ensure data consistency and quality. The inital dataset cleaning pipeline included the following steps:

- **Duplicate Removal**: Cards with identical names, identifiers or reprints across different sets were filtered out and removed, keeping only a single version of the card. Any cards with the same oracle text were also removed.
- **Retain English only Cards**: Although Magic: The Gathering cards are printed in many different languages, for the purpose of this research only cards in English were retained.
- **Handling Missing Values**:
  - Any cards missing essential textual or numerical data, such as cards with missing oracle text or cards that had no power or toughness statistics, were excluded from the dataset.
  - Missing values for categorical attributes like card colour or rarity were imputed or inferred from similar cards of the same set.
- **Card Type and Set Exclusion**: A number of MTG sets (cards within the "UN" sets) were excluded from the working set before moving on as they include joke cards with strange or excessively powerful mechanics that are not intended for play outside of the specific joke sets. Certain types of cards such as tokens and emblems were removed from the set, as these cards generally do not contain any abilities or effects and may have caused issues downstream.
- **Image Validation**: Image URLs were checked for each card as there are different types of card images, and cards that have artwork and information on both sides of the card. Double sided card versions were found and if a standard card image was found this was retrieved instead of the flipped version.
- **Type Consistency**: Columns such as cmc, power and toughness were cast to float data types. All cards with NaN values for oracle_text were replaced with an empty string, and cards with variable power or toughness values, represented as "*" were converted to numeric or NaN.

3.2.3 Oracle Text Preprocessing

A detailed cleaning pipeline was implemented for the oracle_text data as it was considered an essential modality to have clean and ready for embedding generation. A three step cleaning process was designed in order to normalise the text, extract certain card mechanics and then tag these mechanics within the textual data.

The first stage in the cleaning pipeline was the remove_extra_text() function. This function was built to take in the oracle text data and remove any explanation text from the data. Explanation text was found using regex to find certain keywords from within parenthesis, like *reminder* or *defending player*. These such keywords within brackets were removed as they explain an ability offer specific rulings for a card. This function also utilised a mapping of mana symbols to words to remove any instances of symbols like {G}, {R,W} or terms such as {T}. The mapping was used to change these into readable English such as: *colourless* or *untap*.

Secondly, the function clean_oracle_text(), was used to normalise common terms within the text like *draw a card* or *discard a card*. Card mechanics like triggered abilities at the start of a player's end step, are found and marked with a *trigger_when* keyword or *trigger_at_start_of_*; other mechanics like tapping a card were replaced with *tap_ability* keywords to allow for pretrained SentenceTransformers neural network to better understand the semantic meaning when generating embedding vectors for each card. A check for whitespaces was done at the end of each step to remove any extra blank spaces within the text.

Lastly, a function called ability_tagging() was used to checked through a list of keywords that are commonly found on Magic cards to tag abilities that are found across each card. If an evergreen ability within the list if found in the oracle text data, a tag *has_X* will be added to a new column for the card within the dataframe. This explicit ability tagging allows for the model to learn the semantic similarity between cards based on known abilities. Cards that have the same abilities should be located closer together within the data space during embedding generation and label propagation. Other abilities or traits that are found on cards were extracted through the use of the keywords like *each creature* or *search your library for a land*, if such keywords were found; a special tag like *mass_removal* or *ramp* was added to the cards oracle_text to further indicate strengths of individual cards against others.

This pipeline evolved through trial and error, with the initial oracle embeddings being generated on the raw text data. Although this approach worked the performance of the downstream model was improved with each iteration of the cleaning process. The second testing of the pipeline used some basic regex cleaning to removed unnecessary words and eventually grew into the full processing of explanation text. With each iteration of the pipeline the semantic meaning extracted from the text improved and the finally tagging of certain abilities was implemented.

3.2.4 Data Normalisation and Encoding

In order to ensure feature comparability across all modalities, all numeric and categorical data were normalised and encoded using multiple techniques such as LabelEncoder and Min-Max scaling.

- **Numerical Scaling**: All numeric features such as cmc, power and toughness were scaled using Mix-Max normalisation from a range [0, 1]. As seen in figure 6 there are a wide spread of CMC values within the dataset. Scaling here was critical as extreme values would negatively affect the model.

- **Categorical Encoding**: Certain categorical features were encoded using the LabelEncoder from sklearn before having an embedding vector generated for each individual card. The super_types and sub_types of the cards within the dataset required a nuanced approach as separate features required extraction but then further encoding and data clean-up was done. This will be expanded upon in greater detail within the next section - **Feature Engineering**.

- **Feature Filtering**: Non-informative features such as card IDs, names, set codes and URLs were removed to minimise noise and prevent possible data leaks during the downstream model training process.

- **Image Standardisation**: All images were resized to a consistent dimension of 224x224 pixels and normalised, ensuring compatibility with the pre-trained ResNet50 model.

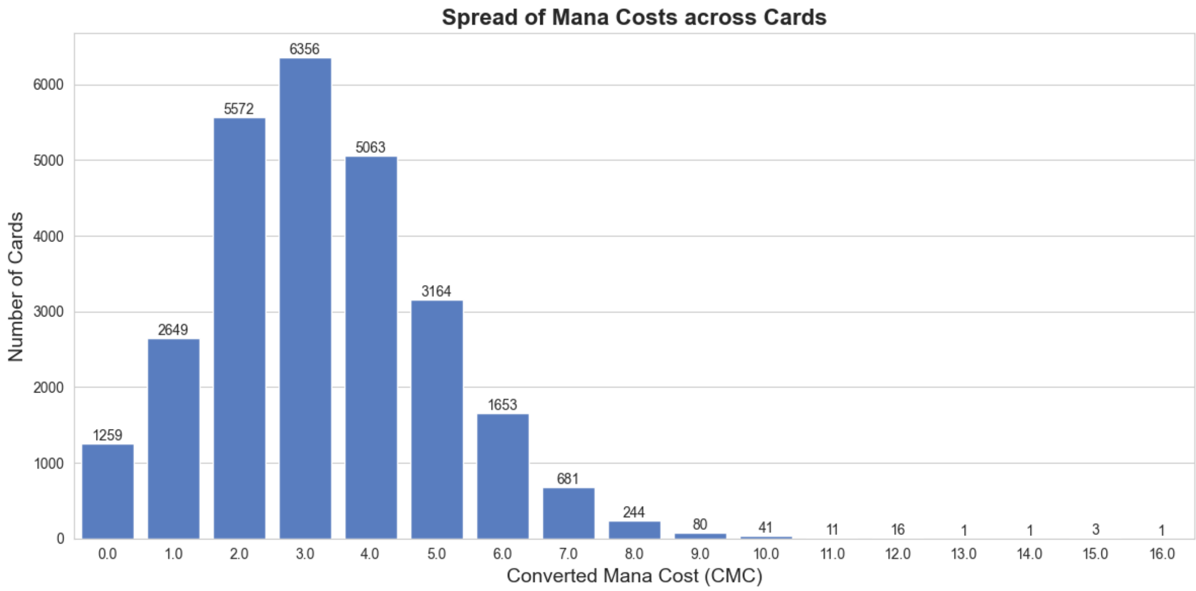

Figure 6: Plot showing the converted mana cost all cards within the dataset. This plot displays the total values after cleaning as some extreme values of up to one million exist in certain removed sets.

3.2.5 Ethical and Reproducibility Considerations

The Scryfall API data is openly available and free for research purposes, ensuring ethical compliance. All preprocessing scripts and data transformation processes are documented and available, allowing the entire data collection process to be transparent, reproducible and auditable as possible.

Image assets retain original artist ownership, and all textual content remains the property of Wizards of the Coast, used under fair use for academic research.

3.3 Feature Engineering

The feature engineering stage transforms all preprocessed data into dense, information-rich representations suitable for multimodal learning. This stage converts text, image, and structured features of each Magic: The Gathering card into a shared, high-dimensional vector space that can be used for label propagation and model training.

The process consists of three main steps: (1) modality-specific feature extraction, (2) dimensionality reduction and normalisation, and (3) multimodal feature fusion. Together, these three steps allow SPELLS to capture semantic, visual, and structural relationships between cards in a single, unified embedding representation.

3.3.1 Textual Feature Engineering

Textual data provides essential contextual information about a card's abilities, effects and gameplay mechanics. The oracle text feature was used to construct a comprehensive textual representation for each card.

The preprocessed field containing the oracle text was ingested by the sentence transformer model all-MiniLM-L6-v2[3]. This transformer model is pre-trained to produce sentence embeddings of a fixed-size. The oracle_text column from the Dataframe is separated out into a Series, and passed to model for vector embedding generation. The model tokenises each input before applying a pooling layer, and returning a dense vector with a size of 384 dimensions per input [32]. These embedding values are then stored locally to avoid rerunning expensive and time consuming computations.

3.3.2 Image Features

The images present on Magic: The Gathering cards often communicate implicit information about rarity, card colours, and thematic-style - attributes that can be correlated with strength. Visual features were extracted using a pre-trained ResNet50[4] convolutional neural network [33], available through the TensorFlow library.

- **Image Preprocessing**: Each image was resized to 224 x 224 pixels, converted to RGB, and normalised using ImageNet mean and standard deviation values.
- **Feature Extraction**: The classification head of ResNet50 was removed, and the model was used as a fixed feature extractor. The reason for this was to get an output from the global average pooling layer, that resulted in a 2048-dimensional vector representing each card image. This embedding captures high-level spacial patterns such as colour distributions and texture cues on each of the cards.

3.3.3 Structural Features

Structured numerical and categorical attributes were transformed into dense feature vectors to complement the text and image representations. SPELLS makes use of feature concatenation to combine vector embeddings from different categorical features into a single representation before the final fusion process.

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4]https://huggingface.co/microsoft/resnet-50

- **Numerical Features**: Attributes such as converted mana cost, power and toughness were retained after the preprocessing stage. After Min-Max normalisation, these continuous variables were concatenated into a single numeric feature vector.
- **One-hot and Binary Flags**: After the initial preprocessing stage new features were extracted from the existing data. Several categorical, and multi-label fields were converted into explicit binary indicators to preserve the interpretability and to avoid losing important information. Information like if a card contains a colour was extracted using binary flags like is_red or is_colourless. New boolean features such as *is_creature* were also encoded to provide direct mechanistic signals that are easy to interpret and are useful as structural model inputs.
- **Label Encoding (ID Mapping)**: Where a small, compact mapping was useful for indexing certain features, such as the newly engineered feature, single_colour. This LabelEncoder mapping was used to generate a single_colour_id; this integer ID was used for embedding lookup, but is not used as input for the training of the model downstream.
- **Categorical Features**: For categorical features fields that were better represented as dense vectors, categorical embeddings were generated instead of implementing one-hot encoding as this would explode the number of columns within the dataset. This approach was used to generated embeddings for the keywords, supertypes and subtypes of cards for example. Each of the embeddings for these features were generated using pretrained SentenceTransformers model all-MiniLM-L6-v2.

3.3.4 Dimensionality Reduction

A hybrid approach to dimensionality reduction was used in order to best visualise the embeddings generated for each of the features. Principle Component Analysis was used to reduce each of the high-dimensional embeddings down to a smaller size. The oracle embeddings were reduced from 384 dimensions, down a size of 50, before t-SNE was utilised to further reduce down to 2 and 3 components for visualisations. The the 2048 dimensional visual embeddings were reduced down to 256 components, and then again for plotting. This approach was chosen in order to preserve the dominant features within each embedding, while reducing computational load, and allowing for visualisation.

3.3.5 Feature Fusion

Once all modality-specific deep representations were obtained they were unified into a single multimodal embedding vector per card:

$$X_i = \left[T_{\text{text}} , I_{\text{image}} , S_{\text{structural}} \right]$$

Prior to fusion each feature block was normalised to unit variance to prevent dominance by high-dimensional embeddings such as the image features.

This fusion approach implements an early-intermediate fusion strategy where modality features are combined before training but after individual representational learning. This design allows SPELLS to capture cross-modal interactions while maintaining flexibility to substitute or extend any single modality module in future work.

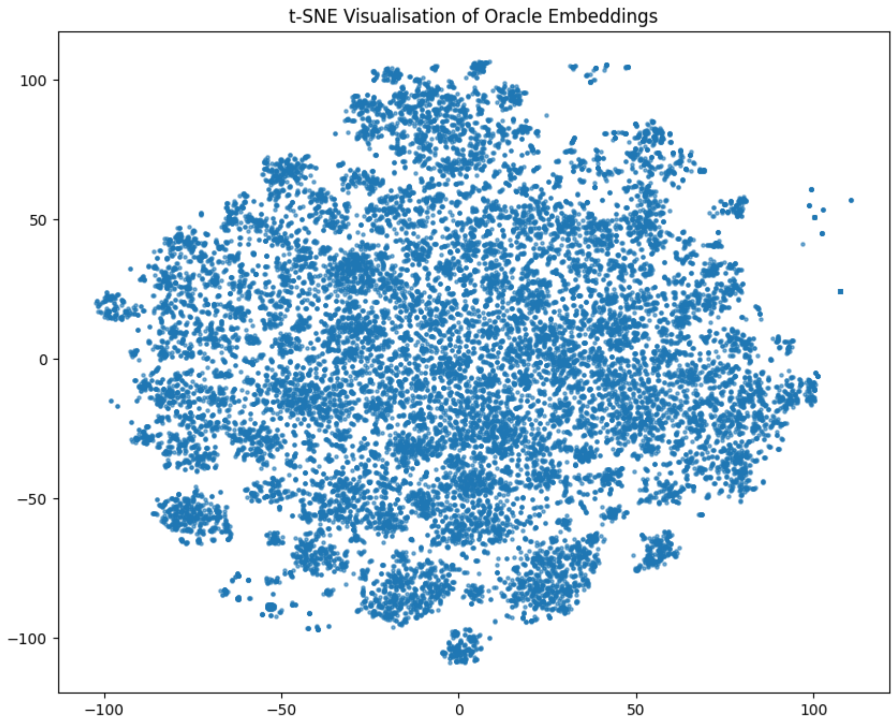

Figure 7: t-SNE Visualisation of Oracle Embeddings

3.4 Manual Labelling Process

As the dataset used for this project did not contain power scores or strength labels a semi-supervised learning approach was utilised in order to generate pseudo labels for each of the cards within the dataset. A power label was manually given to a subset of around 70 cards before the label propagation process started. This small subset of manually annotated cards, although small, was diverse, acting as anchor points for clusters of card groupings.

In order to generate a diverse but even spread of card types and colours for manual labelling a stratification method was implemented to collect samples of card types into different buckets. Cards were filtered based on their converted mana cost (cmc), and samples then split into buckets or chunks based on certain features. A Cartesian product [5] of each sample is created based on color_identity, is_creature, rarity, and cmc. With 605 strata groups created with two cards in each group, a wide spread of cards was created to manually label. K-Means clustering was then used to reduce the 605 groups, with two card samples per group, down to a list of 50 cards. The medoid [6] for each of these 50 clusters were then measured this is the most central data point within each cluster giving the final sample of cards to be manually labelled as the anchor points for label propagation.

These anchors were generated as part of an approach, learned from a method implemented by Lazarou et al in *Adaptive Anchor Label Propagation for Transductive Few-Shot Learning* [34]. The method of creating anchor points as a means of reducing the computational load and improving effectiveness of the label propagation algorithm using

[5] https://www.sciencedirect.com/topics/computer-science/cartesian-product
[6] https://neo4j.com/blog/developer/clustering-graph-data-k-medoids/

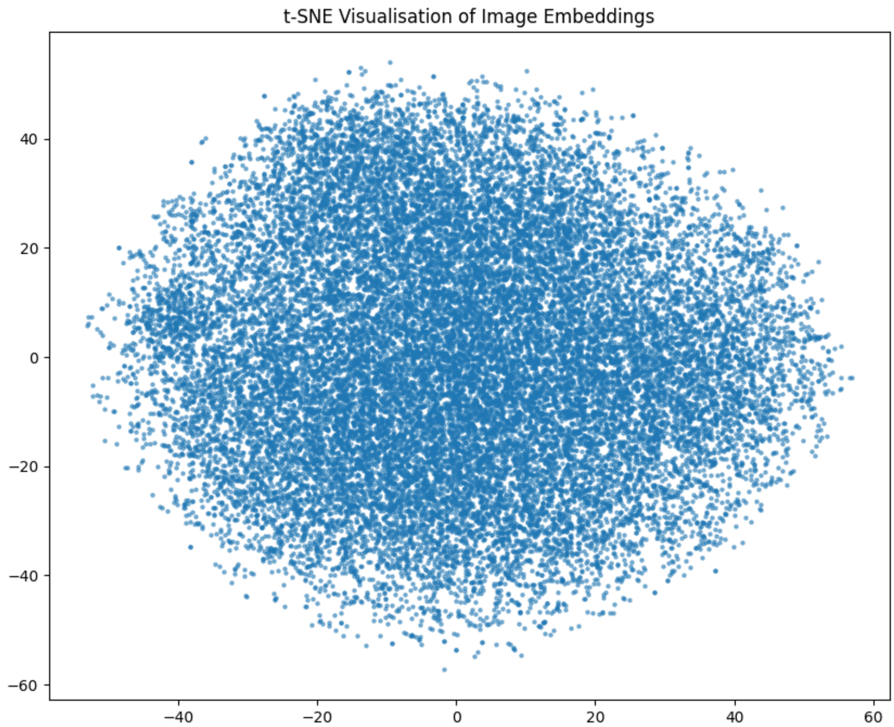

Figure 8: t-SNE Visualisation of Image Embeddings

embeddings to optimise their positions within the constructed graph. This idea of anchor points were used in SPELLS as a method of generating proxy labels for each card based on the unified representation embedding vector from each feature modality.

K-Means clustering was used to generated a number of anchor points M, over the fused embedding representation of the dataset. For each card's embedding within the dataset, the Euclidean distance was generated to capture the distance from a card to a tunable number of nearest anchor points. This small subset of anchors was chosen using K-Means clustering to represent groups of 'typical' cards. This reduced number of cards then has a small number of neighbours checked when measuring the similarity of cards against these clusters, meaning there is no need to have every single card be compared to each other, only a smaller subset of close neighbours. This reduces the computational needs of the system dramatically, from 26,795 * 26,795, which would be slow and memory intensive, to a more manageable comparison of 26,795 * X neighbours (between 5 - 50).

3.4.1 Grading Rubric

To accurately score the subset of sample cards used in the manual labelling process a power_score grading rubric was created. The rubric follows a number of checks used on each card to evaluate the power_score of the card by looking at its features. Each dimension of the rubric is given a score out of a possible 10, with each dimension having a certain weighting, totalling up to 100%, with the total power_score for the annotated card calculated using the following formula:

$$\text{raw\_score}_i = w_1 \cdot M_i + w_2 \cdot I_i + w_3 \cdot L_i + w_4 \cdot F_i + w_5 \cdot S_i - w_6 \cdot R_i$$

Once each score is generated for the specific card to be manually labelled, each score is normalised using min-max scaling to provide a continuous values between 1.0 and 10.0.

- **Mana Efficiency - 20%**: This is the general mana cost to ability ratio. How much does a card cost to cast given its relative power and toughness or number of abilities.
- **Impact - 20%**: The impact of a card depends what a card does when it enters the battlefield or what effects happen when the card is cast. This would include triggered abilities and effects such as board wipes.
- **Long Term Value - 20%**: How long does a card give a returned impact or does the card stay on the battlefield and continually provide benefits. Enchantments cards are good examples of a card type with long term value as they remain in play, usually providing some benefit to a player or prohibit another.
- **Flexibility - 15%**: Does a certain card have a strong utility value or can this card help across many situations. Cards with multiple colour identities can often be useful in many ways, or could be useful when used alone.
- **Synergy - 15%**: Does a card work well with other cards or is it difficult to use without relying on a specific number of complex plays. If a card can be used with lots of other types of cards and build powerful combos it is rated highly under synergy.
- **Restrictions - 10%**: This dimension is weighted the lowest and reduces the overall score of the card. A restriction is put on a card if it requires more colours to play as this is generally more difficult to get out onto the battlefield. Other penalties are given for how situational the card is. If a card only helps to project against a single card in another players deck then points are removed for this as the counter is considered too specific.

3.5 Label Propagation Algorithm

The label propagation algorithm uses an identity matrix I as a safety net for regularisation, ensuring the model doesn't experience any overfitting. This M*M matrix is made up of vectors of 0s with 1s down the diagonal axis. A lambda expression is set as a control for the regularisation strength, telling the model how much to avoid overfitting or not. This parameter was fine-tuned throughout the label propagation process, increasing or reducing the smoothness the labelled data was to be trusted.

3.6 Active Learning

In order to improve the overall quality of the label propagation process, while minimising the number of samples labelled manually, SPELLS incorporates an active learning loop. After the initial label propagation and deep neural network training, the results showed some issues with certain cards getting stuck with final power scores closer to that of the locally annotated groups within the label propagation graph. Although the LP is working as expected and cards of certain types and styles are being clustered correctly, cards with a varity of keywords or card types seemed to be labelled with scores that were either too far above or below an expected value.

Two unlabelled cards that were used as examples of strong and weak cards were: *Oko, Thief of Crowns* and *Mudhole*. Oko is generally considered a very strong card and is banned across multiple game formats, Mudhole on the other hand is a weak card with very few abilities. After the first label propagation Oko was returned with a propogated_score of 1.0 which was definitely the incorrect power label for a card with such powerful features.

The active learning loop was implemented in order to reduce the number of outliers, and extreme values for cards that were expected to get vastly different results to their propagated values. The process included adding a manual check after each label propagation iteration and checking for extreme values. After each iteration a number of cards were looked at manually and evaluated based on the strength labelling rubric. If a card received a propagated value that was dramatically different to it's expected result, the card was re-labelled manually according to the rubric. This process increases the number of manually labelled cards within the dataset and gives the next iteration of label propagation improved data to work with. Table 2 shows an example of the propagated values of cards, with a single manually labelled card *Lightning Bolt*, with a power_score of 9.0.

This active learning - human-in-the-loop - approach was repeated for multiple iterations, with between 20 and 30 cards being manually corrected to ensure outliers were removed, and future label propagation iterations had more accurate anchor points for clustering the cards.

ID	**Name**	**Power score**	**Propagated power**
492	Oko, Thief of Crowns	—	6.373799
801	Lightning Bolt	9.10	6.683396
2770	Black Lotus	10.00	8.300191
4705	Ragavan, Nimble Pilferer	—	5.776669
15836	One with Nothing	—	2.964058
21281	Mudhole	—	5.701787

Table 2: Selected example of cards with a manual power_score label and the propagated_power values for each card after the label propagation process.

3.7 Model Training and Architecture

3.7.1 Feedforward Neural Network

The deep learning neural network used for the regression training within SPELLS is a fully connected feedforward neural network designed to ingest the training data embeddings with the propagated label data for each card of the dataset. The DNN was architected for the regression of continuous values from 1.0 to 10.0 for each magic card within the final set. There are seven layers within the sequential network, four dense layers, making use of progressive downsampling, with 256, 128, 64 and finally a single scalar output layer for regression. Each of the three dense layers above the final output layer utilise the GELU activation function in order to optimise the smoothing of tabular input data. The output layer uses a linear activation for this regression task as a continuous scalar value is required. There is a single batch normalisation layer after the first dense layer of 256 neurons, to speed up the convergence, and two dropout layers help with additional regularisation to prevent overfitting.

Listing 1: Deep Neural Network configuration

```
dnn_model = models.Sequential([
    layers.Dense(256, activation='gelu', input_shape=(input_dim,)
       ),
    layers.BatchNormalization(),
    layers.Dropout(0.15),

    layers.Dense(128, activation='gelu'),
    layers.Dropout(0.10),

    layers.Dense(64, activation='gelu'),
    layers.Dense(1, activation='linear')
])
```

The AdamW optimiser was chosen for this architecture over the standard Adam[7] as it provides good regularisation and often improved generalisation performance. A learning rate of 1e-3 was chosen as a starting rate and was found to produce good results with this architecture. The MSE loss metric was used for this model as it aligns well with regression objectives when training the DNN.

3.7.2 XGBoost

The first of the comparative baseline models to be trained was the XGBoost model. The XGBRegressor was used to train a model on the same data ingested by the DNN before results were evaluated. The XGBoost model had a tuned setup for good performance with the configuration seen in the code. The n_estimators value of 1500 builds a large tree with a learning_rate of 0.05 and a max_depth of 10 for a boost in performance at the cost of speed. With a dataset of nearly 27,000 samples and a large number of trees, the computational times were balanced using the *hist* tree_method, this method speeds up training times and reduces the memory usage of large models through the use of binning [35].

Listing 2: XGBRegressor model configuration

```
xgb = XGBRegressor(
    n_estimators=1500,
    learning_rate=0.05,
    max_depth=10,
    subsample=0.8,
    colsample_bytree=0.8,
    objective='reg:squarederror',
    tree_method='hist'
)
```

3.7.3 Support Vector Machine

A Support Vector Regression model was chosen as the second of the baseline comparisons of the deep neural network for SPELLS. The regression model was imported from the sklearn library in order to predict the power_score for all the cards within the dataset.

[7]Adam vs AdamW

This model makes use of the radial basis function kernel this is a common default setting for non-linear relationships. For this setup, the gamma is set to *scale* as it is the default sklearn value. Again the epsilon value of 0.1 is the default value for this SVR model. The value of C is increased to 10.0 in order to improve regularisation, by applying a higher penalty to training errors in attempt to improve the fitting performance of the model.

Listing 3: XGBRegressor model configuration

```
svm = SVR(
    kernel='rbf',
    C=10.0,
    gamma='scale',
    epsilon=0.1
)
```

4 Evaluation Metrics

To evaluate the performance of the models trained for SPELLS, multiple approaches were used and certain techniques like ablation were utilised to fairly test a wide variety of architectures. Two pillars of evaluation were used in order to test the performance of the models trained for this project. The first pillar of evaluation was to test the models on only the manually labelled dataset. This approach evaluates the consistency between the models outputs and the true human labels. The second pillar of evaluation measures local predictive consistency using a subset of representative cards from clustered groups.

4.1 Baseline Comparison ML Models

In order to conduct a thorough analysis of how the deep neural network performed multiple baseline models were tested to provide a rigours and comparative approach. Support Vector Regression (SVM) and XGBoost models were both trained as comparative baselines to see which of the models produced the best results. Having multiple machine learning models for a detailed comparison of results, provides evidence that the chosen architecture of SPELLS is justified, highlighting strengths and weaknesses of each of the trained models under controlled environments.

4.2 Ablation studies

As SPELLS makes use of multiple modalities for training data it was important to conduct the evaluation using some ablation techniques. Each of the models trained throughout the project were tested using all three modalities: oracle data, images and the structural data, before certain modalities were removed in order to better compare the models performance relative to the modalities being ingested during training.

4.3 Classic Machine Learning Evaluation

To compare the performance of each of the machine learning models traditional evaluation techniques like Mean Absolute Error and R^2 metrics were used. All three models used were trained on each of the fused modalities. The first pillar of evaluation used for SPELLS measures how well the models perform when comparing the manually labelled samples and the predictions outputted by each of the models. One of the main evaluations of this stage was to see how the three very different models could perform on the same fused embeddings, with a solid set of results for all three models, this indicates that the embedding space is likely capturing meaningful structure relevant to gameplay power.

The initial evaluation of all three embeddings combined into a single fused representation are shown in table 3. These results show the DNN performing well with an MAE of 0.8544 and the best R^2 value of the three models evaluated in pillar 1 0.8389. The performance of SVR and XGBoost are also strong with MAE < 0.75, although they both have weaker R^2 compared to the neural network.

Model	**MAE**	R^2
DNN	0.8544	**0.8389**
XGBoost	0.7487	0.6567
SVR	**0.7065**	0.6973

Table 3: Comparison of the three models trained on all three fused modality embeddings. MAE and R^2 performance metrics shown for the first pillar of evaluation.

As mentioned previously, ablation testing was done in order judge the usefulness of the image embeddings within the representation space. As the visual embeddings had a negative effect on the graph-based structure of the label propagation, appropriate testing was done to rule of similar findings on the performance of trained models. A second representation space with the image embeddings removed was evaluated using the same three machine learning models. The results seen in table 4 shows each of the three models trained on fused embeddings using only the Oracle and Structural data.

The Support Vector Regressor produced the best MAE of 0.58 and with a R^2 score of 0.79, the SVR did a good job of modelling the embedding space using the radial basis function. This suggests that the fused embedding space created contained a smooth and kernel friendly structure. The XGBoost model also performed very well on both of these metrics. With an R^2 value of 0.75, this was the lowest of the three evaluated on the manually labelled subset, showing that the XGBoost model struggled the most with variance. Finally, the deep neural network although it had the worst MAE value at 0.92, the DNN performed very well with the highest R^2 score of the models evaluated.

Model	**MAE**	R^2
DNN	0.9264	**0.8046**
XGBoost	0.6196	0.7593
SVR	**0.5811**	0.7934

Table 4: Comparison of the three models trained on Oracle text and Structural embeddings, with the image embeddings removed. MAE and R^2 performance metrics.

From the results shown in table 4 we can see that the DNN actually performed better with the visual features included within the representation space. Although both the SVR and XGBoost models performed significantly better with the visual data removed. The MAE for both improved without the image embeddings < 0.62 with SVR showing a very strong score of 0.58 when evaluated against the manually labelled cards. Surprisingly, the

DNN performed better with the visual data included, meaning that the fused embeddings *do* help the neural network. It is clear that the image embeddings introduced too much visual noise for the XBGoost and SVR models to learn the feature space as effectively.

Model	**Text + Structural**	**Fused**	**Interpretation**
DNN	MAE 0.93 R^2 0.80	MAE 0.85 R^2 0.84	Slightly improved
XGBoost	MAE 0.62 R^2 0.76	MAE 0.75 R^2 0.66	Images degrade model
SVR	MAE 0.58 R^2 0.79	MAE 0.70 R^2 0.69	Images degrade model

Table 5: Performance comparison of models with and without image embeddings.

4.4 Structural Consistency and Local Smoothness

A strong model architecture is only one of the ways of evaluating SPELLS, it is equally important to understand whether or not the system produces consistent and stable predictions across the multimodal embedding space. In order for SPELLS to to function as a card ranking system, similar cards should receive similar predicted power scores. To evaluate this, the second pillar of evaluation looks at local neighbourhood consistency. Groups of cards that are close together within the fused multimodal embedding space are analysed to determine how well the model retains relative ordering and if smooth stable predictions have been produced. This evaluation provides a deeper understanding of how well SPELLS captures the structural relationships between cards.

The embeddings tested were reduced to 64 dimensions using Principle Component Analysis (PCA), in order to reduce noise within the high-dimensional feature space of combined multimodal attributes. This reduction in dimensions also help to preserve the neighbourhood relationships required for assessing the geometry structure of the manifold. A 2D projection of the PCA-reduced embedding space can be seen figure 9, the plot helps to visualise the groups learned using K-Means clustering. These clusters represent the latent card archetypes such as mana-ramp effect, control spells and aggressive creatures.

For each cluster, a representative centroid card was taken and can be seen in figure 10. These centroid cards represent the average structural location of their respective neighbourhood and are then used as anchor card for subsequent evaluation. As these anchor cards within the cluster share similar function and statistical features with their neighbours, a smooth and consistent model should be built showing power scores ordered correctly across these clusters.

Using the evaluation cards as seen in figure 10, the performance of SPELLS to produce a consistent and structurally smooth model was measured using four metrics. Mean Absolute Error, R^2, Spearman correlation[8] and Kendall Tau correlation[9]. Both of the Spearman and Kendall metrics evaluate the structure of the representational space by measuring how well the cards were placed when compared to other cards. The Spearmanr correlation metric looks at whether or not the ranking of cards was preserved from the label propagation and the model training stages of SPELLS. Kendall Tau measures the comparative strength of each card derived from the model and the label propagation.

[8]https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.spearmanr.html
[9]https://docs.scipy.org/doc/scipy-1.16.2/reference/generated/scipy.stats.kendalltau.html

Both metrics give a strong view of how well the models performed and if they were able to learn the global structure of the manifold constructed earlier in the pipeline.

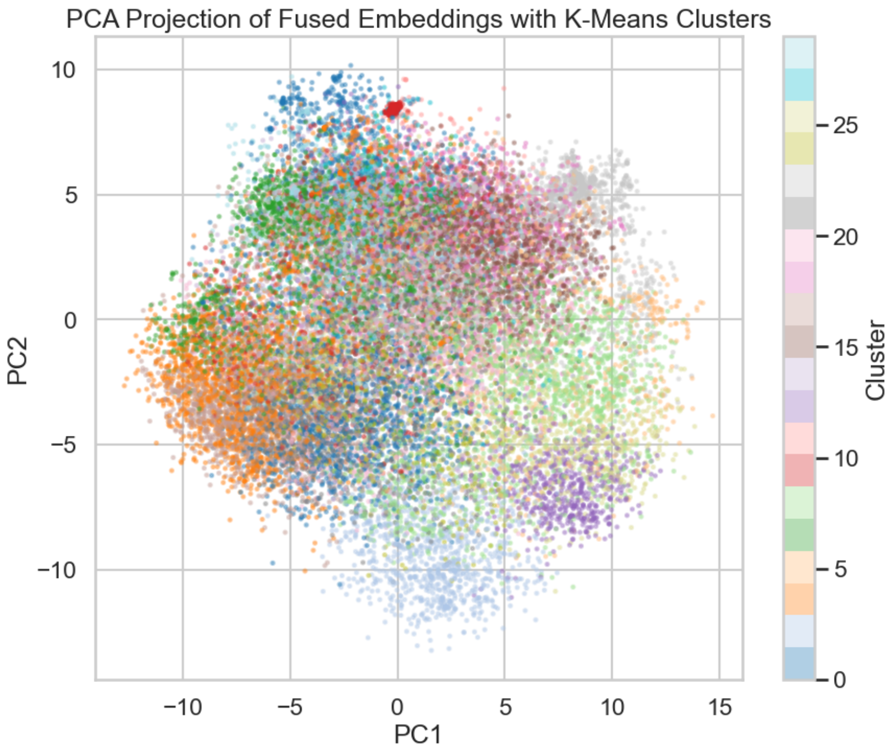

Figure 9: Representation of the fused embedding clusters using PCA for evaluation of SPELLS and measuring of a structurally smooth and consistent model. This plot shows only the Oracle and Structural embeddings.

The Spearman and Kendall metrics are particularly important for this pillar of evaluation, as they assess how well the model can capture rank-ordering relationships within the manifold. The results seen in table 6, show the performance for each of the models using the embedding space with no visual information. Both the deep neural network and XGBoost models had a Spearman correlation value of over 0.96, showing very high agreement between the models and the label propagation labels. These values mean that the cards that were considered strong or weak by label propagation, were also considered strong or weak by these two models. This stage of the evaluation process measures the structural consistency of SPELLS and with the results in table 7, it proves that the current architecture is effective at retaining the structure of the learned representational space. With high scores in both Spearman above 0.96 and Kendall above 0.89, using DNN and XGBoost, SPELLS was able to discern, based on the learned features, which cards are stronger than others.

XGBoost achieved the lowest MAE result with a score of 0.19, suggesting that the model has a strong alignment with the propagated labels, where as DNN had the highest R^2 value of 0.87, indicating that the neural network outperformed the others in modelling variance. The high R^2 value achieved by the DNN is not unexpected as neural networks excel at learning deep structural relationships from within the fused embedding space.The XGBoost model performed very well on the MAE values, showing that it had the closest raw value predictions to the propagated labels. XGBoost also had the best Kendall

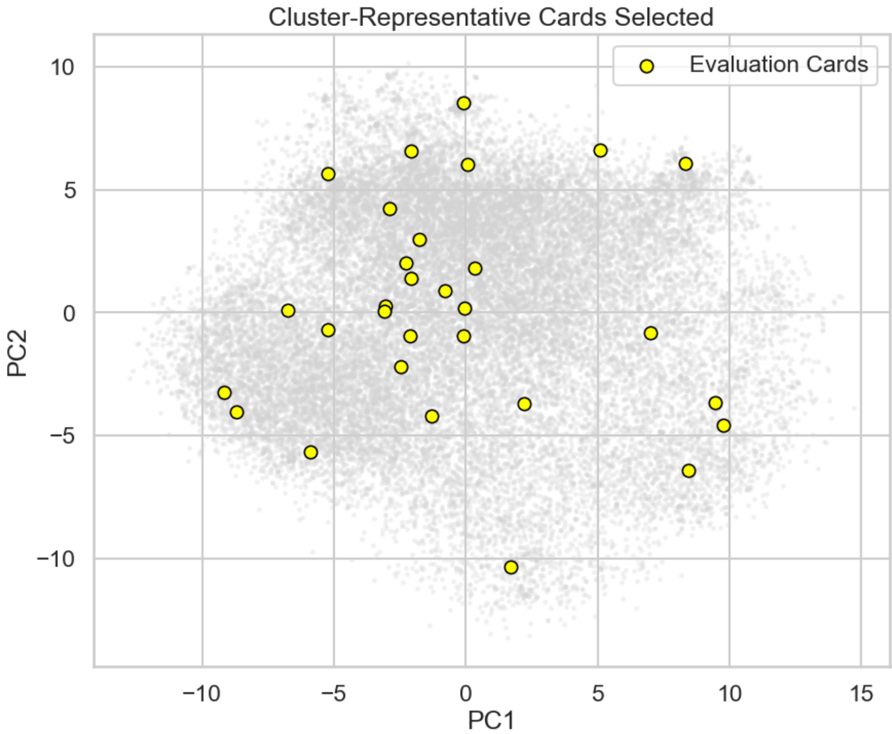

Figure 10: Centroid value cards representing each of the K-Means clusters.

value of 0.93, with a slightly lower results for R^2, showing it's weakness in variance when compared to the DNN.

The Support Vector Machine completely fell apart when evaluating the consistency of the manifold using only textual and structural features. With an R^2 value of 0.2, the highest Mean Absolute Error of 1.4, and the lowest correlation values for both other metrics, SVR was unable to capture the same detailed information from within the high dimensional representation as the other two models. SVR struggled with the multimodal non-linear representational space, confirming that even when using a non-linear kernel, SVM are not suitable for SPELLS.

These results prove that SPELLS, with fused multimodal embeddings and graph-based label propagation, can create a smooth and learnable latent space for downstream training of models and the estimation of latent power for Magic: The Gathering cards.

Model	**MAE**	R^2	**Spearman**	**Kendall**
DNN	0.3757	**0.8737**	**0.9737**	0.8943
XGBoost	**0.1908**	0.8298	0.9689	**0.9310**
SVR	1.4494	0.2342	0.7343	0.5540

Table 6: Evaluation results for the second pillar of evaluation measuring the structural consistency of local neighbourhoods within the embedding space using textual and structural features.

When evaluating the performance of the models using the fused embedding space of all modalities, the results show that the SVR is not suitable at all for SPELLS, as it cannot handle the noisy representation space when images are included. Looking at the results in table 7, the high MAE 1.26 and negative R^2 -0.028 are enough to prove that the SVR collapses under this evaluation. Surprisingly, the XGBoost model performs nearly perfectly in this pillar of evaluation with extremely high scores: MAE= 0.01 and R^2 = 0.99, telling us that the model is extremely consistent and smooth across local neighbourhoods, with virtually zero high variance in ranking of similar cards. Although these values are great, it does not mean that XGBoost performs the best globally, just when looking at local clusters.

The deep neural network also performs very well when the full fused embedding space is evaluated. A low MAE of 0.41 tells us that the DNN is stable and consistent with small variation within neighbourhoods. High Spearman and Kendall values 0.95 and 0.83 respectively, also indicate that the ordering of cards latent power is preserved. As seen in the first pillar of evaluation, the DNN performed better when looking at global accuracy which is important for SPELLS.

Model	**MAE**	**R^2**	**Spearman**	**Kendall**
DNN	0.4140	0.8909	0.9537	0.8299
XGBoost	**0.0108**	**0.9988**	**0.9973**	**0.9862**
SVR	1.2612	-0.0276	0.5248	0.3655

Table 7: Performance metrics of the models using all three modalities: textual, visual and structural embeddings.

5 Discussion

The research of AI and Magic: The Gathering started out as in investigation into whether multimodal semi-supervised learning could infer the latent power levels of MTG cards without the need for gameplay statistics or large-scale manual annotation. SPELLS combines three major components: multimodal embeddings, anchor-based label propagation and supervised regressor models to construct a learnable, continuous power level manifold geometry. The results proves that SPELLS was successful and the objective is feasible and in most cases very effective.

5.1 Multimodal Representation Learning

One of the main contributions of SPELLS to the research of artificial intelligence is the fusing of textual, structural and visual card information into a single unified embeddings space. The oracle text cleaning pipeline, generation of embedding vectors for each of the card's subtype and supertype, and the encoding of semantic keyword information, all contributed to the complex but learnable representation.

The oracle text data seemed to contained the most information of the three modalities tested throughout this research. After extensive cleaning, normalisation and mechanical tagging, SBERT embeddings were able to capture important functional relationships between cards. Details such as damage scaling, token creation and triggered abilities were mapped using the pretrained model. The improvements to this oracle text processing pipeline were directly reflected during the downstream model performance. These changes highlight the sensitivity to the quality of textual representation and the semantic patterns unique to Magic: The Gathering.

The structured metadata like converted mana cost, colour identity and rarity contributed complementary information that the textual data could not capture. These features helped to encode the certain constraints and archetypes that every card has. These relatively clean numerical attributes, making them important contributions to the fused representation.

Artwork embeddings played a smaller role despite containing a lot of rich, visual information. The artwork of a card is not directly associated with relative strength and does not encode gameplay information, this was noted during the effects that image embeddings had on label propagation. The removal of the image embeddings allowed for the propagation of a far more stable graph.

The fusion of three separate modalities was successful in that a single unified representation space with a meaningful structure was built. This was evident after the use of UMAP and PCA visualisations showed cards with similar mechanical identities, mana curves and archetypes clustered nearby one another without supervision. The clustering behaviour validates one of the core assumptions of this research, that multimodal embeddings can encode the manifold of a card's gameplay power. However, the fused feature space is not perfect, some card types, with certain mechanics continue to cluster poorly within the graph-based structure. This is likely a reflection on both the small size of the labelled subset and the complexity of representing all cards with diverse mechanics.

5.2 Effectiveness of Anchor-Based Label Propagation

One of the most difficult challenges of this research was the absence of ground-truth labels within Magic: The Gathering. During the first iteration of SPELLS only 59 cards were labelled manually, representing only 0.22% of the total number of cards within the dataset. After the first iteration of active learning a total of 77 cards had been labelled, meaning 0.29% had now been given labels. If the pipeline was relying on a fully supervised architecture this number of labels would be insufficient for model training, but the semi-supervised approach using anchor-based propagation proved capable of utilising this data sparsity efficiently to infer labels for over 26,000 cards. The final number of cards given manual labels after the corrective iterations was 106 or 0.4%, still only a tiny fraction of the total number of cards.

The qualitative and quantitive behaviour of the propagated labels aligned with the manifold assumption: cards that are close in the manifold embedding space should receive a similar propagated power score. However, when running the label propagation process some cards saw some issues. *Oko, Thief of Crowns* continually received a lower propagated score than expected due to it being embedded nearer clusters of efficient but low powered cards. This issue was not contained to a single card, and highlights the limitation of local neighbourhood consistency issues. Often high-power outliers were rounded down when surrounded by moderately strong neighbours.

Despite some issues with label sparsity and local neighbourhood outliers the propagated labels were well structured and effectively supported the training of multiple downstream models. Through the combination of dimensionality reduction and label propagation strong results with reduced noise in the embedding space were created.

5.3 Downstream Model Performance

Three supervised regressor models were trained for the testing of the SPELLS architecture. A deep neural network, XGBoost and a Support Vector Regressor were empirically tested to give a broad view of performance and results. The first stage of evaluation for these models was the manually annotated subset. On this subset of labelled data all three models performed very well with MAE values less than 1.0. As the power scale is rated between 1.0 and 10.0, these results indicated that the learned representation and propagated labels provided a strong basis for approximating the human labelled values.

- The deep neural network achieved an MAE value of 0.93 and an R^2 of 0.80, showing that it was able to capture a large portion of the variance in manually labelled values.
- XGBoost obtained a lower MAE of 0.62 and a slightly lower R^2 score of 0.76, suggesting that this model captured a high local accuracy, but struggled with mapping the global variance when compared to the DNN.
- Finally SVR achieved the best Mean Absolute Error with 0.58 and an R^2 of 0.79, indicating that even on a relatively small, manually labelled set a linear-model regressor like a Support Vector Regressor can fit the anchor labels competitively.

These results prove that even with a small number of manually labelled samples, the combination of multimodal embeddings and label propagation can provide a set of features that are learnable by a variety of different models. Even with these results the gap between trained models and human labels reflect the difficulty of the task at hand. Manual annotation of card power is a noisy, time consuming process and is often context dependant. SPELLS, on the other hand, attempts to learn a single continuous scale for grading Magic: The Gathering cards.

In contrast the second method of evaluation evaluated the same three models against label propagation. Scores based on a cluster representative sample were drawn from the full dataset. This method of evaluation showed that both the DNN and the XGBoost models demonstrated strong alignment with the propagated manifold geometry, with both having high R^2 values. Their Spearman and Kendall correlation results were also impressive with both achieving above 0.96 and 0.89 respectively. This indicates that both models were highly capable at reconstructing the geometry of the label propagation and could order cards from weak to strong across the clustered structure. Unfortunately the SVR model did not perform well under this evaluation with substantially lower results across all metrics. This suggests that while it can model small manually labelled datasets, it struggles with noisy, dense multimodal representations, and is unable to maintain stable predictions of local clusters.

Looking at these results it shows that SPELLS supports two different but complementary views: meaningful alignment with the manually labelled anchors, and strong consistency with the semi-supervised learning manifold geometry created by the anchor-based label propagation. The DNN and XGBoost models both emerged as the high performers and most reliable models trained across both settings. The Support Vector Regressor was unable to capture the local neighbourhoods of the learned multimodal representation space and is unsuitable for SPELLS.

5.4 Limitations

Despite the promising performance that SPELLS has shown through its results during evaluation there are a number of limitations that keep it from achieving even stronger results across the multiple models trained.

Limited Volume of Human Labelled Data
Only a small number of cards were manually labelled with ground-truth power values. Although this aligns with the semi-supervised design of SPELLS, it does restrict the anchor points available to guide the label propagation process. The quality of the propagated labels depends on these manually labelled anchors; any gaps or biases within the small samples of labelled cards could have a major impact on the effectiveness of SPELLS by completely distorting clustered neighbourhoods in the embedding space.

Label Propagation's Sensitivity to Neighboured Clusters
The anchor-based approach to label propagation used throughout SPELLS provides significant benefits in reducing the computational complexity, but it is sensitive to the local neighbourhoods created within the manifold geometry. Cards that exist within sparse regions of the embedding space, or cards that have noisy embeddings due to a lack of feature information, may be incorrectly pulled towards high confidence anchor nodes nearby in the graph. This was shown in a number of cards which were considered outliers, with propagated power values for powerful cards getting unexpectedly low scores or relatively low powered cards such as *Mudhole* getting a score of 10.0. These behaviours highlight how fragile the graph-based propagation can be when the fused multimodal embeddings are not finely tuned.

Noise in Multimodal Representation Learning Space
Even though a lot preprocessing was done on the dataset the multimodal embeddings still contain noise. The image embeddings often contain irrelevant artwork features that have zero indication of power or ability. The oracle text, although thoroughly cleaned, might still include awkward phrasing or mechanical repetitions. The structural data may well overweight, or underweight certain aspects of the cards categorical and numerical features. Any type of noise within the learned representational space directly influences the geometry of the clustering and propagation of labels, effecting the downstream training of models.

Evaluation Constraints
This first method of evaluation only measures the performance of the models against the manually labelled cards. This means that the R^2 and other metrics are computed against a subset, consisting of less than 1% of the total number of cards. The second approach evaluates models against the propagated_score values generated during label propagation, measuring the internal consistency of the graph structure rather than real-world predictive accuracy. Having a much large sample size of manually annotated cards would likely provide a much stronger evaluation framework, but this was not possible within the scope of this project.

Generalisation Across Formats
The notion of card *power* is still inherently dependant on the format in which a card

is being played in. SPELLS attempts to learn a single universal power metric across all contexts. This simplification could potentially miss format specific interactions and might limit the applicability of SPELLS and predictions in real gameplay situation.

5.5 Implications

Through the implementation and evaluation of SPELLS, I observed that a semi-supervised multimodal learning framework can capture detailed information from card features in order to approximate human judgement of a card's strength even when explicit labels are scarce. The results suggest a number of things: implications for machine learning research and future work into representation learning, also for practical applications in gaming and across the area of game design.

The findings from this research using SPELLS highlights a potential for the partial automation of systems within game design and development pipelines. SPELLS has the potential to a useful and practical tool utilised during the costly and time consuming design process, saving gaming companies hundreds of hours while reducing the need for manual playtesting and heuristic judgement. By accelerating early prototyping stages, SPELLS allows designers to identify structural issues, whether that is proposed cards falling too closely within existing archetypes or through the introduction of unwanted strength spikes within the new set or existing format. This reduction in issues during the development pipeline long before formal playtesting begins gives designers the opportunity to explore creative card possibility quickly with reduced risk.

For game development companies like Wizards of the Coast, a system such as SPELLS represents the potential automation of card design and a reduction in time spent playtesting. This directly correlates with a reduced costs throughout the development lifecycle leading to a possible increase in profit. The ability to estimate power levels automatically allows companies to predict the cards that may be dominant, leading to undesirable effects within the player-base. Avoiding negative community sentiments is critical for companies such as Wizard of the Coast as the financial and reputational costs of retroactive fixes are great.

The results show that the fusing of oracle text, card images and structured data, produces a representation that captures functional relationships between cards. This validates the feasibility of semantic similarity between modalities in complex, rule-based domains. This type of research aligns closely with broader multimodal advances in natural language processing and areas of computer vision.

Although Magic: The Gathering lacks any dataset containing power levels for cards, SPELLS shows that a combination of graph-based label propagation and deep learning can create pseudo-labels from only a small subset of manual labels. This showcases the practicality of semi-supervised learning in specialised, expert-driven fields where the cost of manual annotation is a hurdle to the adoption of artificial intelligence systems.

A tuned version of SPELLS could possibly act as an assistant to game designers through the estimation of new cards and how they compare to existing ones before any playtesting has to occur. This has clear applications in balancing, game simulation and early prototyping. This would be particularly effective when hundreds or even thousands of cards required analysing and evaluation.

As SPELLS relies on the fusion of multimodal and graph-based semi-supervised learning, the architecture could potentially be transferred to other domains. Extending the system to other trading card games such as Hearthstone, Lorcana or Pokemon would increase the number of samples SPELLS would have to train on. Other such board games with structured components could also be benefit from SPELLS. Video game systems with highly complex relationships between items and character abilities could make use of SPELLS with only minor changes. Another area where the framework could be extended is that of recommender systems with scarce human labelled data. The generic but modular pipeline implemented by SPELLS has the potential to generalise well and could be applied to other industries.

As the project was being designed and implemented in real-time, the importance of high quality representational data was clear. Improvements to the oracle text cleaning pipeline, the supertype and subtype embeddings and changes made to the label propagation processes produced meaningful performance gains. These changes resulted in the clear observation that semi-supervised learning is only as good as the feature space that it is propagated through. Label propagation cannot produce effective results if the fused embeddings are of a low quality. Through the changes made and understanding that semantically aligned embeddings were essential for successful label propagation, it was clear that with such a small sample of ground-truth labels, an approach to correcting outliers was required. The introduction of active learning greatly improved the results of both the neural network and the propagated values. This demonstrated that iterative human corrections were a powerful step towards effective labelling, especially when errors made in the propagation process can negatively effect neighbourhoods within the graph structure.

6 Conclusion and Future Work

This research introduced SPELLS, a semi-supervised multimodal learning framework designed to estimate the latent power level of Magic: The Gathering cards, using encoded features from within the cards themselves, without the need for statistical gameplay data. The work shows that meaningful predictive performance is possible by combining human expertise and automated representation learning, even in complex domains where rule text, card art and structured features all contribute to conceptual card strength.

The results of this research indicate that the multimodal embeddings captured complex relationships between cards, with the deep neural network and XGBoost models performing well when evaluated against both the annotated ground-truth labels, and on a stratified subset of representative cards taken from clusters within the dataset. The best performing models achieved low MAE scores and high rank correlations when tested. Although label propagation introduced some localised outliers leading to incorrectly labelled cards, the implementation of iterative human correction improved the propagated labels and the downstream performance of all models.

This work also highlights a number of broader implications. Firstly, that multimodal fusion is suited to structured, rules-based games with rich feature spaces. Secondly, semi-supervised learning offers practical solutions for domains where expert annotations are costly, time consuming and scarce. Finally SPELLS shows how human-in-the-loop refinement integrates perfectly with semi-supervised learning in order to see improvements to pseudo label quality and predictive accuracy.

The research done here provides a strong foundation for future exploration of Magic: The Gathering and artificial intelligence. Expanding of the manually labelled dataset for a much stronger propagation of anchors could help to provide better performance. Utilising a large language model, such as Llama-3.1-8B for the processing of oracle text could potentially further improve performance of SPELLS as the increase in understanding of complex semantic data is essential for success. A huge improvement to the architecture of SPELLS would be to add a level of interpretability that could give a much clearer understanding of what embeddings are effecting the model and how certain features may not be as important as previously thought.

In conclusion, SPELLS demonstrates that semi-supervised multimodal learning can successfully infer the latent power of Magic: The Gathering cards. This research bridges the gap between human understanding and machine reasoning in complex game systems. This framework is scalable, extensible and can be tuned for the analysis of other trading card games. SPELLS lays the groundwork for future research into explainable AI, data efficient models in gaming, video game design and beyond.

References

[1] A. Churchill, S. Biderman, and A. Herrick, “Magic: The gathering is turing complete,” Apr 2019. [Online]. Available: https://arxiv.org/abs/1904.09828v2

[2] N. Brown and T. Sandholm, “Superhuman ai for heads-up no-limit poker: Libratus beats top professionals,” *Science*, vol. 359, no. 6374, p. 418–424, Dec 2017. [Online]. Available: https://science.sciencemag.org/content/359/6374/418.full

[3] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, Henrique, J. Raiman, T. Salimans, J. Schlatter, and J. Schneider, “Dota 2 with large scale deep reinforcement learning,” *arXiv (Cornell University)*, Jan 2019.

[4] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, and A. S. Vezhnevets, “Grandmaster level in starcraft ii using multi-agent reinforcement learning,” *Nature*, vol. 575, no. 7782, p. 350–354, Oct 2019. [Online]. Available: https://storage.googleapis.com/deepmind-media/research/alphastar/AlphaStar_unformatted.pdf

[5] C. D. Ward and P. I. Cowling, “Monte carlo search applied to card selection in magic: the gathering,” p. 9–16, Sep 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5286501

[6] M. Swiechowski, T. Tajmajer, and A. Janusz, “Improving hearthstone ai by combining mcts and supervised learning algorithms,” p. 1–8, Aug 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8490368

[7] T. Bertram, J. Furnkranz, and M. Muller, “Predicting human card selection in magic: The gathering with contextual preference ranking,” *2021 IEEE Conference on Games (CoG)*, Aug 2021.

[8] H. N. Ward, B. Mills, D. J. Brooks, D. Troha, and A. S. Khakhalin, “Ai solutions for drafting in magic: the gathering,” p. 1–8, Aug 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9619100

[9] A. Summerville and M. Mateas, “Mystical tutor: A magic: The gathering design assistant via denoising sequence-to-sequence learning,” *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 12, no. 1, p. 86–92, Jun 2021.

[10] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CALD-02-107, jun 2002.

[11] W. Guo, J. Wang, and S. Wang, “Deep multimodal representation learning: A survey,” *IEEE Access*, vol. 7, p. 63373–63394, 2019.

[12] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, “A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets,” *The Visual Computer*, Jun 2021.

[13] M. A. Manzoor, S. Albarri, Z. Xian, Z. Meng, P. Nakov, and S. Liang, "Multimodality representation learning: A survey on evolution, pretraining and its applications," *arXiv (Cornell University)*, Jan 2023.

[14] M. Campbell, A. Hoane, and F.-h. Hsu, "Deep blue," *Artificial Intelligence*, vol. 134, no. 1-2, p. 57–83, Jan 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370201001291

[15] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, p. 484–489, Jan 2016. [Online]. Available: https://www.nature.com/articles/nature16961

[16] A. Bud, A. Nicholson, I. Zukerman, and D. Albrecht, "A hybrid architecture for strategically complex imperfect information games," p. 42–45, Jan 2003.

[17] T. Bertram, J. Fürnkranz, and M. Müller, "Learning with generalised card representations for "magic: The gathering"," *2021 IEEE Conference on Games (CoG)*, p. 1–8, Aug 2024.

[18] A. K. Hoover, J. Togelius, S. Lee, and S. Fernando, "The many ai challenges of hearthstone," 2019. [Online]. Available: https://arxiv.org/abs/1907.06562

[19] H. N. Ward, B. Mills, D. J. Brooks, D. Troha, and A. S. Khakhalin, "Ai solutions for drafting in magic: the gathering," p. 1–8, Aug 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9619100

[20] P. Garcia-Sanchez, A. Tonda, G. Squillero, A. Mora, and J. J. Merelo, "Evolutionary deckbuilding in hearthstone," *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, Sep 2016.

[21] P. I. Cowling, C. D. Ward, and E. J. Powley, "Ensemble determinization in monte carlo tree search for the imperfect information card game magic: The gathering," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 4, p. 241–257, Dec 2012.

[22] M. Świechowski and T. Tajmajer, "A practical solution to handling randomness and imperfect information in monte carlo tree search," *Annals of Computer Science and Information Systems*, Sep 2021.

[23] V. Tomar, S. Sharma, S. Arora, and A. D. Tyagi, "A comprehensive analysis of techniques and applications in multimodal deep learning," *International Conference on Computing, Sciences and Communications*, p. 1–5, Oct 2024.

[24] M. Lymperaiou and G. Stamou, "A survey on knowledge-enhanced multimodal learning," *Artificial Intelligence Review*, vol. 57, no. 10, Sep 2024.

[25] Z. Yan, S. Li, Z. Duan, and Y. Zhao, "Robust deep semi-supervised learning with label propagation and differential privacy," *Frontiers in Computer Science*, vol. 5, May 2023.

[26] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," *arXiv:1904.04717 [cs]*, Apr 2019. [Online]. Available: https://arxiv.org/abs/1904.04717

[27] M. Farrokhi Maleki and R. Zhao, "Procedural content generation in games: a survey with insights on emerging llm integration," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 20, no. 1, p. 167–178, Nov 2024.

[28] Q. Liu, J. Hu, Y. Xiao, X. Zhao, J. Gao, W. Wang, Q. Li, and J. Tang, "Multimodal recommender systems: A survey," *ACM Computing Surveys*, Sep 2024.

[29] Z. Peng, G. Zheng, and W. Huang, "Semi-supervised learning with close-form label propagation using a bipartite graph," *Symmetry*, vol. 16, no. 10, p. 1312, Oct 2024.

[30] J. Lee, Y. Kim, and S. B. Kim, "Noise-robust graph-based semi-supervised learning with dynamic shaving label propagation," *Applied soft computing*, vol. 142, p. 110371–110371, Jul 2023.

[31] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," Sep 2020. [Online]. Available: https://arxiv.org/abs/1802.03426

[32] N. Reimers and I. Gurevych, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*, Aug 2019. [Online]. Available: https://arxiv.org/pdf/1908.10084

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv*, Dec 2015.

[34] M. Lazarou, Y. Avrithis, G. Ren, and T. Stathaki, "Adaptive anchor label propagation for transductive few-shot learning," *2023 IEEE International Conference on Image Processing (ICIP)*, p. 331–335, Oct 2023.

[35] T. Chen and C. Guestrin, "Xgboost: a scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, vol. 1, no. 1, p. 785–794, Aug 2016.