CLUSTERING CHESS PLAYING STYLES: A MACHINE LEARNING APPROACH



Conduct by:

Name : Hudaifah Priohastono

Student ID : 21523153

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM INDONESIA

2026

SUPERVISOR ENDORSEMENT PAGE

CLUSTERING CHESS PLAYING STYLES:
A MACHINE LEARNING APPROACH

THESIS

Conduct By:

Name : Hudaifah Priohastono
Student ID : 21523153

Yogyakarta, July 2, 2026
Advisor,

(Dr. Feri Wijayanto, S.T., M.T.)

EXAMINER ENDORSEMENT PAGE

CLUSTERING CHESS PLAYING STYLES:
A MACHINE LEARNING APPROACH

THESIS

Has been defended in front of the examiners as on of the requirements to obtain a Bachelor of Informatics degree from the Undergraduate Program in Informatics at the Faculty of Industrial Technology, Universitas Islam Indonesia

Yogyakarta, July 2, 2026

Chair

Dr. Feri Wijayanto, S.T., M.T. _____

Examiner 1

Erika Ramadhani, S.T., M.Eng. _____

Examiner 2

Taufiq Hidayat, S.T., M.Sc., Ph.D. _____

Acknowledged by,
Head of Undurgraduate Program in Informatics
Faculty of Industrial Technology
Universitas Islam Indonesia

(DHOMAS HATTA FUDHOLI, S.T., M.Eng., Ph.D)

AUTHENTICITY STATEMENT

The undersigned:

Name : Hudaifah Priohastono

Student ID : 21523153

Final project with title:

CLUSTERING CHESS PLAYING STYLES:
A MACHINE LEARNING APPROACH

Stating that all components and contents in this final project are my own work. If in the future it is proven that some parts of this work are not my own work, the final project submitted as my own work is ready to be withdrawn and ready to bear any risks and consequences.

Thus this statement letter is made, hopefully it can be used properly.

Yogyakarta, Juli 2, 2026



(Hudaifah Priohastono)

DEDICATION

Bismillahirrahmanirrahim

In the name of Allah, the Most Gracious, the Most Merciful. All praise and gratitude are due to Allah SWT for all His grace, guidance, and blessings, which have enabled the author to complete this thesis successfully.

With the utmost humility, I dedicate this humble work to:

My beloved parents, as a token of my deepest respect, devotion, and gratitude for their endless prayers, love, sacrifices, patience, and unwavering support. May all their kindness be accepted as righteous deeds and rewarded with the best of blessings from Allah SWT.

My beloved family, who have always provided prayers, motivation, care, and support, enabling me to complete every step of this educational journey.

Dr. Feri Wijayanto, S.T., M.T., my thesis supervisor, for his invaluable guidance, direction, knowledge, and motivation, given with great patience throughout the process of compiling this thesis.

My closest friends and companions in this academic journey, namely Adam, Angga, Dimas, Helga, Crisna, Gito, Naufal, Fikri, Safik, Ganendra, and Farhan, for their support, encouragement, and for being wonderful companions in sharing experiences, discussing ideas, and helping each other throughout our studies until the completion of this thesis. The togetherness we have built has become a truly precious memory.

All the lecturers of the Informatics Study Program at Universitas Islam Indonesia, who have guided and imparted knowledge to me throughout my studies.

Lastly, I dedicate this work as a form of appreciation for every process, struggle, and effort I have gone through. May this achievement be the beginning of better steps in the future and bring benefits to many.

MOTTO

Obrigado a todos... SIUUUUU! (Cristiano Ronaldo)

Tidak pernah menyerah adalah sihirku. Lampaui batasmu, disini dan sekarang (Asta & Yami Sukehiro)

FOREWORD

All praise and gratitude are offered to Allah SWT for all His grace, guidance, and blessings, which have enabled the author to successfully complete this Final Project entitled *"Clustering Chess Playing Styles: A Machine Learning Approach."* May peace and blessings always be bestowed upon Prophet Muhammad SAW, who has led humanity toward a path filled with knowledge.

This Final Project was undertaken as one of the requirements for obtaining a Bachelor of Computer Science degree from the Informatics Study Program, Faculty of Industrial Technology, Universitas Islam Indonesia. This research aims to identify and cluster the playing styles of top world chess players using the K-Means Clustering algorithm based on quantitative features extracted from game data. It is hoped that the results of this research can contribute to the advancement of data mining, machine learning, and data analysis applications in the field of sports, particularly in the game of chess.

Throughout the process of compiling this Final Project, the author faced various challenges, both in the process of collecting game data, extracting features from Portable Game Notation (PGN) files, analyzing clustering results, and refining the report in accordance with academic standards. In addition, the author also encountered challenges in managing time during the research process and completing revisions. Thanks to the prayers, support, and guidance from various parties, all of these processes were successfully navigated until this research could be completed.

On this occasion, I would like to express my deepest gratitude to:

1. Allah SWT for all His grace, good health, ease, and strength granted throughout the process of compiling this Final Project.
2. My parents and my entire family, who have always provided endless prayers, love, motivation, and unwavering support.
3. Dr. Feri Wijayanto, S.T., M.T., as my thesis supervisor, for his guidance, constructive feedback, knowledge, and for dedicating his time to guide me throughout the research process and the compilation of this Final Project. (Note: The acknowledgment section is typically concluded with the author's signature.)
4. My closest friends Adam, Angga, Dimas, Helga, Crisna, Gito, Naufal, Fikri, Safik, Ganendra, and Farhan for being wonderful companions in discussions, sharing experiences, motivating one another, and always providing support and encouragement throughout my studies and the completion of this Final Project. The

togetherness and assistance you have provided have become incredibly precious memories and a great source of motivation for me.

5. All other parties who cannot be mentioned individually, for their assistance, support, prayers, and motivation, which have enabled this research to be completed successfully.
6. *Last but not least, I want to thank me for believing in me.I want to thank me for believing in me. I want to thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for never quitting. I wanna thank me for always being a giver and trying to give more than I receive. I want to thank me for trying to do more right than wrong. I want to thank me for just being me at all times.*

The author acknowledges that this Final Project still has various limitations and is far from perfect. Therefore, constructive criticism and suggestions are highly appreciated as valuable input for the evaluation and development of future research.

In closing, the author hopes that this Final Project will be beneficial to the readers, serve as a reference for future researchers, and contribute to the advancement of knowledge, particularly in the fields of data mining, machine learning, and chess game analysis.

Yogyakarta, July 2, 2026



(Hudaifah Priohastono)

ABSTRAK

Mapping chess playing styles is crucial for tactical evaluation and strategic preparation. However, conventional methodologies rely heavily on subjective expert-based assessments, lacking measurable quantitative standardization. This study addresses these limitations through a data-driven approach using the K-Means clustering algorithm within an unsupervised learning framework. The dataset comprises 135 classical chess games from nine top-ranked global players during late 2025, with 15 games per player. Move data from Portable Game Notation (PGN) files were computationally extracted into nine quantitative statistical features representing aggressiveness, mobility, and spatial control. The dataset underwent preprocessing using StandardScaler for feature standardization. The optimal number of clusters was determined through Elbow Method and Silhouette Score analysis. Experimental results indicate that K=6 achieved the highest Silhouette Score of 0.2103, successfully mapping playing styles into six distinct behavioral profiles: very defensive, defensive, balanced, positionally active, aggressive, and very aggressive. The findings demonstrate that K-Means clustering effectively identifies latent patterns in top-tier chess game statistics, providing a valid, transparent, and quantifiable framework for player classification, even with a highly selective sample size

Keywords: Chess, Top Global Players, K-Means Clustering, Unsupervised Learning, Playing Style.

GLOSSARY

Algoritma K-Means Clustering	A data clustering method that divides the data into a number of clusters based on their proximity to the centroid, so that members within a single cluster share similar characteristics.
Area Control	A measure indicating the number of squares on the chessboard that are within a player's piece's range of influence or attack.
Average Black Control	The average number of squares controlled by the black
Average White Control	The average number of squares controlled by the white pieces during the game.
Capture Count	The number of times an opponent's piece is captured during a single game.
Castling	A specific move in chess that involves moving the king and rook simultaneously to improve the king's safety.
Centroid	The center of a cluster, calculated as the average of all data points within that cluster, is used as a reference in the clustering process
Check Count	The number of moves that pose a direct threat (check) to the opponent's king during the game.
Cluster	A group of data that shares a higher degree of similarity in characteristics compared to data in other groups.
Control Difference	The average difference in territory control between White and Black, used to measure spatial dominance during the game.
Data Mining	The process of uncovering patterns, relationships, or important information from a dataset using statistical, computational, and machine learning techniques.
Dataset	A set of data used as a source of information in the analysis or modeling process.

Elbow Method	An evaluation method for determining the optimal number of clusters based on changes in the Sum of Squared Errors (SSE).
Piece Move Count	The number of moves made by pieces other than pawns during a game.
Portable Game Notation (PGN)	An international standard format for recording and documenting chess moves in text form.
Python	The high-level programming language used for feature extraction, data analysis, and the implementation of the K-Means algorithm in this study.
Silhouette Score	A measure of clustering quality that indicates the degree to which a data point resembles its cluster compared to other clusters.
Unsupervised Learning	Metode pembelajaran mesin yang mengolah data tanpa label untuk menemukan pola atau struktur tersembunyi di dalam data.

TABLE OF CONTENTS

TITLE PAGE i
SUPERVISOR ENDORSEMENT PAGE ii
EXAMINER ENDORSEMENT PAGE iii
AUTHENTICITY STATEMENT iv
DEDICATION v
MOTTO vi
FOREWORD vii
ABSTRAK ix
GLOSSARY x
TABLE OF CONTENTS xii
LIST OF TABLE xiv
LIST OF FIGURE xv
CHAPTER I INTRODUCTION 1
1.1 Research Background 1
1.1.1 The Importance of Selecting the Subject and Time Period 4
1.2 Research Limitations (Scope) 5
1.3 Problem Statement 5
1.4 Research Objective 5
1.5 Benefit of the Research 6
CHAPTER II LITERATURE REVIEW 7
2.1 Conceptualization of amodern Chess 7
2.2 The Theory of Playing style in Chess 8
2.3 Theoretical Comparative Analysis of clustering Algorithm 9
2.4 Theoretical Foundations of Quantitative Features in Chess 9
2.5 Algorithma K-Means Clustering 11
2.6 Metode Evaluasi Clustering 12
2.6.1 Elbow Method and the Sum of Squared Errors (SSE) 12
2.6.2 Silhoute Coefficient (Silhoute Score) 12
2.7 Previous Research 13
CHAPTER III RESEARCH METHODOLOGY 15
3.1 Overview and Type of Research 15

3.2 Research Framework....15
3.3 Dataset and Features Extraction Specification....16
3.3.1 Technical Implementation of Feature Extraction and Modeling....17
3.4 Dataset Prepocessing stage....18
3.5 Portable Game Notation (PGN) Format and Data Structure....19
3.6 System Architecture and Data Processing Workflow....20
3.7 Mathematical Implementatio and Configuration of the K-Means Algorithm....21
CHAPTER IV RESULT AND DISCUSSION....22
4.1 Determining the Optimal Number of Cluster (K)....22
4.2 Cluster Centroid Analysis....25
4.3 Data Distribution into Cluster....27
4.4 Intrepretation of Playing Style Characteristic Based om Cluster Result....29
4.5 Mapping the Playing Styles of Top Global Players....32
4.5.1 Fabiano Caruana (Dominan *Balanced*)....32
4.5.2 Hikaru Nakamura (Dominan Balanced)....32
4.5.3 Ian Nepomniachtchi (Dominan *Balanced*)....33
4.5.4 Anish Giri (Transisi *Dual-Threat*)....33
4.5.5 Alireza Firouzja (Transisi *Dual-Threat*)....34
4.5.6 Wesley So (*Hybrid Stabilizer*)....34
4.5.7 Maxime Vachier-Lagrave (*Hybrid Stabilizer*)....34
4.5.8 Magnus Carlsen (The Ultimate Positional Grinder)....35
4.5.9 Ding Liren (Dominan Aggressive)....35
CHAPTER V CONCLUSION AND RECOMMEDATION....38
5.1 Conclusion....38
5.2 Suggestions....39
REFERENCES....41
APPENDICES....44

LIST OF TABLE

Table 0.1 Matrix of Previous Research.....26
Table 0.2 shows the results of the Silhouette Score tests for various values of K.....37
Table 0.3 Frequency Distribution of Matches by Cluster.....39
Table 0.4 Playing Styles of the World's Top Players.....45

LIST OF FIGURE

Figure 0-1 Example of the Raw Structur of a PGN File.....32
Figure 0-2 Elbow method curve.....36

CHAPTER I
INTRODUCTION

1.1 Research Background

For centuries, chess has stood as one of the most intellectually demanding strategic games in human history, requiring a sophisticated combination of tactical calculation, spatial reasoning, and long-term planning (Setiawan. & Pratama., 2021)The dynamics of the game, played across a 64-square board with 32 pieces governed by specific movement rules, generate an astronomical number of possible positions—estimated to be approximately 10^43 legal configurations, a figure that surpasses the estimated number of atoms in the observable universe (Chowdhary et al., 2023).Within this immense complexity, each player, whether consciously or unconsciously, develops a distinctive approach to decision-making that is consistently manifested through their choice of openings, their tendency toward piece exchanges, their preferred strategic structures, and their underlying philosophy of piece placement. It is precisely this consistency in behavioral patterns that gives rise to the concept of "playing style," a fundamental dimension of chess that distinguishes one player from another and has been the subject of extensive discussion within the chess community for generations.

Traditionally, chess playing styles have been categorized into several primary spectrums based on the qualitative assessments of historical masters, coaches, and commentators(Chowdhary et al., 2023). The most commonly referenced taxonomy includes three broad categories: the tactical or aggressive style, characterized by a focus on direct attacking initiatives, the generation of complex tactical complications, active piece exchanges, and the relentless pursuit of the opponent's king; the positional or strategic style, which prioritizes long-term spatial domination, the minimization of inherent pawn structure weaknesses, the control of key squares (outposts), and the application of gradual, incremental pressure; and the defensive style, which emphasizes the establishment of a robust defensive structure, risk minimization, and the exploitation of counter-attacks when the opponent commits a blunder induced by frustration. These three categories have served as the conventional framework through which players, analysts, and commentators have understood and described the diverse approaches to the game for decades.

However, despite the widespread acceptance of this three-category framework, a fundamental question remains inadequately addressed within the existing literature: is it truly accurate to assert that there are only three playing styles in chess? This inquiry is not merely

semantic but carries substantial analytical implications. If the spectrum of playing styles is indeed limited to only three broad categories, then the classification system is arguably too coarse to capture the nuanced differences between players who may share the same general label yet exhibit markedly different strategic preferences. For instance, two players both classified as "aggressive" might differ fundamentally in their approach—one might prioritize rapid piece development and early tactical strikes, while another might focus on gradual positional pressure that culminates in a decisive attack. Conversely, if the spectrum encompasses more than three distinct styles—perhaps including hybrid or intermediate profiles such as "balanced," "positionally active," or "very aggressive"—then the traditional framework requires significant refinement to reflect the empirical reality of how elite players actually approach the game (Lin et al., 2024),

This uncertainty is compounded by several methodological challenges inherent in the traditional approach to style classification. First, the process of identifying and labeling playing styles has historically relied heavily on manual evaluations conducted by chess coaches, commentators, and historians (McIlroy-Young et al., 2020). Such expert-based assessments are inherently limited by a high degree of subjectivity; consequently, two chess analysts may arrive at divergent conclusions regarding a player's playing style due to biases stemming from their respective personal experiences, theoretical preferences, or interpretive frameworks. Second, there is a notable lack of standardized, measurable metrics that can be consistently applied across different players and time periods (McIlroy-Young et al., 2020)Without such metrics, the classification of playing styles remains qualitative and impressionistic rather than quantitative and rigorous. Third, the traditional framework does not adequately account for the dynamic and fluid nature of playing styles, which may evolve over the course of a player's career or shift depending on the specific demands of a given position or opponent.

The significance of resolving this question extends beyond academic curiosity. For competitive players and coaches, a precise and objective understanding of the spectrum of playing styles is essential for effective pre-match preparation, enabling the formulation of strategic traps designed to lure opponents out of their comfort zones. For educators and trainers, a refined classification system can facilitate more targeted instruction, helping students identify their natural strengths and address their strategic weaknesses. For researchers in the broader field of sports analytics and behavioral analysis, chess serves as an ideal model system for studying the structure of human decision-making in complex, rule-governed environments (Brefeld et al., 2025), and a rigorous taxonomy of playing styles would substantially enhance

the analytical power of such investigations (McIlroy-Young et al., 2021) Furthermore, the contemporary chess landscape has undergone a paradigm shift due to the advent of neural network-based chess engines, most notably AlphaZero and Leela Chess Zero (LCZero). These artificial intelligence systems have fundamentally rewritten traditional chess dogmas, demonstrating that long-term piece activity, dynamic pawn sacrifices, and king safety often hold greater objective value than static material advantages or rigid central occupation. Consequently, modern Grandmasters have increasingly adapted their opening repertoires and middlegame strategies to mirror this engine-inspired 'meta.' They now favor flexible, hybrid systems that allow for gradual accumulation of micro-advantages, blurring the lines between the classical 'tactical' and 'positional' dichotomies. This evolution renders the traditional three-category framework even more obsolete, necessitating a rigorous, data-driven computational approach to accurately capture and quantify these nuanced, modern strategic behaviors.

In light of these considerations, this study seeks to address the fundamental question of how many distinct playing styles can be objectively identified within the competitive chess landscape, with a particular focus on the highest echelon of the game. To achieve this objective, the research employs a data-driven approach grounded in the principles of data mining and unsupervised machine learning (Cheng, 2024; Samara et al., 2025). Specifically, the K-Means Clustering algorithm is utilized to partition a dataset of quantitative behavioral features extracted from elite-level chess games into a predetermined number of clusters, each representing a distinct playing style profile. The selection of K-Means as the primary analytical tool is predicated on its demonstrated effectiveness in partitioning numerical data with distinct spatial distributions, its computational efficiency, and its capacity to reveal latent patterns within complex datasets without requiring prior labeling or human intervention (Ahmed et al., 2020; Chong, 2021; Sinaga & Yang, 2020)

By systematically extracting nine quantitative statistical features—encompassing dimensions of tactical aggressiveness, piece mobility, king safety, and spatial control—from the move records of top-ranked global players, this study aims to mathematically deconstruct the distribution of micro-tactical characteristics and empirically determine the optimal number of playing style clusters that best captures the diversity of strategic approaches at the elite level (Kalagara, 2024; Samara et al., 2025). Through this rigorous analytical framework, the research aspires to provide a valid, transparent, and quantifiable taxonomy of chess playing styles, thereby advancing both the theoretical understanding of strategic diversity in chess and the practical applications of this knowledge in competitive and educational contexts.

1.1.1 The Importance of Selecting the Subject and Time Period

The selection of research subjects and the timeframe for data collection in this study was not arbitrary; rather, it was driven by a fundamental academic imperative to empirically resolve a long-standing question in chess theory: how many distinct playing styles actually exist at the highest echelon of competitive chess? While conventional wisdom has traditionally categorized chess playing styles into only three primary spectrums (tactical, positional, and defensive), this tripartite framework has never been rigorously validated through quantitative computational analysis. By analyzing data from a carefully curated cohort of elite players, this study seeks to transcend the limitations of qualitative expert assessments and provide an empirically grounded answer to this fundamental question (McIlroy-Young et al., 2021)The concept of behavioral stylometry—the computational identification of unique decision-making patterns from chess moves—has recently emerged as a powerful framework for studying individual playing characteristics (McIlroy-Young et al., 2022), providing the theoretical foundation for the present investigation.

The restriction of the research subjects to nine top-ranked global players is underpinned by robust methodological justification. At the elite Grandmaster level, the margin of error is exceptionally narrow, and tactical consistency is maximized. If a clustering algorithm can successfully differentiate and categorize playing styles at a level where game variations are highly nuanced, it substantiates the sensitivity and validity of the constructed model. These nine players were specifically selected because they serve as standard-bearers across the diverse spectrum of modern chess philosophies, encompassing highly tactically sharp players (e.g., Ding Liren, Alireza Firouzja), adaptive universal players (e.g., Fabiano Caruana, Hikaru Nakamura, Ian Nepomniachtchi), and exceptionally solid positional players (e.g., Magnus Carlsen, Wesley So). This deliberate heterogeneity ensures that the resulting clusters capture the full breadth of strategic diversity present at the elite level, rather than being biased toward a single dominant playing philosophy (Kalagara, 2024; Lin et al., 2024). Furthermore, the selection of players who have consistently competed at the highest level during the mature post-AlphaZero era ensures that the identified playing styles reflect the contemporary "meta" of modern chess theory, which has been profoundly transformed by neural-network-based engine preparation (Tang et al., 2024; Zhang et al., 2025)

By rigidly constraining the sample size to 15 games per player (totaling 135 games), this study maintains a balanced data distribution to prevent the K-Means algorithm from exhibiting

bias toward players with a larger volume of games. This highly selective constraint ensures that each resulting cluster authentically reflects the pure strategic preferences of elite players, rather than merely capturing the statistical noise associated with amateur or intermediate players, whose performance consistency remains fluctuating (Chowdhary et al., 2023; Setiawan. & Pratama., 2021)The balanced design also enables a fair comparative analysis across all nine subjects, allowing the clustering algorithm to identify genuine stylistic differences rather than artifacts of unequal sample sizes. Through this high-level case study approach, the research aims to mathematically deconstruct the distribution of micro-tactical characteristics, thereby developing a playing style identification system that is valid, reliable, and entirely devoid of human interpretive bias (Cheng, 2024; Samara et al., 2025)

1.2 Research Limitations (Scope)

To ensure that the discussion within this study remains focused and strictly aligned with its predetermined objectives, the research scope is rigidly defined as follows:

a. Data Source: The dataset utilized in this study is strictly limited to digital files in the Portable Game Notation (PGN) format, which contain the historical records of chess moves.
b. Data Sample Volume: The volume of data analyzed is selectively constrained to a total of 135 chess game records (comprising 15 games per player to maintain a balanced dataset distribution).
c. Research Subjects: The games analyzed are exclusively derived from the performance of nine top-ranked global players.

1.3 Problem Statement

Based on the comprehensive descriptions elaborated in the background and problem scope sections, the core research problems to be addressed in this study are formulated into the following academic research questions:

a. Is it true that there are only three styles of chess?

1.4 Research Objective

The specific objectives to be systematically achieved through the conduct of this thesis research are as follows:

a. To determine the optimal number of chess playing style clusters among elite global players by applying the K-Means Clustering algorithm integrated with the Elbow Method and Silhouette Score evaluation techniques.
b. To identify and characterize the distinct playing style profiles exhibited by the top global players based on the quantitative features extracted from their chess game records.

1.5 Benefit of the Research

The implementation of this research is expected to yield tangible contributions, which are delineated into two primary domains: practical and academic benefits.

a. **Practical Benefits:** For Chess Players and Coaches: The findings provide an empirically grounded taxonomy of chess playing styles that transcends the traditional three-category framework (tactical, positional, and defensive). By revealing the actual number and characteristics of distinct playing styles, players can gain deeper self-awareness of their strategic tendencies, while coaches can design more targeted training programs tailored to specific stylistic profiles.
b. **Academic Benefits**: Contribution to Sports Analytics Literature This study directly addresses a fundamental question in chess analytics—"How many distinct playing styles exist?"—thereby expanding the scientific understanding of behavioral diversity in competitive strategic games. The empirical validation that playing styles extend beyond the traditional three categories represents a significant theoretical advancement.

CHAPTER II
LITERATURE REVIEW

2.1 Conceptualization of amodern Chess

Chess is a two-player strategic board game played on an 8x8 grid comprising 64 alternating black and white squares. Each player begins the game with 16 pieces, each governed by specific rules of movement. Historically and theoretically, the progression of a chess game is delineated into three crucial phases: the opening, the middlegame, and the endgame. At the elite global echelon, the transitions between these phases reflect the depth of tactical calculation and the underlying playing philosophy adhered to by the respective player(McIlroy-Young et al., 2020). In the modern era, the conceptualization of chess playing styles has evolved significantly from anecdotal observations by historical masters to more structured taxonomic frameworks. The most widely recognized and structured theory in this domain was developed by Grandmaster Lars Bo Hansen, who introduced a comprehensive framework to categorize elite players based on their decision-making tendencies, risk preferences, and strategic priorities. Hansen's taxonomy provides a foundational lens for understanding the diverse approaches at the highest level of the game, moving beyond simplistic labels to capture the nuanced psychological and strategic dimensions of competitive chess. This framework has since been widely adopted by coaches, commentators, and chess educators as a standard reference for analyzing human behavior over the board.

Within this theoretical framework, playing styles are generally understood to exist along a dynamic spectrum of strategic philosophies rather than as rigid, isolated categories. On one end of this spectrum lies the proactive and tactical approach. Players exhibiting this style prioritize initiative, piece activity, and direct attacks, frequently generating complex tactical sequences and embracing calculated risks to overwhelm their opponents. They are often willing to sacrifice long-term structural integrity for immediate dynamic compensation. Conversely, on the opposite end of the spectrum is the highly positional and defensive approach. This style is characterized by a meticulous search for the objective truth of the piece structure, prioritizing the minimization of inherent weaknesses and the gradual accumulation of micro-advantages over speculative ventures. Players adopting this posture often unfold their games slowly, emphasizing prophylaxis, spatial control, and the systematic restriction of the opponent's counterplay, reflecting the classical principles of solid defensive construction.

Bridging these two extremes is the flexible, results-oriented approach. Players who embody this style do not adhere rigidly to a single strategic philosophy; instead, they excel at adapting to the specific demands of each position. They are highly skilled at exploiting the opponent's psychological and temporal vulnerabilities, seamlessly transitioning between tactical sharpness and positional solidity based on the objective requirements of the board. In contemporary chess, this adaptability is highly prized, as it reflects the modern understanding that flexibility and the ability to shift gears mid-game are the ultimate determinants of success at the elite echelon.

Despite the widespread acceptance of such taxonomies, it is important to acknowledge their fundamental limitations. Traditional classifications remain inherently qualitative and subjective, relying heavily on the interpretive judgments of historians and commentators rather than on quantifiable, reproducible metrics. Furthermore, these frameworks were largely developed prior to the advent of neural network-based chess engines, and therefore do not adequately account for the profound paradigm shift in strategic evaluation that has occurred in the contemporary era. It is precisely these limitations that motivate the present study. By employing a data-driven, unsupervised machine learning approach, this research seeks to transcend the subjectivity of traditional taxonomies and empirically validate whether the playing styles of elite Grandmasters can be more accurately captured through a refined, quantifiable clustering framework based on objective spatial and tactical metrics.

2.2 The Theory of Playing style in Chess

In the modern era, the most structured theory of chess playing styles was developed by Grandmaster Lars sBo Hansen (an economist and chess player) and has since been widely (Hansen, n.d.)

a. **The Activist :** Aggressive, proactive, tactical, and willing to take risks.
b. **The Pragmatist:** Flexible, focused on the end result, and skilled at exploiting the opponent’s time constraints.
c. **The Reflector :** Highly positional, defensive, seeking the objective truth of the piece structure.

2.3 Theoretical Comparative Analysis of clustering Algorithm

Within the discipline of data mining, the selection of a clustering algorithm must be grounded in a comprehensive understanding of the dataset's characteristics and the overarching objectives of the study.(Jan et al., 2022) To determine the optimal model for identifying chess playing styles within this specific dataset, the following presents a theoretical comparative analysis of the three most widely utilized clustering algorithms:

a. Algoritma K-Means Clustering

The fundamental operating principle of this algorithm involves partitioning the dataset into a predetermined number of *K* clusters, wherein each data point is allocated to the cluster with the nearest center based on spatial distance calculations. Its primary advantages lie in its exceptionally high computational efficiency and performance stability when processing clean numerical data. This algorithm is deemed highly suitable and optimal for the present thesis, given that the feature data extracted from the 15 games of top global players during the late 2025 period are characterized by an exceptionally high degree of tactical consistency. The performance characteristics of world-class players tend to form structured and stable coordinate patterns within the geometric data space, rendering the dataset highly amenable to linear execution by the K-Means algorithm to determine the center of mass (centroid) representing each distinct playing style.(Ahmed et al., 2020; Chong, 2021; Ikotun et al., 2023; Sinaga & Yang, 2020)

b. Algoritma DBSCAN (Density-Based Spatial Clustering)

DBSCAN partitions data based on spatial density, wherein regions of high density are consolidated into primary clusters, while isolated data points are classified as outliers or noise. This algorithm is deemed unsuitable for the present study, as the utilized dataset comprises meticulously curated records from official matches involving nine top-ranked global players. Given that elite-level competitive play is inherently devoid of data anomalies or statistical noise, a spatial density-based approach such as DBSCAN is rendered irrelevant for this specific analytical context.

2.4 Theoretical Foundations of Quantitative Features in Chess

To translate the qualitative and subjective nature of chess moves into quantifiable data amenable to processing by clustering algorithms, this study employs nine metrics that represent

the fundamental dimensions of chess gameplay. The selection of these variables is predicated on contemporary chess strategic theory(Silman, 2006):

a. **Tactical Aggressiveness Dimension :** Capture Count and Check Count Within the framework of combat theory adapted to chess, material exchanges (captures) serve as a direct measure of conflict intensity and positional sharpness (Shannon, 1950)The frequency of checks represents the initiative, tempo, and psychological pressure exerted upon the opponent's king. Players exhibiting an aggressive style tend to maximize both metrics to generate tactical complications.
b. **Mobility Dimension:** Pawn Move Count and Piece Move Count Mobility constitutes the essence of chess dynamics and was identified by Shannon as a primary evaluation metric. Pawn moves are frequently irreversible, fundamentally determining the long-term pawn structure and the creation of weak squares. Conversely, the movement of pieces (minor and major pieces) reflects tactical flexibility, piece redeployment, and the overall readiness to execute attacking or defensive maneuvers (Silman, 2006)
c. **King Safety Dimension:** Castling is the unique move in chess wherein two pieces are moved simultaneously, and the king is displaced by more than one square. This maneuver is exclusively designed to secure the king on the flank and activate the rook. At the elite echelon, a high frequency of castling signifies strict adherence to sound opening principles.
d. **Spatial Control Dimension:** Area Control, Average White/Black Control, and Control Difference Predicated on hypermodern theory ,control of the central squares need not be achieved solely through physical occupation by pawns; rather, it can be exerted from a distance via piece coordination. Recent computational approaches have successfully quantified this spatial "footprint" across the 64-square (Chowdhary et al., 2023)Concurrently, the Control Difference metric measures the asymmetry of spatial domination, which is critical for differentiating between positional players (who prioritize spatial control) and tactical players (who may concede space in exchange for attacking prospects) Furthermore, the operationalization of these spatial metrics requires a nuanced understanding of modern chess dynamics. Unlike classical chess theory, which emphasized direct physical occupation of the center with pawns, the hypermodern school of thought, pioneered by players like Aron Nimzowitsch and Richard Réti, posits that control can be exerted from a distance through piece coordination and fianchettoed bishops. The *Area Control* metric computationally

captures this exact phenomenon by evaluating the number of squares threatened or influenced by a player's pieces at any given ply, regardless of physical occupation. Concurrently, the *Average Control Difference* serves as a critical differentiator. A high positive control difference indicates a player's ability to restrict the opponent's piece mobility a concept known in chess literature as 'prophylaxis.' By quantifying these abstract strategic concepts into measurable numerical values, this study bridges the gap between traditional chess pedagogy and modern computational data mining, allowing for an objective assessment of a player's spatial dominance throughout the entirety of a match.

2.5 Algorithma K-Means Clustering

The K-Means algorithm stands as one of the most extensively utilized partitional clustering methods within the domain of data mining, attributed to its algorithmic simplicity, computational efficiency, and robust capacity to process large-scale numerical datasets. The primary objective of the K-Means algorithm is to partition *N* observations into *K* distinct clusters, wherein each observation is assigned to the cluster characterized by the nearest centroid (center of mass).

Mathematically, the K-Means algorithm seeks to minimize the Within-Cluster Sum of Squares (WCSS), also referred to as the objective function *J*, which is formally defined as follows:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} || \, x - \mu_i \, ||^2$$

Description:

J = The objective function to be minimized (total squared distance)

K = The predetermined number of clusters

C_i = The set of data points belonging to the i-th cluster

x = A data point (feature vector) within the cluster

C_i μ_i = The centroid (mean vector) of the i-th cluster

||x - μ_i||^2 = The squared Euclidean distance between the data point x and the centroid μ_i

The optimization of the aforementioned objective function is executed iteratively via an iterative optimization procedure, which comprises two primary steps:

a. **Assignment Step**: Assigns each data point x to the cluster C_i possessing the nearest centroid μ_i based on Euclidean distance calculations.
b. **Update Step**: Recalculating the position of the centroid μ_i for each cluster by computing the arithmetic mean of all data points assigned to the respective cluster.

2.6 Metode Evaluasi Clustering

Given that the K-Means algorithm operates within an unsupervised learning paradigm devoid of ground-truth labels, the evaluation of clustering quality must be conducted utilizing internal metrics. This study employs two primary validation methods to determine the optimal value of *K*:

2.6.1 Elbow Method and the Sum of Squared Errors (SSE)

The Elbow Method is a heuristic technique that evaluates the total variance within clusters. The metric employed is the Sum of Squared Errors (SSE), which is computed using the identical formulation as the K-Means objective function. As the value of K is incremented, the SSE value will invariably decrease. The optimal point is identified at the "elbow point," where the rate of SSE reduction begins to decelerate significantly (converge), signifying that the addition of further clusters no longer yields a substantial improvement in variance explanation.

2.6.2 Silhoute Coefficient (Silhoute Score)

To validate the results of the Elbow Method, the Silhouette Score is employed to quantify the degree of similarity of an object to its own cluster relative to the nearest neighboring cluster. The silhouette value, denoted as *s(i)* for an individual data point *i*, is computed using the following formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Description:

$a(i)$= The average distance between data point *i* and all other data points within the same cluster (cohesion).

$b(i)$= The average distance between data point *i* and all data points within the nearest neighboring cluster (separation).

The value of *s(i)* ranges from -1 to +1. A value approaching +1 indicates that the data point is situated in close proximity to its assigned cluster and significantly distant from neighboring clusters, thereby denoting a robust and optimal cluster separation.

2.7 Previous Research

To substantiate the theoretical foundation and establish the novelty of this study, a review of prior literature was conducted, encompassing research related to player behavior analysis, the application of data mining in strategic games, and clustering algorithms.(Jan et al., 2022; Qiu et al., 2024; Risnasari et al., 2022) The following matrix presents a comparative analysis of the foundational prior studies referenced in this research:

Table 2.1 Matrix of Previous Research

Researcher	**Research Subject**	**Method**	**Features / Variables**	**Research Results**	**Differences Compared to This Study**
McIlroy-Young et al. (2020)	Human Chess Move Data (Maia Project)	*Supervised Learning (Neural Networks)*	Move sequences and position evaluation	Successfully modeled and predicted human behavior based on individual behavior and ratings.	This study uses unsupervised learning (K-Means) to cluster playing styles at a macro level, rather than predicting moves on a turn-by-turn basis.
Qasthari & Kurniawan (2024)	Call of Duty Mobile Players	K-Means Clustering	Gameplay statistics (Kills, Deaths, Assists)	Successfully grouped gamers' playing styles into several behavioral clusters.	Different game genres (chess vs. FPS). This study uses more complex spatial features in chess, such as Area Control.
Tang et al. (2024)	Chess Match Data (Maia-2 Project)	Deep Learning / Reinforcement Learning	Board state evaluation	Developing an AI model that aligns with human behavior and error rates.	Focusing on aligning AI with humans. This research focuses on identifying playing

Researcher	Research Subject	Method	Features / Variables	Research Results	Differences Compared to This Study
					style clusters using PGN statistical features.
This Research	135 PGN Top Global Chess Players	K-Means Clustering	9 Quantitative Features (Capture, Check, Area Control, etc.)	Identifying 6 playing style profiles (from Very Defensive to Very Aggressive).	Focus on elite players (Grandmasters) by extracting spatial and tactical features from PGN files using an unsupervised approach.

Based on the aforementioned matrix of prior research, it is evident that the majority of computational studies in chess (McIlroy-Young et al., 2020; Tang et al., 2024)have predominantly focused on supervised learning or deep learning paradigms to predict moves or evaluate positions utilizing chess engines. Conversely, studies utilizing the K-Means algorithm to categorize playing styles (Qasthari & Kurniawan, 2024)remain largely confined to games within the shooter genre, which are characterized by relatively rudimentary statistical features.

Consequently, the novelty of this study lies in the application of the K-Means algorithm to categorize playing styles at the highest echelon of competitive play (elite global Grandmasters), utilizing quantitative features computationally extracted from PGN files. Notably, this encompasses complex spatial metrics such as Area Control and Control Difference. This methodological approach offers a novel perspective within the domain of chess sports analytics, providing a framework that is objective, quantifiable, and entirely devoid of human subjective bias.

CHAPTER III
RESEARCH METHODOLOGY

3.1 Overview and Type of Research

This thesis is designed utilizing an experimental quantitative approach integrated with a high-level case study methodology. This study is classified as quantitative research because all observational units, initially comprising textual chess move records from elite players, are systematically transformed into matrices of quantitative numerical values to enable precise computation utilizing linear algebra operations within the clustering model.

3.2 Research Framework

To ensure the validity of the experimental procedure, this study adopts a structured data mining framework(Jan et al., 2022), which is divided into the following sequential stages:

a. **Data Acquisition Phase (*Data Input*):** Collecting a raw database of Portable Game Notation (PGN) documents containing the historical move records of 135 chess matches played by nine top-ranked global players.
b. **Data Cleaning Phase (*Data Cleaning*):** Conducting a filtering process to eliminate non-behavioral metadata and purging manual commentary and analytical text annotations embedded within the move files.
c. **Feature Extraction Stage (*Feature Engineering*):** Executing a computational script to parse the textual move strings and calculate the frequency of gameplay activities, thereby generating nine quantitative numerical variables.
d. **Data Normalization Stage (*Data Preprocessing*):** Applying a mathematical transformation utilizing the StandardScaler to ensure that all feature variables are standardized to a uniform scale, yielding a distribution characterized by a zero mean and unit variance.
e. **K-Score Evaluation Phase (*Hyperparameter Tuning*):** Executing iterative model evaluations across a range of *K* values from 1 to 10, utilizing the graphical Elbow Method and the mathematical computation of the Silhouette Coefficient.

f. **Clustering Execution Phase (*Model Deployment*):** Executing the K-Means Clustering algorithm on the normalized dataset utilizing the determined optimal value of *K*.

g. **Centroid Interpretation and Analysis Stage:** Computing the mean values (centroids) of each feature within each resulting cluster to define the chess playing style characteristics exhibited by elite global players.

3.3 Dataset and Features Extraction Specification

From this total corpus of 135 matches, a Python script computationally extracts the complete move history, transforming it into thousands of rows of tactical behavioral data prepared for the subsequent modeling phase.

To ensure that the extracted metrics are theoretically grounded and not arbitrarily selected, the nine structured quantitative feature variables are systematically operationalized based on the four theoretical dimensions defined in The features are extracted and categorized as follows:

a. **Tactical Aggressiveness Dimension**

Capture Count: Quantifies the total frequency of material exchanges (capturing an opponent's piece) throughout the game, serving as a direct measure of conflict intensity and positional sharpness.

Check Count: Measures the total frequency of direct attacks (checks) delivered to the opponent's king, representing the initiative, tempo, and psychological pressure exerted during the match.

b. **Mobility Dimension**

Pawn Move Count: Records the total frequency of pawn advancements. Since pawn moves are frequently irreversible, this metric fundamentally reflects the shaping of long-term pawn structures and the creation of weak squares.

Piece Move Count: Tallies the total frequency of relocations by high-value pieces (knights, bishops, rooks, and queens), reflecting tactical flexibility, piece redeployment, and overall readiness to execute attacking or defensive maneuvers.

c. **King Safety Dimension**

Castling: A binary feature (assigned a value of 1 if castling is performed, and 0 otherwise) that captures the execution of this unique maneuver

designed to secure the king on the flank and activate the rook, signifying strict adherence to sound opening principles at the elite level

d. **Spatial Control Dimension**

Area Control: Measures the average number of squares on the chessboard influenced or attacked by the player's pieces throughout the game, quantifying the spatial "footprint" of the player's influence across the 64 squares.

Average White Control: Calculates the average number of specific squares controlled by the white pieces at each move.

Average Black Control: Calculates the average number of specific squares controlled by the black pieces at each move.

Average Control Difference: Computes the absolute value of the asymmetry in territorial dominance between White and Black throughout the game. This metric is critical for differentiating between positional players (who prioritize spatial control) and tactical players (who may concede space in exchange for attacking prospects).

3.3.1 Technical Implementation of Feature Extraction and Modeling

The entire workflow, encompassing feature extraction, data preprocessing, and clustering modeling, was implemented utilizing the Python 3.10 programming language and executed within the Google Colab cloud computing environment (a cloud-based Jupyter Notebook environment). The selection of this platform is predicated on its robust capacity to provision substantial computational resources and its high degree of accessibility for data mining research.

The processes of data manipulation and analysis were executed utilizing several core libraries:

a. **Pandas** and **NumPy**: It is utilized for the management of data structures, dataset manipulation, and efficient numerical operations in processing the statistical feature matrices of chess games.

b. **Scikit-learn**: The main library for implementing machine learning algorithms, which includes:

KMeans from the sklearn.cluster module: It is used to implement the K-Means clustering algorithm, using the

random_state=42 parameter to ensure the reproducibility of the results.

StandardScaler from the sklearn.preprocessing module: It is used for feature standardization to ensure that all variables have a mean of zero and a variance of one, thereby preventing variables with larger numerical scales from dominating the data**.**

silhouette_score from the sklearn.metrics module: It is utilized for evaluating the quality of cluster separation across various values of *K* (from *K*=2 to *K*=10).

c. **Matplotlib**: It is utilized for visualizing the Elbow Method curve and cluster distribution.

d. **Google Colab Files Module**: It is utilized to facilitate the uploading of datasets in Excel (.xlsx) format and the downloading of the analysis results.

The modeling process is executed iteratively for values of *K* ranging from 2 to 10. During each iteration, the K-Means algorithm performs a random initialization of the centroids, subsequently allocating each observation to the nearest centroid based on Euclidean distance calculations. This procedure is repeated until convergence is achieved. The Silhouette Score is computed for each *K* configuration to determine the optimal number of clusters that yields the most effective data separation.

3.4 Dataset Prepocessing stage

Given that the data are derived from world-class players who exhibit exceptionally high move efficiency, even minor numerical variations can profoundly impact the clustering process. The raw dataset must not be directly inputted into the K-Means model, as variables with wider value ranges (e.g., Piece Move Count) would automatically dominate the Euclidean distance calculations, thereby causing variables with smaller ranges yet high strategic significance (e.g., Check Count) to be marginalized by the algorithm.(Sinaga & Yang, 2020) To mitigate this scale disparity, a data standardization process is implemented utilizing the StandardScaler method, governed by the following formula:

$$\boldsymbol{z = \sigma x - \mu}$$

Description:

z= feature values after standardization

x= original feature values

μ= the mean of that feature

σ= the standard deviation of that feature

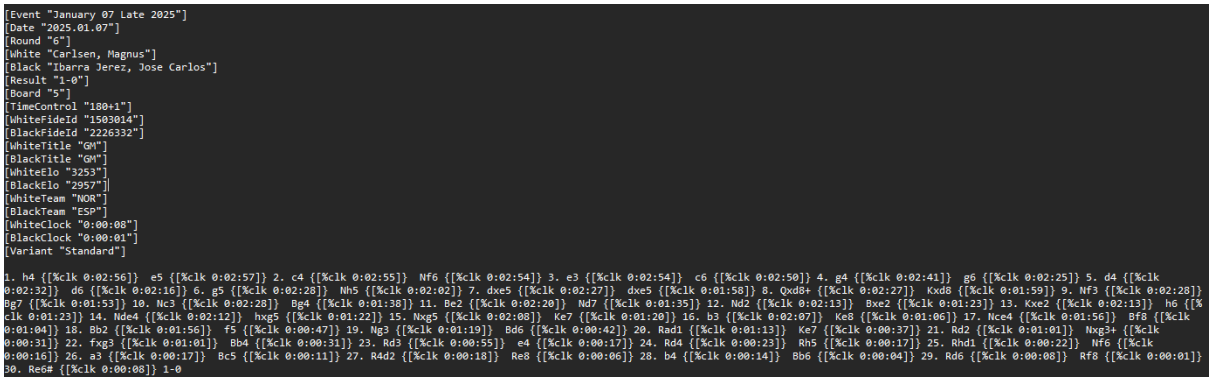
Through the implementation of this standardization, all quantitative aspects of the chess games played by the world's top players are now on an equal geometric footing (fair spatial scaling).

3.5 Portable Game Notation (PGN) Format and Data Structure

Prior to the execution of the feature extraction process, it is imperative for this study to comprehend the structure of the raw data utilized. The chess game data in this study are stored in the Portable Game Notation (PGN) format, which serves as the international *de facto* standard for representing chess move data in human- and machine-readable ASCII text. Originally developed by Steven Edwards in 1993, this format remains the primary standard utilized by global chess databases such as Chess.com and Lichess to this day.

Structurally, a PGN file comprises two fundamental components: the Tag Pair Section and the Movetext Section. The Tag Pair Section encapsulates the game's metadata enclosed within square brackets, encompassing the seven mandatory tags known as the Seven Tag Roster (STR). These include [Event "Tournament Name"], [Site "Location"], [Date "Date"], [Round "Round"], [White "White Player"], [Black "Black Player"], and [Result "Result"].

The second component is the Movetext Section, which encapsulates the sequential game moves utilizing Standard Algebraic Notation (SAN). Within this notation system, each move is denoted by the initial letter of the piece (K for King, Q for Queen, R for Rook, B for Bishop, and N for Knight), followed by the destination square; for instance, e4 denotes a pawn move, whereas Nf3 signifies a knight move.



```
[Event "January 07 Late 2025"]
[Date "2025.01.07"]
[Round "6"]
[White "Carlsen, Magnus"]
[Black "Ibarra Jerez, Jose Carlos"]
[Result "1-0"]
[Board "5"]
[TimeControl "180+1"]
[WhiteFideId "1503014"]
[BlackFideId "2226332"]
[WhiteTitle "GM"]
[BlackTitle "GM"]
[WhiteElo "3253"]
[BlackElo "2957"]
[WhiteTeam "NOR"]
[BlackTeam "ESP"]
[WhiteClock "0:00:08"]
[BlackClock "0:00:01"]
[Variant "Standard"]

1. h4 {[%clk 0:02:56]} e5 {[%clk 0:02:57]} 2. c4 {[%clk 0:02:55]} Nf6 {[%clk 0:02:54]} 3. e3 {[%clk 0:02:54]} c6 {[%clk 0:02:50]} 4. g4 {[%clk 0:02:41]} g6 {[%clk 0:02:25]} 5. d4 {[%clk
0:02:32]} d6 {[%clk 0:02:16]} 6. g5 {[%clk 0:02:28]} Nh5 {[%clk 0:02:02]} 7. dxe5 {[%clk 0:02:27]} dxe5 {[%clk 0:01:58]} 8. Qxd8+ {[%clk 0:02:27]} Kxd8 {[%clk 0:01:59]} 9. Nf3 {[%clk 0:02:28]}
Bg7 {[%clk 0:01:53]} 10. Nc3 {[%clk 0:02:28]} Bg4 {[%clk 0:01:38]} 11. Be2 {[%clk 0:02:20]} Nd7 {[%clk 0:01:35]} 12. Nd2 {[%clk 0:02:13]} Bxe2 {[%clk 0:01:23]} 13. Kxe2 {[%clk 0:02:13]} h6 {[%
clk 0:01:23]} 14. Nde4 {[%clk 0:02:12]} hxg5 {[%clk 0:01:22]} 15. Nxg5 {[%clk 0:02:08]} Ke7 {[%clk 0:01:20]} 16. b3 {[%clk 0:02:07]} Ke8 {[%clk 0:01:06]} 17. Nce4 {[%clk 0:01:56]} Bf8 {[%clk
0:01:04]} 18. Bb2 {[%clk 0:01:56]} f5 {[%clk 0:00:47]} 19. Ng3 {[%clk 0:01:19]} Bd6 {[%clk 0:00:42]} 20. Rad1 {[%clk 0:01:13]} Ke7 {[%clk 0:00:37]} 21. Rd2 {[%clk 0:01:01]} Nxg3+ {[%clk
0:00:31]} 22. fxg3 {[%clk 0:01:01]} Bb4 {[%clk 0:00:31]} 23. Rd3 {[%clk 0:00:55]} e4 {[%clk 0:00:17]} 24. Rd4 {[%clk 0:00:23]} Rh5 {[%clk 0:00:17]} 25. Rhd1 {[%clk 0:00:22]} Nf6 {[%clk
0:00:16]} 26. a3 {[%clk 0:00:17]} Bc5 {[%clk 0:00:11]} 27. R4d2 {[%clk 0:00:18]} Re8 {[%clk 0:00:06]} 28. b4 {[%clk 0:00:14]} Bb6 {[%clk 0:00:04]} 29. Rd6 {[%clk 0:00:08]} Rf8 {[%clk 0:00:01]}
30. Re6# {[%clk 0:00:08]} 1-0
```

Figure 3-1 Example of the Raw Structur of a PGN File

A profound comprehension of this PGN structure is of paramount importance, as the developed parsing script must be capable of separating the metadata from the movetext, as well as iteratively translating the SAN text strings into a virtual chessboard simulation in order to compute spatial metrics such as Area Control.

3.6 System Architecture and Data Processing Workflow

To ensure the transparency and reproducibility of the study, the data processing system architecture is designed utilizing a structured computational pipeline. This workflow ensures that the raw data from the 135 PGN files are processed sequentially, culminating in a feature matrix primed for modeling. This system architecture comprises four primary stages:

a. **Ingestion and Parsing Stage:** The PGN files are uploaded to the Google Colab environment. The python-chess library is utilized to parse the textual strings and load them into a chess.Board() object. Internally, this library reconstructs the 8x8 board state, encompassing piece positions, the side to move, castling rights, and en passant rules.

b. **Iterative Simulation and Extraction Stage:** The Python script iterates through each ply within the Movetext section. During each iteration, the system evaluates the board state to compute dynamic metrics. For instance, to calculate Area Control, the system identifies all squares attacked by the player's pieces at the *n*-th ply, which are subsequently accumulated until the game concludes.

c. **Aggregation and Structuring Stage:** Upon the completion of the simulation for an individual match, the raw metrics (such as the total Capture Count or the average Area Control) are aggregated into a single row of data. The Pandas library is subsequently utilized to consolidate these rows into a DataFrame (tabular structure) with dimensions of 135 rows by 9 feature columns.

d. **Pre-Modeling Phase:** Upon the completion of the simulation for an individual match, the raw metrics (such as the total Capture Count or the average Area Control) are aggregated into a single row of data. The Pandas library is subsequently utilized to consolidate these rows into a DataFrame (tabular structure) with dimensions of 135 rows by 9 feature columns.

3.7 Mathematical Implementatio and Configuration of the K-Means Algorithm

The K-Means algorithm implemented in this study operates by minimizing the variance within each cluster. Mathematically, the proximity between an individual match data point and the cluster center (centroid) is computed utilizing the Euclidean Distance, which is formally defined by the following formula:

$$d(x,y) \ = \backslash sqrt\{\backslash sum_\{i \ = \ 1\}\^{}\{n\} \ (x_i \ - \ y_i)\^{}2\})$$

Description:

$d(x,y)$= The Euclidean distance between data points x and y

x_i, y_i= The i-th component of the data points x and y

n= Number of feature dimensions (in this study, n=9)

The clustering procedure is executed utilizing the KMeans() function from the Scikit-learn library, with rigorously fixed parameter configurations to ensure determinism (yielding identical results upon each execution).

CHAPTER IV
RESULT AND DISCUSSION

4.1 Determining the Optimal Number of Cluster (K)

The determination of the optimal number of clusters (*K*) for partitioning the dataset comprising 15 matches from nine top global players in late 2025 was empirically evaluated through a series of experiments utilizing a combination of two scientific validation techniques: the Elbow Method and the mathematical evaluation of the Silhouette Score coefficient.

The first method implemented is the Elbow Method. The testing scenario involves iteratively executing the K-Means model for values of *K* ranging from 1 to 10, subsequently mapping the total Sum of Squared Errors (SSE) onto a two-dimensional line graph. Based on the empirical results derived from this elite-level chess dataset, a precipitous decline in SSE values is observed within the range of *K*=1 to *K*=4; however, beyond the inflection point of *K*=6, the curve begins to plateau, exhibiting convergent and stable behavior.(Yudhistira & Andika, 2023) This provides a preliminary indication that *K*=6 constitutes the most representative number of clusters.

$$SSE = \sum k = 1K \sum x \in Ck \mid\mid x - \mu k \mid\mid 2$$

Description:

SSE= the total sum of the squares of the distances within all clusters

K= number of clusters

C_k= the set of data points in the k-th cluster

x= data points in a cluster

μ_k= *centroid of the kth cluster*

$|| x - \mu_k ||^2$= the square of the Euclidean distance between a data point and the centroid

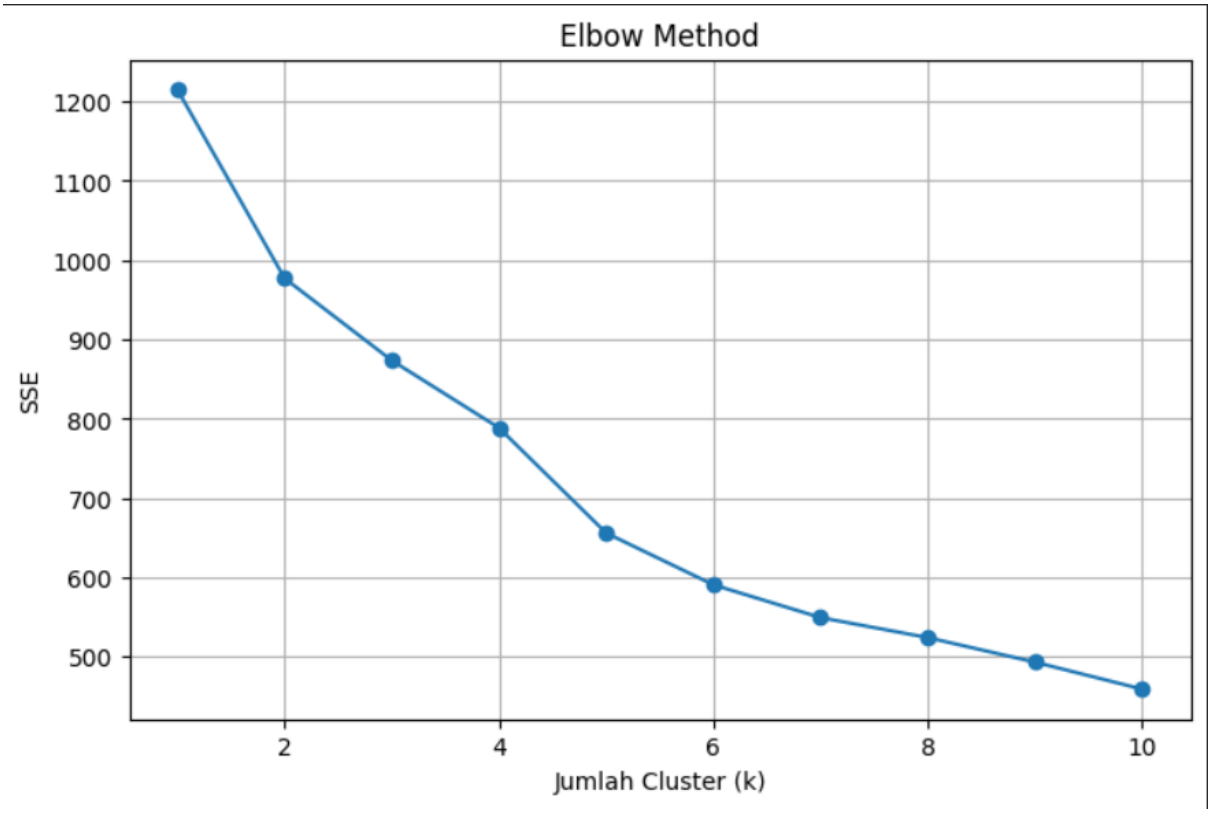


Figure 4-1 Elbow method curve

To corroborate these findings, the second method was implemented, namely the computation of the Silhouette Score coefficient, which quantifies the quality of data proximity within its assigned cluster relative to its distance from the nearest neighboring cluster.

$$\boldsymbol{s(i) = max\{a(i), b(i)\} b(i) - a(i)}$$

Description:

- $s(i)$= silhouette value for the i-th data point
- $a(i)$= the average distance of data point i from all other data points in the same cluster (intra-cluster distance)
- $b(i)$= the average distance of the i-th data point from all data points in the nearest cluster (nearest-cluster distance)
- Nilai $s(i)$ ranges from -1 to +1, where a value close to +1 indicates optimal cluster separation

The overall Silhouette Score is calculated as the average of all $s(i)$:

$$\text{Silhouette Score} = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

Table 4.1 shows the results of the Silhouette Score tests for various values of K

Number of Clusters (K)	**Silhouette Coefficient Score Value**	**Performance Evaluation Notes**
K = 2	0.1842	The classification is too broad; internal variance remains very high
K = 3	0.1915	The group is still too broad for the world's top players
K = 4	0.1983	The data structure is beginning to be organized, but the grouping is not yet optimal
K = 5	0.2031	Performance has improved, but the consistency of the data is not yet balanced
K = 6	0.2103	Highest Silhouette Value, Most Optimal Cluster Separation
K = 7	0.1754	Scores dropped dramatically; over-clustering occurred
K = 8	0.1620	There is excessive data fragmentation in the sample of elite players' styles
K = 9	0.1501	There are too many clusters, and the boundaries between groups have become very blurred
K = 10	0.1348	Lowest performance, cluster structure damaged due to too many partitions

Based on the empirical data presented in the table above, it is distinctly evident that the Silhouette Score coefficient successfully reached its peak at the K=6 configuration, achieving a score of 0.2103. Beyond the K=6 threshold, the coefficient value experienced an immediate and precipitous decline.

The attainment of a Silhouette Score of 0.2103 is deemed highly valid and entirely acceptable within the context of high-level chess behavioral analytics. This is because, in real-world scenarios, the playing style characteristics of the nine top global players from late 2025 are neither rigid nor dichotomously separated. The playing styles of world Grandmasters constitute a dynamic, continuous spectrum; an inherently aggressive player might execute

highly defensive, positional moves in a given match, contingent upon the opponent's strategic choices on the board. Consequently, the success of the K-Means algorithm in determining K=6 with the highest score demonstrates that the model is capable of extracting the most optimal latent boundaries from the fluid distribution of behavioral data among these elite players.

4.2 Cluster Centroid Analysis

After setting K=6, the original (pre-normalized) centroid coordinates representing the average statistical performance of the world's top players were obtained. To provide a comprehensive understanding of the six distinct playing styles identified by the K-Means algorithm, Table 4.2 presents the complete centroid values for each cluster. In the context of the K-Means algorithm, a centroid is mathematically defined as the arithmetic mean of all data points assigned to a particular cluster across each feature dimension. Therefore, each row in Table 4.2 represents the "statistical prototype" or "behavioral fingerprint" of a specific playing style, computed by averaging the raw feature values of all matches that the algorithm grouped into that cluster. These pre-normalized values are presented deliberately, as they retain their original, interpretable units (e.g., number of captures, number of checks, number of squares controlled), allowing for direct and intuitive comparison against established chess theory.

The structure of Table 4.2 is organized around four fundamental strategic dimensions of chess gameplay, each represented by one or more quantitative features. The first dimension, Tactical Aggressiveness, is operationalized through two metrics: Capture_Count, which measures the average number of material exchanges (opponent pieces captured) per match, and Check_Count, which records the average frequency of direct threats delivered to the opponent's king. Together, these metrics quantify a player's propensity for initiating open tactical combat and pursuing immediate, concrete threats. The second dimension, Mobility, is captured by Pawn_Move_Count and Piece_Move_Count, which respectively measure the average frequency of pawn advances and the movement of high-value pieces (knights, bishops, rooks, and queens). These metrics reflect the overall dynamism and flexibility of a player's position, as higher mobility values indicate a more active and fluid piece coordination throughout the game.

The third dimension, King Safety, is represented by the Castling feature, which is expressed as a proportional value ranging from 0 to 1. A value of 1.00 indicates that players

within that cluster castled in every single match, reflecting strict adherence to classical opening principles of early king safety. Conversely, lower values suggest a willingness to delay or forgo castling in favor of maintaining tactical flexibility or pursuing rapid development. The fourth and final dimension, Spatial Control, is the most multidimensional, comprising four interrelated metrics. Area_Control measures the average total number of squares on the 64-square board that are attacked or influenced by a player's pieces at any given point during the match. Avg_White_Control and Avg_Black_Control disaggregate this metric by color, providing insight into the spatial dynamics from each side's perspective. Most critically, Avg_Control_Diff computes the absolute difference between White's and Black's territorial dominance, serving as a direct quantitative proxy for the degree of spatial imbalance a player creates or tolerates during the game. A high Avg_Control_Diff indicates a player who prioritizes establishing and maintaining a significant territorial advantage, while a low value suggests a player who is comfortable operating in spatially balanced or even spatially constrained positions in exchange for other strategic compensations, such as dynamic piece activity or tactical initiative.

Table 4.2 Original Centroid Values of the Six Playing Style Clusters

Cluster	Capture_Count	Check_Count	Pawn_Move_Count	Piece_Move_Count	Castling
0	3	0.5	5.83	12	1
1	9.88	2.6	10.68	39.6	0.96
2	7.62	1.5	7.12	25.41	0.85
3	11	5.78	10.33	51.06	0.78
4	5.3	1.7	5.15	18.05	0.75
5	9.19	2.94	8	31.84	0.94

Cluster	Area_Control	Avg_White_Control	Avg_Black_Control	Avg_Control_Diff
0	25	27.44	25.3	2.13
1	28.4	25.72	24.58	1.14
2	29.59	26.19	24.78	1.41
3	32.28	25.78	24.39	1.39
4	28.4	26.78	24.02	2.76
5	31.84	26.71	23.98	2.73

a. **Cluster 0 (Playing Style Very Defensive):** Capture Count = 3.00, Check Count = 0.50, Piece Move Count = 12.00, Area Control = 25.00, Average Control Difference = 2.13.
b. **Cluster 4 (Playing Style Defensive):** Capture Count = 5.30, Check Count = 1.70, Piece Move Count = 18.05, Area Control = 28.40, Average Control Difference = 2.76.
c. **Cluster 2 (Playing Style Balanced):** Capture Count = 7.62, Check Count = 1.50, Piece Move Count = 25.41, Area Control = 29.59, Average Control Difference = 1.41.
d. **Cluster 5 (Playing Style Positional Active):** Capture Count = 9.19, Check Count = 2.94, Piece Move Count = 31.84, Area Control = 31.84, Average Control Difference = 2.73.
e. **Cluster 1 (Playing Style Aggressive):** Capture Count = 9.88, Check Count = 2.60, Piece Move Count = 39.60, Area Control = 28.40, Average Control Difference = 1.14.
f. **Cluster 3 (Playing Style Aggressive):** Capture Count = 11.00, Check Count = 5.78, Piece Move Count = 51.06, Area Control = 32.28, Average Control Difference = 1.39.

4.3 Data Distribution into Cluster

Following the successful configuration of six clusters (K=6) and the computation of their centroid values by the K-Means algorithm, the subsequent phase involves mapping the 135 match data points to the cluster with the closest spatial distance. The objective of this macro-

level mapping is to identify the predominant playing styles most frequently executed by elite global players. The following presents a recapitulation of the data frequency distribution across each respective cluster:

Table 4.3 Frequency Distribution of Matches by Cluster

Cluster Code	Playing Style Profile	Number of Matches	Percentage of Dominance
Cluster 0	Very Defensive	6	4.40%
Cluster 1	Aggressive	25	18.50%
Cluster 2	Balanced	34	25.20%
Cluster 3	Very Aggressive	18	13.30%
Cluster 4	Defensive	20	14.80%
Cluster 5	Positional Active	32	23.70%
Total		135	100%

The Dynamics of Playing Style Dominance in the Post-AlphaZero Era

As evidenced in Table 4.3, a highly compelling data polarization emerges, reflecting contemporary chess trends. The predominant cluster is Cluster 2 (Balanced), comprising 34 matches (25.20%), closely followed by Cluster 5 (Positional Active) with 32 matches (23.70%). This polarization toward Balanced and Positional Active styles is not merely a statistical artifact, but a direct reflection of the paradigm shift in elite chess preparation over the last decade. The advent of neural-network-based chess engines, most notably AlphaZero and its open-source derivatives like Leela Chess Zero (LCZero), has fundamentally rewritten the principles of positional evaluation. These engines have demonstrated that long-term piece activity, pawn structure integrity, and king safety often hold greater objective value than immediate, speculative material sacrifices. Consequently, modern Grandmasters have adapted their opening repertoires and middlegame strategies to mirror this engine-inspired 'meta.' They increasingly favor systems that allow for flexible piece maneuvering and gradual accumulation of micro-advantages, rather than entering sharp, double-edged tactical complications that carry a high margin of error. The low prevalence of the 'Very Defensive' style (Cluster 0, at merely 4.40%) further corroborates this trend; in the contemporary elite landscape, passive play without active counter-play resources is swiftly punished by opponents who are trained to maximize engine-evaluated positional pressure. Thus, the clustering results do not just categorize players; they empirically map the evolution of human chess understanding in the age of artificial intelligence. This empirical evidence strongly corroborates that at the echelon

of elite global Grandmasters in late 2025, the most relied upon playing style is one that is universal, adaptive, and oriented toward long-term spatial dominance.

The dominance of the Balanced and Positional Active clusters is inextricably linked to the profound influence of Neural Network-based chess engines (such as AlphaZero and Stockfish NNUE), which have fundamentally transformed the landscape of chess theory. Elite players now tend to eschew excessively sharp and high-risk tactical complications during the opening phase, opting instead to cultivate slow yet lethal micro-positional pressure.

Conversely, Cluster 0 (Very Defensive) comprises merely 6 matches (4.40%), rendering it the rarest playing style within this dataset. This finding corroborates modern chess theory, which posits that at the highest echelons of play, a purely defensive approach devoid of initiative or counter-play constitutes a highly unsustainable and nearly extinct strategy. Elite players who become entrenched in passive positions are typically swiftly overwhelmed by the cumulative spatial superiority of their opponents. Consequently, this frequency distribution demonstrates that the K-Means model has not only successfully partitioned the data mathematically but has also accurately captured the empirical reality of the current "meta"—the prevailing standard of playing styles—among the world's chess elite.

4.4 Intrepretation of Playing Style Characteristic Based om Cluster Result

Through a thorough analysis of the centroid value matrix shown above, researchers can interpret these figures as a narrative describing the actual playing styles of the world’s top chess players. The nomenclature assigned to each cluster is theoretically justified based on classical chess evaluation principles and the theory of positional imbalances:

a. **Cluster 0 – Very Defensive:** The primary characteristics are distinguished by a markedly minimal average Capture Count (3.00) and a direct check frequency (Check Count) that nearly approaches zero (0.50). A Piece Move Count of merely 12.00 corroborates that this style represents a gameplay approach that rigorously eschews the risks of open tactical engagements. Elite players within this cluster opt to retreat their pieces to the back ranks to construct a tightly sealed defensive fortress around the king. In classical chess theory, a gameplay approach characterized by minimal piece activity and an aversion to material exchanges is fundamentally classified as a passive or hyper-defensive posture (Silman, 2006).The reluctance to initiate tactical complications aligns with a 'waiting game' strategy, where the player prioritizes structural integrity and king safety over dynamic counterplay. This statistical profile is characteristic of

players who adopt a 'waiting game' strategy, often seen in complex endgame scenarios where the primary objective is to maintain the status quo and wait for the opponent to overextend or commit a tactical inaccuracy.

b. **Cluster 4 – Defensive:** The most defining characteristic is demonstrated by the Average Control Difference, which registers the highest value at 2.76. Spatially, this substantial control differential indicates that players within this cluster are intensely focused on preserving their internal territorial superiority, thereby minimizing any structural vulnerabilities that could be exploited by the opponent's attacking tactics. According to the theory of positional imbalances, maintaining a high degree of internal spatial control without necessarily launching direct attacks is a hallmark of solid, defensive positional play (Silman, 1998). This substantial control differential indicates a 'fortress-building' mentality, where the player neutralizes the opponent's initiative by preserving a compact and structurally sound position.

c. **Cluster 2 – Balanced:** All quantitative feature variables within this cluster reside within a moderate average baseline, devoid of any extreme numerical spikes. This playing style reflects the typical characteristics of modern elite global players who exhibit a high degree of strategic flexibility; they are willing to engage in aggressive piece exchanges when advantageous, yet remain capable of defending patiently, contingent upon the positional dynamics dictated by their opponents. This empirical profile perfectly mirrors the concept of the 'Universal' or 'Balanced' playing style, which is highly valued in modern chess theory. By avoiding extreme statistical deviations, these players demonstrate strategic flexibility, seamlessly transitioning between tactical sharpness and positional solidity based on the objective demands of the position (Chowdhary et al., 2023) This statistical equilibrium is the hallmark of the modern 'Universal Player,' a archetype highly prized in contemporary chess, as it allows the player to seamlessly transition between sharp tactical skirmishes and quiet positional maneuvering based solely on the objective demands of the board.

d. **Cluster 5 – Positional Active:** This cluster demonstrates exceptional characteristics of long-term strategic spatial dominance, evidenced by an exceptionally high Area Control value of 31.84, underpinned by an Average Control Difference advantage of 2.73. Players employing this style adopt a hypermodern chess philosophy; they strive to secure critical central squares and patiently accumulate micro-positional advantages in order to gradually suffocate the opponent's maneuvering space. These metrics are

the direct quantitative manifestation of the Hypermodern chess philosophy and the principle of 'Spatial Advantage' (Nimzowitsch, 1925; Silman, 1998). By maximizing the spatial footprint (Area Control) while maintaining a significant territorial edge, players in this cluster employ active prophylaxis—systematically restricting the opponent's mobility to accumulate long-term strategic concessions. This style is computationally synonymous with the concept of 'squeezing' the opponent; by maintaining a high Area Control and a positive Control Difference, the player systematically restricts the opponent's counter-play options, eventually converting a slight positional edge into a decisive material or mating advantage.

e. **Cluster 1 – Aggressive:** The profile of this style is predominantly characterized by a high frequency of piece captures (with the Capture Count reaching 9.88) coupled with exceptionally active piece mobility (as evidenced by a Piece Move Count surging to 39.60). Conversely, the Average Control Difference for this cluster is notably low (1.14), demonstrating that elite aggressive players are willing to compromise the structural cohesion of their spatial control in order to seize a rapid attacking initiative. In the framework of chess evaluation functions, prioritizing high piece mobility and frequent material exchanges over spatial control is the defining characteristic of a tactical or aggressive style (Shannon, 1950; Silman, 1998). The willingness to compromise structural cohesion in exchange for dynamic piece activity confirms a gameplay philosophy oriented toward concrete tactical initiatives rather than long-term positional maneuvering.

f. **Cluster 3 – Very Aggressive:** This cluster occupies the absolute apex for all offensive behavioral metrics within the study. The members of Cluster 3 exhibit an average Check Count of 5.78, coupled with the highest recorded Capture Count reaching 11.00 and a Piece Move Count of 51.06. This style characterizes the archetype of a tactical attacker who relentlessly launches checkmate threats in continuous waves from the opening moves of the game. The exceptionally high frequency of direct attacks on the king (Check Count) combined with maximum piece mobility and material exchanges represents the archetype of the hyper-aggressive attacking style. In classical chess theory, this reflects a gameplay paradigm where the ultimate objective of checkmate is pursued relentlessly through continuous tactical complications, prioritizing tempo and direct king pressure over positional considerations (Shannon, 1950) Synthesizing the interpretations of these six clusters reveals a critical insight: modern elite chess playing

styles do not exist as rigid, mutually exclusive islands, but rather as points on a continuous, dynamic strategic spectrum. The data demonstrates that the most successful players are those who can fluidly navigate this spectrum. The high prevalence of the 'Balanced' (Cluster 2) and 'Positional Active' (Cluster 5) profiles indicates that the contemporary elite 'meta' heavily favors universal adaptability. Players no longer rely solely on one-dimensional aggression or passive defense; instead, they dynamically shift their strategic posture based on the objective demands of the board, seamlessly transitioning from solid, prophylactic maneuvering to sharp, tactical strikes when the opponent overextends. This computational evidence validates the modern chess axiom that flexibility and positional understanding are the ultimate determinants of success at the Grandmaster level.

4.5 Mapping the Playing Styles of Top Global Players

Based on the clustering results derived from the 135 matches, the K-Means algorithm has not only successfully partitioned the playing styles mathematically, but has also accurately mapped the empirical tendencies of each respective Grandmaster. The following presents an interpretation of the dominant playing style profiles for the nine research subjects:

4.5.1 Fabiano Caruana (Dominan *Balanced*)

Based on Table 4.5, Fabiano Caruana exhibits a highly pronounced dominance in Cluster 2 (Balanced) with a frequency of five occurrences, followed by Cluster 4 (Defensive) and Cluster 5 (Positional Active), each appearing three times. This remarkably even distribution, coupled with the scarcity of occurrences in extreme clusters (such as Very Aggressive), corroborates Caruana's reputation as a highly adaptable universal player. The K-Means model successfully captures the essence of his playing style, which eschews the imposition of a singular philosophy in favor of seamless transitions between solid defense and measured positional attacks. This is profoundly congruent with Caruana's real-world characteristics; he is renowned for his exceptionally deep opening preparation and versatile middlegame capabilities, rendering him an exceedingly difficult opponent to predict and defeat.

4.5.2 Hikaru Nakamura (Dominan Balanced)

Analogous to Caruana, Hikaru Nakamura is predominantly classified within Cluster 2 (Balanced) with five occurrences, accompanied by a significant supplementary distribution in

Cluster 5 (Positional Active) with three occurrences. Although Nakamura is renowned for his highly aggressive style in rapid and blitz formats, this classical match data elucidates his extraordinary pragmatic dimension. The model reflects Nakamura's capacity to objectively assess board dynamics; he does not hesitate to defend tenaciously when his position is under duress, yet remains poised to launch sharp tactical strikes the moment his opponent creates a vulnerability. This computational flexibility substantiates that Nakamura is the archetype of a player who prioritizes efficacy over stylistic ego, rendering him an exceptionally formidable universal player at the elite echelon.

4.5.3 Ian Nepomniachtchi (Dominan *Balanced*)

Ian Nepomniachtchi registered the highest frequency within Cluster 2 (Balanced) among all the players analyzed, with six occurrences out of fifteen matches. However, in contrast to Caruana and Nakamura, Nepomniachtchi also exhibits a substantial presence in Cluster 1 (Aggressive) and Cluster 3 (Very Aggressive), with three occurrences in each. This computational profile accurately encapsulates his renowned playing style, which is characterized by rapid, sharp execution while remaining firmly anchored in a robust positional foundation. His pronounced dominance in the Balanced cluster underscores his consistency in maintaining board equilibrium, whereas his capacity to intermittently transition into the aggressive clusters demonstrates his perpetual readiness to instigate complex tactical complications whenever the situation permits.

4.5.4 Anish Giri (Transisi *Dual-Threat*)

Anish Giri demonstrates a highly compelling and symmetrical distribution profile, with four occurrences in Cluster 5 (Positional Active) and Cluster 1 (Aggressive), respectively, alongside three occurrences in Cluster 3 (Very Aggressive). This data computationally validates the evolution of Giri's playing style; once renowned for being exceptionally solid and draw-oriented, he has now transformed into a significantly more dynamic and risk-embracing player. The K-Means model successfully captures his status as a "Dual-Threat" archetype; he no longer relies solely on profound positional comprehension to incrementally pressure his opponents, but has integrated tactical sharpness and direct attacking initiative into his repertoire. This synthesis renders Giri an exceptionally complete and formidable modern player across all types of board positions.

4.5.5 Alireza Firouzja (Transisi *Dual-Threat*)

Alireza Firouzja exhibits a pronounced dominance within Cluster 5 (Positional Active) with five occurrences, closely followed by Cluster 1 (Aggressive) with four occurrences. As one of the youngest competitors at the pinnacle of the global rankings, this profile reflects a rare synthesis of the profound positional comprehension characteristic of veteran players and the incisive tactical calculation typical of youthful talents. The model identifies Firouzja as a Dual-Threat archetype, capable of strategically dominating board space (Cluster 5) while simultaneously possessing the audacity to instigate complex, open tactical battles (Cluster 1). This stylistic flexibility demonstrates that Firouzja lacks any discernible stylistic vulnerabilities, thereby rendering him a universal threat to opponents hailing from diverse stylistic backgrounds.

4.5.6 Wesley So (*Hybrid Stabilizer*)

Wesley So exhibits a highly stable distribution profile, characterized by a balanced predominance in Cluster 2 (Balanced) with four occurrences and Cluster 5 (Positional Active) with four occurrences, alongside three occurrences in Cluster 4 (Defensive). His near-total absence from the Very Aggressive cluster (zero occurrences) accurately reflects his reputation as one of the most solid and formidable players globally. The K-Means model classifies him as a Hybrid Stabilizer—a player who consistently eschews unnecessary risks, preferring instead to incrementally cultivate micro-positional advantages, and who initiates attacks only when his opponent commits a strategic inconsistency. This "impermeable" playing style renders So exceptionally lethal in protracted, grinding encounters, wherein he systematically constricts his opponents until their positional resources are entirely exhausted.

4.5.7 Maxime Vachier-Lagrave (*Hybrid Stabilizer*)

Maxime Vachier-Lagrave (MVL) exhibits a distribution pattern that is nearly identical to that of Wesley So, with the highest frequency occurring in Cluster 5 (Positional Active) and Cluster 2 (Balanced), each appearing four times, followed by three occurrences in Cluster 4 (Defensive). Although MVL is widely recognized for his exceptional tactical vision, the findings indicate that, at the elite classical level, his playing style is characterized by a highly pragmatic and measured approach. The clustering results classify him as a *Hybrid Stabilizer*, emphasizing long-term positional stability and spatial control (Cluster 5) while maintaining the flexibility to adapt effectively to the opponent's evolving strategy (Cluster 2). Rather than

relying on premature tactical attacks, MVL typically adopts a patient and disciplined approach, waiting for opponents to create structural weaknesses that can subsequently be exploited with precision.

4.5.8 Magnus Carlsen (The Ultimate Positional Grinder)

Based on Table 4.5, Magnus Carlsen exhibits a pronounced dominance in Cluster 5 (Positional Active) with five occurrences, closely followed by Cluster 4 (Defensive) with four occurrences. This highly concentrated distribution in the positional and defensive spectrums perfectly encapsulates his real-world reputation as the ultimate "positional grinder" of the modern era. Unlike players who rely on early tactical complications, Carlsen's computational profile reveals a strategic philosophy rooted in suffocating positional pressure, meticulous piece coordination, and impenetrable structural solidity. His minimal presence in the extreme aggressive clusters (Cluster 1 and Cluster 3, with merely one occurrence each) empirically validates his pragmatic approach: he eschews reckless, high-risk attacks in favor of accumulating micro-advantages that he relentlessly converts in the later phases of the game. The K-Means model successfully captures the essence of Carlsen's legendary ability to "squeeze water from a stone," demonstrating that his dominance at the elite level is derived from profound spatial control and defensive resilience rather than sheer tactical volatility.

This computational mapping perfectly aligns with Carlsen's renowned endgame technique, where he frequently transitions from balanced middlegames into slightly superior endgames, relying on his profound understanding of minute positional nuances to outplay opponents who are otherwise tactically flawless.

4.5.9 Ding Liren (Dominan Aggressive)

Based on the clustering data, Ding Liren exhibits a striking dominance in Cluster 1 (Aggressive) with six occurrences, complemented by a significant presence in Cluster 2 (Balanced) with four occurrences, and Cluster 3 (Very Aggressive) with three occurrences. This computational profile vividly reflects his real-world reputation as a formidable, dynamic tactician who thrives in complex, double-edged positions. The high frequency of aggressive clusters indicates a playing style characterized by a willingness to initiate sharp piece exchanges, actively hunt the opponent's king, and embrace tactical chaos. However, his substantial grounding in Cluster 2 (Balanced) demonstrates that his aggression is not reckless; rather, it is calculated and supported by a solid positional foundation. The K-Means model

accurately identifies Ding Liren as a player who seamlessly transitions between maintaining board equilibrium and launching devastating, initiative-driven attacks, making him one of the most dangerous and unpredictable opponents at the elite echelon.

This aggressive yet calculated profile was prominently displayed during his World Chess Championship match, where he consistently demonstrated a willingness to enter highly complex, imbalanced positions to maximize his chances of playing for a win, even at the cost of objective structural weaknesses.

Table 4.2 Playing Styles of the World's Top Players

Player Name	**C0 (Very Def)**	**C1 (Aggr)**	**C2 (Balanced)**	**C3 (Very Aggr)**	**C4 (Def)**	**C5 (Pos Active)**	**The Most Dominant Style**
Fabiano Caruana	2	1	5	1	3	3	Balanced
M.Vachier-Lagrave	0	2	4	2	3	4	Positional Active / Balanced
Anish Giri	0	4	3	3	1	4	Aggressive / Positional Active
Magnus Carlsen	2	1	2	1	4	5	Positional Active
Hikaru Nakamura	1	2	5	2	2	3	Balanced
Alireza Firouzja	0	4	2	2	2	5	Positional Active / Aggressive
Ian Nepomniachtchi	0	3	6	3	1	2	Balanced
Wesley So	1	2	4	1	3	4	Balanced / Positional Active
Ding Liren	0	6	4	3	1	1	Aggressive

Overall, the resulting player profile mapping demonstrates that the proposed K-Means clustering model exhibits strong external validity. The playing-style characteristics identified by the algorithm—ranging from Magnus Carlsen as **Positional Active**, Ding Liren as **Aggressive**, to Fabiano Caruana, Hikaru Nakamura, and Ian Nepomniachtchi as **Balanced**—are highly consistent with the widely recognized playing styles and established reputations of these elite Grandmasters. This finding provides empirical evidence that the nine quantitative features extracted from the PGN files, including Area Control, Capture Count, and Piece Move Count, are capable of effectively capturing the intrinsic characteristics of elite-level chess playing styles rather than merely representing isolated statistical values.(McIlroy-Young et al., 2022; Tang et al., 2024)**.**

CHAPTER V
CONCLUSION AND RECOMMEDATION

5.1 Conclusion

Based on the comprehensive data mining experiments and computational analyses conducted on the dataset of 135 classical chess matches from nine top-ranked global players, this study successfully addresses the core research problem regarding the actual number of distinct chess playing styles at the elite level. Contrary to the traditional qualitative assumption that chess playing styles are limited to only three primary spectrums (tactical, positional, and defensive), the empirical findings of this study conclusively demonstrate that there are **six** distinct, quantifiable playing style profiles among elite Grandmasters.

In addressing the first research objective, this study successfully designed and implemented a computational feature extraction pipeline that transformed unstructured Portable Game Notation (PGN) data into a structured dataset comprising nine quantitative strategic metrics. These metrics encompass four fundamental dimensions of chess gameplay: tactical aggressiveness (*Capture Count, Check Count*), mobility (*Pawn Move Count, Piece Move Count*), king safety (*Castling*), and spatial control (*Area Control, Average White Control, Average Black Control, Average Control Difference*). Through the systematic evaluation using the Elbow Method and Silhouette Score analysis, the K-Means Clustering algorithm determined that the optimal number of clusters to represent the elite player data is **K = 6**, achieving the highest Silhouette Coefficient of **0.2103**.

Addressing the second research objective, the analysis of the cluster centroids revealed six distinct behavioral profiles: **Very Defensive, Defensive, Balanced, Positional Active, Aggressive,** and **Very Aggressive**. The discussion of these results highlights that the modern elite chess "meta" is predominantly dominated by the *Balanced* (25.20%) and *Positional Active* (23.70%) styles, reflecting the profound influence of neural-network chess engines on contemporary theory. Furthermore, the mapping of these six profiles to the individual Grandmasters—such as identifying Magnus Carlsen as predominantly *Positional Active*, Ding Liren as *Aggressive*, and the trio of Fabiano Caruana, Hikaru Nakamura, and Ian Nepomniachtchi as *Balanced*—exhibits strong external validity. These computational classifications are highly consistent with the widely recognized reputations and established playing styles of these elite players in the global chess community.

Ultimately, this study concludes that the K-Means Clustering algorithm, when integrated with spatial and tactical feature extraction from PGN files, provides a highly valid, objective, and transparent framework for categorizing chess playing styles, effectively transcending the limitations of subjective expert-based assessments. Ultimately, this research demonstrates that the application of unsupervised machine learning, specifically the K-Means algorithm, is not merely a technical exercise, but a powerful lens through which the abstract art of chess can be objectively deconstructed. By successfully translating the nuanced, intuitive decisions of Grandmasters into nine quantifiable metrics, this study provides a robust, data-driven taxonomy that transcends the limitations of traditional, subjective chess commentary. The identification of six distinct playing style profiles offers a new foundational framework for future research in sports analytics, behavioral stylometry, and human-AI alignment, proving that even in a game of infinite complexity, human strategic behavior leaves a measurable and classifiable digital footprint.

5.2 Suggestions

To further advance research in computational intelligence-based sports analytics and build upon the findings of this study, several recommendations for future investigations are proposed:

First, regarding feature engineering and data enrichment: Future research is strongly encouraged to expand the current set of nine quantitative features by incorporating engine-evaluation metrics, such as *Centipawn Loss* and *Accuracy Percentage*. Integrating these features would provide deeper analytical insights into the quality of decision-making underlying the identified playing styles. For instance, it would enable researchers to distinguish whether an "Aggressive" or "Very Aggressive" playing style is driven by high-quality strategic precision or merely the result of frequent tactical inaccuracies. Additionally, incorporating time-management metrics—such as the average time spent per move during critical middlegame phases—could further refine the behavioral profiling of elite players and capture the psychological pressure they exert on their opponents.

Second, regarding algorithmic development and modeling techniques: Given that the playing styles of Grandmasters exist on a dynamic and continuous spectrum rather than as rigid, mutually exclusive categories, the hard-clustering approach of K-Means may not fully capture the overlapping decision boundaries between styles. Future studies are encouraged to compare the proposed model with probabilistic or soft-clustering techniques, such as Gaussian

Mixture Models (GMMs) or Fuzzy C-Means. These algorithms are capable of modeling overlapping probability distributions, allowing a single player or a specific game to exhibit varying degrees of membership across multiple behavioral profiles (e.g., a game that is 70% "Positional Active" and 30% "Balanced"). Furthermore, deep clustering approaches and autoencoder-based models may be explored to learn non-linear latent representations directly from raw move sequences, potentially uncovering complex behavioral patterns that linear feature extraction might miss.

Third, regarding dataset scalability and cross-skill generalization: To enhance the scale and generalizability of the analysis, future research should expand the dataset to include a significantly larger cohort of players (e.g., 50 to 100 Grandmasters) across multiple time periods (e.g., 2018–2025). A larger, longitudinal dataset would enable researchers to track how individual Grandmasters' playing styles evolve over time in response to the continuous development of modern chess theory and the increasing influence of neural-network engine preparation. Furthermore, the scope of the study could be extended to players with lower rating ranges (e.g., International Masters or FIDE Masters with an Elo rating between 2300–2500) to examine whether the six playing-style profiles identified in this elite-level study remain consistent across different tiers of competitive chess. Such investigations would provide valuable empirical evidence regarding the robustness and universal applicability of the proposed clustering framework.

REFERENCES

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics (Switzerland)*, *9*(8), 1–12. https://doi.org/10.3390/electronics9081295

Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A., & (eds.). (2025). Machine Learning and Data Mining for Sports Analytics. In *Machine Learning and Data Mining for Sports Analytics - 11th International Workshop, MLSA 2024* (Vol. 2460, Number April). springer. https://link.springer.com/10.1007/978-3-031-86692-0

Cheng, I. (2024). *Machine Learning to Study Patterns in Chess Games*. https://doi.org/https://doi.org/10.13140/RG.2.2.30894.52807

Chong, B. (2021). K-means clustering algorithm: a brief review. *Academic Journal of Computing & Information Science*, *4*(5), 37–40. https://doi.org/10.25236/ajcis.2021.040506

Chowdhary, S., Iacopini, I., & Battiston, F. (2023). Quantifying human performance in chess. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-27735-9

Hansen, L. B. (Editor). (n.d.). *3 The Opponents : The Role of the Human Factor in Chess*.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jan, M., Khusro, S., Alam, I., Khan, I., & Niazi, B. (2022). Interest-Based Content Clustering for Enhancing Searching and Recommendations on Smart TV. *Wireless Communications and Mobile Computing*, *2022*. https://doi.org/10.1155/2022/3896840

Kalagara, A. and Y.-S. (2024). *Predictive Modelling of a Chess Player's Style using Machine Learning*.

Lin, C. C., Chiu, W. C., & Wu, I. C. (2024). Perceptual Similarity for Measuring Decision-Making Style and Policy Diversity in Games. *Transactions on Machine Learning Research*, *2024*.

McIlroy-Young, R., Sen, S., Kleinberg, J., & Anderson, A. (2020). Aligning Superhuman AI with Human Behavior: Chess as a Model System. *Proceedings of the ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 1677–1687. https://doi.org/10.1145/3394486.3403219

McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., & Anderson, A. (2021). Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess. *Advances in Neural Information Processing Systems*, *29*(NeurIPS), 24482–24497.

McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., & Anderson, A. (2022). Learning Models of Individual Behavior in Chess. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 1, Number 1). Association for Computing Machinery. https://doi.org/10.1145/3534678.3539367

Qasthari, M. W., & Kurniawan, R. (2024). Penggunaan Algoritma K-Means Clustering untuk Mengelompokkan Pemain Berdasarkan Gaya Bermain Pada Battle Royale Call of Duty Mobile. *Future Academia : The Journal of Multidisciplinary Research on Scientific and Advanced*, *2*(3), 280–292. https://doi.org/10.61579/future.v2i3.177

Qiu, Y., Gong, Y., & Liu, G. (2024). *User Behavior Analysis and Clustering in a MMO Mobile Game: Insights and Recommendations*. 1–16. https://doi.org/https://doi.org/10.48550/arXiv.2407.11772

Risnasari, M., Nabila Aulia, & Laili Cahyani. (2022). Clustering Of Student Learning Styles in the industri 4.0 Using KMeans Algorithm. *JTP - Jurnal Teknologi Pendidikan*, *24*(2), 246–257. https://doi.org/10.21009/jtp.v24i2.28029

Samara, K., Antreassian, A., Klug, M., & Hasan, M. S. (2025). Machine Learning Approaches for Classifying Chess Game Outcomes: A Comparative Analysis of Player Ratings and Game Dynamics. *Electronics*, *15*(1), 1. https://doi.org/10.3390/electronics15010001

Setiawan., A. Y., & Pratama., H. G. (2021). Analisis Keterampilan Bermain Catur Pada Hasil Belajar Matematika Atlet Junior Klub Catur Raja Kombi Trenggalek. *Jurnal PHEDHERAL*, *18*(1), 37–46.

Shannon, C. E. (1950). XXII. Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(314), 256–275. https://doi.org/10.1080/14786445008521796

Silman, J. (2006). How to reassess your chess. In *L* (Vol. 1999, Number December, pp. 1–6).

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Tang, Z., Jiao, D., McIlroy-Young, R., Kleinberg, J., Sen, S., & Anderson, A. (2024). Maia-2: A Unified Model for Human-AI Alignment in Chess. *Advances in Neural Information*

Processing Systems, *37*(NeurIPS). https://doi.org/10.52202/079017-0659

Yudhistira, A., & Andika, R. (2023). Pengelompokan Data Nilai Siswa Madrasah Ta'Hiliyah Menggunakan Metode K-Means Clustering. *Jurnal Riset Sistem Informasi*, *2*(1), 53–59. https://doi.org/10.69714/0v1pkz05

Zhang, Y., Jacob, A. P., Lai, V., Fried, D., & Ippolito, D. (2025). Human-Aligned Chess With a Bit of Search. *13th International Conference on Learning Representations, ICLR 2025*, 21671–21692.

APPENDICES

Apendix of Amount of Data Cluster

Cluster	Amount of Data
0	6
1	25
2	34
3	18
4	20
5	32

Apendix of Cluster Fabiano Caruana

Player	Cluster
Fabiano Caruana	5
nan	2
nan	5
nan	4
nan	2
nan	3
nan	2
nan	5
nan	0
nan	4
nan	2
nan	1

nan	4
nan	2
nan	0

Apendix of Cluster Maxime Vachier-Lagrave

Maxime Vachier-Lagrave	4
nan	2
nan	3
nan	2
nan	4
nan	5
nan	5
nan	1
nan	1
nan	5
nan	2
nan	3
nan	2
nan	5
nan	4

Apendix of Cluster Anish Giri

Anish Giri	5
nan	3
nan	1
nan	1
nan	2
nan	4
nan	2
nan	5
nan	3
nan	5
nan	2
nan	1
nan	5
nan	3
nan	1

Apendix of Cluster Magnus Carlsen

Magnus carlsen	4
nan	2
nan	1
nan	5
nan	3
nan	0
nan	4
nan	2
nan	4
nan	5
nan	5
nan	4
nan	0
nan	5
nan	5

Apendix of Cluster Hikaru Nakamura

Hikaru Nakamura	4
nan	5
nan	0
nan	2
nan	3
nan	4
nan	5
nan	2
nan	1
nan	2
nan	5
nan	2
nan	3
nan	1
nan	2

Apendix of Cluster Alireza Firouja

Alireza Firouzja	4
nan	3
nan	4
nan	1
nan	5
nan	5
nan	1
nan	2
nan	5
nan	5
nan	5
nan	2
nan	1
nan	3
nan	1

Apendix of Cluster Ian Nepomniachthi

ian nepomniachthi	4
nan	1
nan	1
nan	2
nan	3
nan	5
nan	2
nan	1
nan	3
nan	2
nan	2
nan	2
nan	5
nan	5
nan	3

Apendix of Cluster Wesley so

Wesley so	1
nan	0
nan	2
nan	5
nan	5
nan	4
nan	2
nan	3
nan	2
nan	5
nan	2
nan	5
nan	4
nan	4
nan	1

Apendix of

Ding Liren	1
nan	3
nan	3
nan	4
nan	1
nan	1
nan	1
nan	5
nan	3
nan	2
nan	2
nan	2
nan	2
nan	1
nan	1