Data-Driven Approaches Using Explainable AI for Industrial Applications

A thesis submitted during 2025 to the University of Hyderabad in partial fulfillment of the requirements for the award of the Ph.D. degree in Computer Science

by

MUDAVATH RAVI

Enrollment No. 19MCPC07

School of Computer and Information Sciences

University of Hyderabad

P.O. Central University, Gachibowli

Hyderabad – 500046, Telangana, India

CERTIFICATE

This is to certify that the thesis entitled **Data-Driven Approaches Using Explainable AI for Industrial Applications** submitted by **Mudavath Ravi** bearing **Reg. No. 19MCPC07**, in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy in Computer Science**, is a bona fide work carried out by him under our supervision and guidance.

The thesis is free from plagiarism and has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma. The student has the following publications before submission of the thesis for adjudication and has produced the necessary evidence:

1. **A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI.** *2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India*, pp. 1–6, 2022. (Scopus) DOI: 10.1109/I2CT54291.2022.9824265.

2. **Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations.** *In: Multi-disciplinary Trends in Artificial Intelligence (MIWAI 2024), Lecture Notes in Computer Science, Vol. 15431*, Springer, Singapore, 2025. (Scopus) DOI: 10.1007/978-981-96-0692-4_33.

3. **Leveraging LIME Explainability and Gustafson-Kessel Fuzzy Clustering for Resume Grouping and Text Summarization.** *Communicated to: Knowledge-Based Systems (SCIE Journal), Elsevier*, Revised and Resubmitted, 2025.

4. **Hybrid Thresholding for Enhanced Performance and Interpretability in Fraud Detection: Integrating LIME and SHAP for Trustworthy AI-Based Decision Making.** *CMC–Computers, Materials & Continua*, Article ID: CMC-65304. (SCIE) Accepted for publication, 2025.

5. **Evolution of AI-Driven Decision Making With Decision Support Systems, Expert Systems, Recommender Systems, and XAI.** *IETE Technical Review*, Taylor & Francis, 2025. (SCIE) DOI: 10.1080/02564602.2025.2512086.

Further, the student has passed the following courses towards the fulfillment of the coursework requirement for Ph.D.:

Course Code	Name	Credits	Pass/Fail
CS402	Algorithms	4	Pass
CS800	Research Methods in Computer Science	4	Pass
CS844	Pattern Recognition	3	Pass
CS862	Machine Learning	3	Pass

Atul Negi

Prof. Atul Negi
(Supervisor)
School of Computer and Information Sciences
University of Hyderabad
Hyderabad – 500046, India

Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46 (India)

Atul Negi

Prof. Atul Negi
(Dean)
School of Computer and Information Sciences
University of Hyderabad
Hyderabad – 500046, India

DEAN
SCHOOL OF COMPUTER & INFORMATION SCIENCES
UNIVERSITY OF HYDERABAD
HYDERABAD - 500046, T.S. INDIA

DECLARATION

I, **Mudavath Ravi**, hereby declare that this thesis entitled **Data-Driven Approaches Using Explainable AI for Industrial Applications**, submitted by me under the guidance and supervision of **Prof. Atul Negi**, is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Date: 01/09/2025

M. Ravi 01/09/25
Signature of the Student

Name: Mudavath Ravi

Reg. No.: 19MCPC07

1/9/25

Counter signed by Supervisor

Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46. (India)

SDG Goals

Thesis Title: Data-Driven Approaches Using Explainable AI for Industrial Applications

Student Regd. No.: 19MCPC07

Supervisor: Prof. Atul Negi

School: School of Computer and Information Sciences

Among 17 goals (https://sdgs.un.org/goals), under the following SDGs, the work incorporated in the thesis will be addressed:

SDG 4 – Quality Education

In today's digital ecosystem, learners and researchers often struggle to access reliable, relevant, and high-quality academic resources amidst an overwhelming influx of online information. This thesis addresses this challenge by leveraging Explainable AI (XAI) to enhance systems for academic content recommendation and resume screening. By embedding interpretability into AI-driven platforms, the proposed methods ensure that users receive accurate and context-aware guidance, supporting informed educational and career choices. This contributes toward equitable access to quality educational tools, fulfilling the vision of SDG 4.

SDG 9 – Industry, Innovation, and Infrastructure

As industries increasingly rely on AI to streamline operations, detect anomalies, and automate decisions, the demand for transparent and trustworthy systems becomes critical. This research supports SDG 9 by developing explainable AI models for applications such as credit card fraud detection and automated recruitment. By focusing on interpretability, the proposed solutions enable domain experts to understand and trust model decisions. This fosters innovation, promotes responsible AI deployment, and strengthens digital infrastructure across diverse sectors including finance, human resources, and e-governance.

SDG 16 – Peace, Justice, and Strong Institutions

M Rao
01/09/25

Atul Negi
Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46, (India)

The integration of AI into decision-making processes demands mechanisms to ensure fairness, accountability, and transparency. This thesis contributes to SDG 16 by proposing interpretable AI frameworks capable of justifying predictions in critical domains such as fraud detection and candidate evaluation. Through the use of XAI techniques like LIME and SHAP, the research promotes bias mitigation, enhances institutional accountability, and supports the creation of ethical, auditable AI systems. These efforts help build trust in AI-driven governance and reinforce strong, transparent institutions.

M. Ravi
Ravi Mudavath
01/09/2025

Counter signed by Supervisor

Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46 (India)

Acknowledgements

I owe my journey to five individuals who initially pushed me forward and continuously motivated me until I reached my goal: my father, **M. Thaarya**; my advisor, **Prof. Atul Negi**; **Prof. Sameen Fatima**; my school teacher, **C. Ramurthy**; and my wife, **N. Jyothsna**.

I doubt I would have found the motivation to pursue a PhD without their unwavering support. Despite my frequent disagreements with my father and my resistance to his guidance, I am deeply grateful for his persistence. My mother, **M. Ramulamma**, has always had a unique perspective—believing that staying single is perfectly fine as long as one earns a PhD. I take immense pride in having such supportive parents and hope they are pleased that I have finally reached this milestone. I also extend my gratitude to **Prof. Atul Negi**, who introduced me to the academic world and whose guidance and encouragement laid the foundation for this journey.

I am sincerely grateful to my Doctoral Research Committee (DRC) members, **Prof. Durga Bhavani**, **Dr. Wilson Naik**, and **Dr. Arun Kumar Das** for their invaluable guidance and insightful contributions to this thesis. I especially thank **Prof. Durga Bhavani** for consistently encouraging me to stay focused on my research and for her candid reminders at times stern that my Ph.D. should remain my top priority. I am also deeply indebted to my supervisor, **Prof. Atul Negi**, whose passionate engagement with research—even to the extent of dreaming about dominating sets—guided me toward the field of mathematical machine learning morphology. I sincerely appreciate his mentorship and the intellectual depth he has brought to my research journey.

Throughout my PhD journey, I faced numerous challenges, both professionally and personally. I am deeply grateful to my colleagues who stood by me, offering unwavering support and encouragement in countless ways. I would also like to

extend my heartfelt thanks to my **Monachary, G. Maduri, KH. Salman, Aruna, Rupa , Maduri Lata, Adarsh, Pranitha, Rajesh, Rakesh, and Teja** — for the countless moments of laughter and camaraderie.

A special note of appreciation goes to Monachary and Teja. Monachary was always there as a sounding board for my ideas and never hesitated to bring me lunch whenever I stayed on campus to complete my work. Danish, on the other hand, willingly joined me in endless paper-reading sessions, helping me navigate the complexities of approximation algorithms—an endeavor that would have been incredibly difficult alone. Their generosity and support made a significant difference in my journey. I would like to express my deepest gratitude to my dear friend B. Rajesh for his unwavering financial support and the immense confidence he instilled in me throughout this journey. His belief in my potential played a vital role in the successful completion of my thesis and in shaping this important chapter of my life.

My sincere thanks also go to Thulasiram and Prasanth, whose valuable suggestions and constant motivation encouraged me to pursue my Ph.D. Their guidance and encouragement at the right moment gave me the courage to take this step forward. I am truly blessed to have such friends by my side.

I thank **Prof. Atul Negi** for the numerous discussions we had on approximation algorithms. He listened to my ideas on clustering and, in his unique style, helped me understand the key concepts of clustering. I am also grateful to **Prof. Sanjay Chitnis**, who, despite being at REVA University, Bangalore, provided invaluable guidance through email. His ideas and suggestions were crucial in shaping the conference paper I wrote at the beginning of my PhD journey. In his characteristic gentle manner, he pointed out my lack of knowledge in recommender systems and introduced me to the concept of baggage problems in airports. During the semester I spent incorporating his ideas, reading papers on recommender systems and web scraping, my understanding of algorithms expanded tremendously. I am deeply grateful for his time, guidance, and patience.

I thank my brother and sister, **M. Raju** and **Pooja**, as well as my brothers-in-law, **R. Ravi** and **Ragavendra**, and my dear daughter, **Deepanwitha**, for their unwavering support during my illness. I am also deeply grateful to my wife's family—my

father-in-law **N. Badru**, mother-in-law **N. Rukunamma**, and brothers-in-law **N. Ragavendra** and **N. Jithendar** whose constant encouragement and readiness to help have been a source of strength throughout my life.

Ravi Mudavath

Abstract

The rapid advancement of Artificial Intelligence (AI) has led to its widespread adoption across various industrial domains, creating a pressing need for systems that are not only accurate but also interpretable and trustworthy. This thesis, titled "Data-Driven Approaches Using Explainable AI for Industrial Applications," presents an investigation into the integration of Explainable AI (XAI) techniques across diverse AI-driven decision-making paradigms within industrial contexts.

Beginning with a comparative analysis of Expert Systems (ES), Recommender Systems (RS), and modern XAI frameworks, the work traces the architectural evolution from rule-based logic to data-driven black-box models, emphasizing the critical importance of explainability in such transitions.

In the domain of recruitment automation, the thesis introduces an interpretable resume categorization framework that integrates the K-Nearest Neighbors (KNN) algorithm with Local Interpretable Model-agnostic Explanations (LIME), illustrating how transparent predictions can foster fairness and trust in hiring systems. Extending into the financial security sector, a hybrid thresholding methodology for fraud detection is proposed, combining LIME and SHAP explanations. This approach achieves a balance between model performance and interpretability through informed threshold optimization using ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) and PR-AUC (Precision–Recall – Area Under the Curve) metrics.

The broader evolution of AI-based decision-making is further explored through a comparative review of Decision Support Systems (DSS), ES, RS, and XAI, offering a roadmap for the design of next-generation intelligent systems centered on explainability. A case study on software defect prediction demonstrates the practical utility of integrating XAI with predictive analytics in software engineering.

Finally, the thesis proposes a novel method that fuses Gustafson-Kessel fuzzy clustering with Sentence-BERT embeddings to semantically group resumes in an unsupervised setting. By applying LIME and SHAP, the interpretability of cluster memberships is enhanced, enabling transparent profiling of applicants. This approach bridges the gap between unsupervised learning and explainability, highlighting impactful use cases in human resource management.

Overall, this work contributes to the growing body of research aimed at making AI systems in industrial settings not only powerful and scalable, but also comprehensible, ethical, and aligned with human decision-making.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

Chapter 1

Introduction

Industrial applications play a critical role in building the next generation of intelligent, transparent, and trustworthy systems within real-world environments. In the era of digital transformation, the integration of Artificial Intelligence (AI), machine learning, and data science has profoundly reshaped various industrial sectors, including finance, manufacturing, logistics, and human resource management. These technologies are increasingly deployed to automate complex tasks, optimize operations, and support data-driven decision-making. However, as AI models become more advanced, they often operate as black boxes, making it difficult to understand the reasoning behind their decisions. This raises a critical question: *Can these AI-driven systems be trusted to make fair, ethical, and interpretable decisions in sensitive industrial environments?* As industries adopt data-driven methods, the challenge is not only to improve performance but also to ensure that AI systems are transparent and accountable. The need for explainable AI (XAI) has become crucial to address concerns regarding fairness, bias, and accountability, especially in high-stakes areas like recruitment, fraud detection, and predictive maintenance. In such contexts, it is vital that these systems provide clear, understandable reasons for their decisions, fostering trust and ensuring ethical application.

Modern industrial systems increasingly leverage data-driven technologies to optimize efficiency, reliability, and automation across operations. These technologies, spanning areas such as banking, recruitment, and software engineering, integrate data acquisition, analysis, and decision-making. The convergence of the Internet of Things (IoT), cyber-physical systems (CPS), and AI has significantly transformed industrial operations [27], [64], [244], [297].

While traditional Information Technology (IT) systems emphasize the confidentiality, integrity, and availability (CIA) triad, data-centric industrial systems introduce new challenges.

1. INTRODUCTION

In modern enterprises, data is a strategic asset, and effective data management is key to driving competitive advantage—especially for innovation-driven firms [157], [47]. Research shows that data-driven decision-making correlates with measurable gains in productivity and profitability [27].

Machine Learning (ML) has become central to enabling intelligent automation in industrial applications [196]. With the rise of Industry 4.0, advanced ML methods—such as deep learning, reinforcement learning, and ensemble techniques—are increasingly embedded in operational pipelines. These approaches support adaptive control, real-time optimization, and continuous learning from sensor or system-generated data, enabling the development of autonomous and intelligent systems.

However, these benefits come with a cost. The black-box nature of many AI systems raises significant concerns about transparency, accountability, and trustworthiness—particularly in high-stakes or regulated environments. As a result, the demand for interpretable and explainable AI systems has grown substantially [21].

This need has driven the emergence of Explainable Artificial Intelligence (XAI) [403], [193], which aims to bridge the gap between complex AI models and human interpretability. XAI techniques—such as *Local Interpretable Model-agnostic Explanations (LIME)* [310] and *SHapley Additive exPlanations (SHAP)* [216]—generate human-understandable explanations for algorithmic decisions, enhancing transparency and fostering trust. Surveys by Adadi and Berrada [5] and Arrieta et al. [29] highlight the growing importance of XAI, particularly in domains where decision-making must be auditable and reliable. Recent research [30, 385, 413] further emphasizes the utility of XAI in tasks such as feature selection, model debugging, and fairness auditing.

In industrial settings, explainability is essential. It empowers domain experts, auditors, and stakeholders to understand, validate, and trust the outcomes of AI-based systems. This thesis investigates how data-driven approaches, when combined with explainable AI techniques, can be applied across diverse industrial applications—from recruitment automation and fraud detection to unsupervised clustering of resumes—toward building systems that are not only high-performing but also interpretable, ethical, and transparent.

1.0.1 Scope and Limitations

This thesis is situated within the broader domain of industrial applications of Artificial Intelligence, with a specific focus on data-driven decision-making processes. It deliberately excludes areas related to manufacturing operations and other industrial domains that do not directly involve data-centric decision systems.

The primary scope of this research lies in the integration of Explainable Artificial Intelligence (XAI) techniques to enhance interpretability, transparency, and trust in classical machine learning models applied to industrial tasks. The thesis presents interpretable frameworks leveraging methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), applied to use cases including resume classification, financial fraud detection, and semantic clustering of candidate profiles.

The work emphasizes the use of classical machine learning techniques—such as K-Nearest Neighbors (KNN), fuzzy clustering, and Term Frequency–Inverse Document Frequency (TF-IDF)—prioritizing transparency and human-in-the-loop interaction over the raw performance of complex black-box models. It also explores hybrid thresholding strategies to manage challenges like class imbalance, and provides a comparative analysis of decision-making paradigms, from traditional Expert Systems (ES) and Recommender Systems (RS) to modern explainability-driven approaches.

However, this thesis does not consider deep learning methods, including neural networks and transformer-based architectures. These models, although powerful, are excluded due to their limited interpretability within the scope of this research. By focusing solely on interpretable, classical AI methodologies, this thesis aims to contribute to the development of transparent and accountable AI systems for critical industrial applications.

1.1 Motivational Points and Problem Formulation

Opaque “black-box” models in AI-based decision-making have raised critical concerns about fairness, accountability, and transparency—especially in safety-critical and regulation-sensitive industrial domains. This is particularly true for high-stakes applications such as financial fraud detection and intelligent recruitment, where unjustified or biased decisions can lead to significant ethical, legal, and operational consequences.

Explainable Artificial Intelligence (XAI) has emerged as a response to these challenges by aiming to make AI decisions more interpretable and trustworthy for human stakeholders.

Surveys and reviews [5, 29, 30], as well as regulatory frameworks like the European Union's General Data Protection Regulation (EU GDPR), underscore the growing need for explainability across sectors such as finance, healthcare, recruitment, and automation.

Despite this growing attention, many academic advances in XAI remain disconnected from real-world industrial workflows. This gap between theoretical models and practical deployment motivates this thesis, which aims to bridge explainability research with operationally relevant, data-driven industrial applications. The following key objectives guide the research:

- Conducting a comparative analysis of traditional intelligent systems—such as Decision Support Systems, Expert Systems, and Recommender Systems—with modern XAI frameworks, focusing on their transparency and usability in industrial contexts [306];
- Developing interpretable AI models for real-world applications such as financial fraud detection and AI-driven hiring, where fairness, trust, and compliance are of paramount importance [301, 302];
- Introducing explainability into unsupervised learning pipelines by integrating semantic embeddings (e.g., Sentence-BERT), fuzzy logic-based clustering, and local explanation tools like LIME and SHAP for applications in resume profiling and candidate grouping [304];
- Aligning technical model performance with human-centric values—by connecting quantitative evaluation metrics to ethical, legal, and accountability considerations [385, 413].

This thesis aims to design *trustworthy and interpretable AI systems* that meet both operational demands and ethical expectations in industrial environments. By leveraging tools such as LIME [310], SHAP [216], semantic embeddings, and fuzzy clustering, it demonstrates how explainability can be effectively embedded into a variety of industrial AI solutions.

Targeted studies in fraud analytics, intelligent hiring, software defect prediction, and semantic clustering of resumes, are used in this work to investigate whether AI systems can be designed to be both high-performing and explainable. The central hypothesis driving this research is that *explainability is not a trade-off, but a prerequisite for building secure, fair, and human-aligned AI systems in industrial applications*.

1.1.1 Problem Statement

The domain of industrial applications presents complex challenges for deploying trustworthy AI solutions. Classification accuracy is often hindered by data imbalance [165], high dimensionality, and the unstructured nature of real-world data. Given its data-driven nature, the domain exhibits the following key characteristics:

- Extreme class imbalance in financial fraud detection scenarios;
- High-dimensional, unstructured textual data in intelligent recruitment systems;
- A critical need for domain-specific explanations to ensure human trust and regulatory compliance;
- An absence of unified, human-centered evaluation frameworks to assess the effectiveness of explainable AI solutions [385, 413].

The central research question guiding this thesis is:

> *How can Explainable Artificial Intelligence (XAI) be effectively integrated into industrial applications to support interpretable, transparent, and reliable decision-making, without sacrificing predictive performance or domain relevance?*

To address this question, the thesis develops tailored XAI pipelines that combine both global and local interpretability methods (e.g., SHAP and LIME), unsupervised semantic clustering using fuzzy logic and contextual embeddings (e.g., Sentence-BERT), and application-specific workflows for fraud analytics and AI-assisted hiring. The work also revisits traditional paradigms—expert systems, recommender systems, and decision support systems—by reinterpreting them through the lens of modern XAI [303, 306].

Ultimately, this research underscores the position that **explainability is not merely an added feature, but a foundational requirement** for responsible, human-aligned, and operationally effective AI systems in industrial environments.

1.1.1.1 Abstracted Definition

To mathematically formalize the problem, we define an industrial dataset $\mathcal{D}$ as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{1.1}$$

where N denotes the total number of instances, $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector (e.g., representing a credit card transaction, a resume embedding, or sensor data), and $y_i \in \mathcal{Y}$ is the corresponding target label, with $\mathcal{Y}$ being a discrete or continuous output space depending on the task (classification, regression, or clustering).

The primary objective is to learn a predictive function:

$$f : \mathbb{R}^d \to \mathcal{Y} \tag{1.2}$$

optimized via a suitable loss function $\mathcal{L}(f(x_i), y_i)$ to maximize performance metrics such as accuracy, precision, recall, F1-score, or AUC-ROC.

However, in high-stakes industrial applications, accuracy alone is insufficient. It is critical that model predictions are accompanied by explanations that are transparent, trustworthy, and actionable. To support this, we define an explanation function:

$$E : \mathbb{R}^d \times \mathcal{Y} \to \mathcal{M} \tag{1.3}$$

where $\mathcal{M}$ represents the space of interpretable explanations—such as feature attributions, decision rules, example-based rationales, or textual narratives—offering insight into why $f(x)$ was predicted for a given x.

The research focus is on the joint optimization of predictive accuracy and explanation quality. This is formally described as:

- The predictive model $f : \mathcal{X} \to \mathcal{Y}$ is trained to maximize predictive performance:

$$\mathcal{P}(f) = \frac{1}{N} \sum_{i=1}^{N} \delta\big(f(x_i), y_i\big) \tag{1.4}$$

  where

$$\delta(a,b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{otherwise} \end{cases} \tag{1.5}$$

  is an indicator function that checks prediction correctness.

- The explanation function E aims to maximize explanation quality $\mathcal{Q}(E)$, defined as:

$$\mathcal{Q}(E) = \underbrace{\text{Interpretability}(m)}_{\substack{\text{simplicity, sparsity, rule length,} \\ \text{user comprehension}}} + \underbrace{\text{Fidelity}(f, E)}_{\substack{\text{how well } E(x, f(x)) \\ \text{approximates the behavior of } f}} \tag{1.6}$$

  where $m = E(x, f(x)) \in \mathcal{M}$ is the explanation for prediction $f(x)$.

In the industrial sector, Equations 1.3, 1.4, 1.5, and 1.6 are highly complex to prove theoretically. Therefore, in my thesis, I adopt a practical, application-oriented approach instead of a purely theoretical one.

These formulations follow the foundational principles proposed by Ribeiro et al. [310], emphasizing the trade-off between fidelity and interpretability in surrogate models. They also align with Lundberg and Lee [216], who introduce additive feature attributions through SHAP, and reflect broader explainability challenges discussed by Doshi-Velez and Kim [107].

In this context, interpretability refers to the ease with which a human can understand the model's reasoning, while fidelity refers to how accurately the explanation reflects the true behavior of f.

Key Challenges Addressed:

- Tackling data imbalance, noise, and heterogeneity typical in fraud detection and process monitoring, which hinder both prediction and explanation.
- Extending explainability to supervised models (e.g., KNN with LIME) and unsupervised methods (e.g., fuzzy clustering with contextual embeddings like Sentence-BERT).
- Combining local (instance-specific) and global (model-level) explanations to serve diverse stakeholder needs.
- Ensuring computational efficiency and real-time applicability of explanation methods within industrial pipelines.

In this research aims to develop and validate explainable AI frameworks that balance predictive performance $\mathcal{P}(f)$ with explanation quality $\mathcal{Q}(E)$, thus enabling reliable, fair, and human-aligned decision-making in industrial applications.

1.2 Key Contributions and Thesis Outline

This thesis, titled *"Data-Driven Approaches using Explainable AI for Industrial Applications"*, advances the field of Explainable AI (XAI) by exploring its application in industrial settings. The key contributions of this work are as follows:

- **Development of a Transparent and Fair Automated Resume Categorization Framework**: This work introduces a novel automated framework for resume categorization, combining *K-Nearest Neighbors (KNN)* with *LIME* (Local Interpretable Model-agnostic Explanations). The framework not only improves recruitment efficiency but also ensures transparency and fairness in the decision-making process, making it interpretable and adaptable to real-world hiring practices [302] in chapter 3.

- **Explainable Fuzzy Clustering Framework for Unsupervised Text Analysis**: This contribution presents an explainable *fuzzy clustering* framework integrating *Gustafson-Kessel clustering*, *Sentence-BERT embeddings*, and *LIME*. The framework enables interpretable unsupervised analysis of text data, providing valuable insights for industrial applications, particularly in scenarios involving large-scale, unstructured data [304] in chapter 4.

- **Hybrid Thresholding Method for Credit Card Fraud Detection**: A novel hybrid method that integrates *LIME* and *SHAP* (SHapley Additive exPlanations) to enhance credit card fraud detection. This method balances interpretability with predictive accuracy, specifically addressing challenges posed by imbalanced datasets, and demonstrates the value of data-driven explainability in high-stakes decision-making like fraud detection [301] in chapter 5.

- **Evolutionary Study of AI-Driven Decision Support**: This contribution traces the evolution of AI-driven decision support systems, from traditional *Decision Support Systems (DSS)* and Expert Systems to modern Explainable AI, supported by case studies on software defect prediction. This work demonstrates how data-driven methods have progressively integrated explainability to improve decision-making in industrial environments [303]. in chapter 6.

Collectively, these contributions advance the state-of-the-art in Explainable AI by addressing transparency, fairness, and usability challenges, enabling more trustworthy AI-driven decision-making in industrial contexts.

The structure of this thesis is organized as follows:

- **Chapter 1: Introduction**
  This chapter presents the motivation, background, problem statement, research objectives, and significance of the study. It also outlines the scope and contributions of the thesis.

- **Chapter 2: Literature Review**
  A comprehensive review of relevant literature on Decision Support Systems, Expert Systems, Recommender Systems, and Explainable AI techniques. This chapter discusses the evolution of AI-driven decision-making and identifies gaps that the thesis aims to address.

- **Chapter 3: Explainable AI in Automated Resume Categorization**
  This chapter explores an explainable KNN-based framework for resume categorization, integrating LIME to enhance transparency in recruitment decisions. Aligned with Industrial Applications, the system balances accuracy and interpretability, supporting fair and trustworthy automation in hiring processes.

- **Chapter 4: Explainability in Unsupervised Learning: Resume Grouping and Summarization**
  This chapter proposes an interpretable framework for resume grouping using Gustafson-Kessel fuzzy clustering and Sentence-BERT embeddings, enhanced with LIME-based explanations. Aligned with Industrial Applications, the approach supports transparent, data-driven profiling and semantic summarization in talent acquisition systems.

- **Chapter 5: Hybrid Thresholding for Fraud Detection Using XAI**
  This chapter introduces a hybrid thresholding methodology combining LIME and SHAP techniques for improved fraud detection performance and interpretability in industrial applications.

- **Chapter 6: Evolution of AI-Driven Decision-Making Systems**
  This chapter explores the evolution of intelligent systems—including Decision Support

Systems, Expert Systems, Recommender Systems, and XAI—through the lens of Industrial Applications. A case study on software defect prediction highlights how these systems enhance decision-making from data processing to actionable insights in industrial settings.

- **Chapter 7: Conclusion and Future Work**
  This chapter summarizes the thesis contributions, discusses overall findings and implications for industrial applications, highlights limitations, and outlines directions for future research in Explainable AI.

List of Publications

Journal Articles

1. Ravi, M., Negi, A. (2025). Evolution of AI-Driven Decision Making With Decision Support Systems, Expert Systems, Recommender Systems, and XAI. *IETE Technical Review*. Taylor & Francis, DOI: 10.1080/02564602.2025.2512086.

2. Ravi, M., Negi, A. (2025). Hybrid Thresholding for Enhanced Performance and Interpretability in Fraud Detection: Integrating LIME and SHAP for Trustworthy AI Based Decision Making. *Computers, Materials & Continua (CMC)*. ISSN: 1546-2226. Accepted for Publication.

3. Ravi, M., Negi, A. Leveraging LIME Explainability and Gustafson-Kessel Fuzzy Clustering for Resume Grouping and Semantic Text Summarization. Communicated to *Knowledge-Based Systems* (SCIE Journal, Revised and Resubmitted), Elsevier.

Conference Proceedings

1. Ravi, M., Negi, A. (2022). A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI. *Proceedings of the 2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, Mumbai, India, April 7-9, 2022. IEEE. DOI: 10.1109/I2CT54291.2022.9824265

2. Ravi, M., Negi, A. (2025). Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations. In: Sombattheera, C., Weng, P., Pang, J. (eds) *Multi-disciplinary Trends in Artificial Intelligence. MIWAI*

2024. Lecture Notes in Computer Science, vol. 15431. Springer, Singapore. DOI: 10.1007/978-981-96-0692-4_33

In the following chapter, we present a comprehensive review of the conceptual foundations of Explainable Artificial Intelligence (XAI) and its relevance to industrial domains. Particular emphasis is placed on their integration within the context of industrial applications and the specific problem domains addressed in this thesis, such as fraud detection and AI-driven recruitment.

Chapter 2

Literature Survey on Explainable AI and Industrial Applications

2.1 Introduction

This chapter presents a comprehensive review of important literature related to data-driven approaches using Explainable Artificial Intelligence (XAI) for industrial applications. It explores foundational work in interpretable machine learning, transparent decision-making systems, and the integration of XAI techniques across domains such as fraud detection, resume categorization, and fuzzy clustering. The objective is to highlight the prevailing research trends, uncover critical limitations in current methodologies, and establish a conceptual foundation for the data-driven, explainability-focused contributions presented in this thesis.

2.2 Industrial Applications and Decision-Making Systems

Broadly speaking, industrial domains are interested about the application of intelligent technologies within industrial environments. The environments would integrate sensors, software systems, and data analytics to enhance operational efficiency and support decision-making processes. Traditional systems, such as Decision Support Systems (DSS) and Expert Systems (ES) [370], represent early efforts toward automating industrial decision making.

Recommender Systems (RS), often based on collaborative filtering and content-based approaches, have become increasingly prevalent in industrial applications. However, they have been deprecated due to a lack of interpretability [312]. In recent years, there has been a growing

emphasis on developing explainable recommender systems and transparent decision-making interfaces to enhance user trust and accountability.

2.3 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) seeks to enhance the transparency and interpretability of machine learning models, enabling humans to better understand, trust, and manage automated decisions. Ribeiro et al. [310] introduced LIME, a model-agnostic technique that explains individual predictions by learning an interpretable model in the vicinity of the instance being explained. Similarly, Lundberg and Lee [216] proposed SHAP values, a unified framework grounded in cooperative game theory, to assign consistent and theoretically sound feature importance scores.

These methods are particularly critical in high-stakes domains where fairness, accountability, and transparency are essential. However, most existing XAI approaches are primarily tailored to supervised learning models, such as classifiers and regressors. Their application to unsupervised learning tasks, including clustering, remains limited and is an active area of research.

2.3.1 Where Did XAI Start?

Explainable Artificial Intelligence (XAI) originated from the need to understand, trust, and control AI systems, especially as they increasingly support or automate decision-making in high-risk domains. The foundational roots of XAI can be traced back to the early development of *Expert Systems (ES)* in the 1970s and 1980s. Systems such as *MYCIN* and *DENDRAL* were developed using explicit IF-THEN rule-based logic, making their reasoning processes inherently transparent and interpretable [60]. These systems were able to trace the rules fired during inference, allowing users to understand how conclusions were reached, almost reminiscent of expert human reasoning.

However, the shift from symbolic AI to statistical and connectionist paradigms in the 1990s, particularly with the advent of machine learning (ML) and deep learning (DL), introduced complex, black-box models such as neural networks and ensemble classifiers. Although these models achieved superior predictive accuracy, they lacked interpretability, making it difficult for end-users to comprehend, or justify the outputs—particularly in problematic domains such as healthcare, finance, law, and defense.

This growing concern led to a renewed focus on explainability. A significant milestone in the evolution of modern XAI was the launch of DARPA's Explainable Artificial Intelligence (XAI) program in 2016 [150]. The initiative aimed to develop machine learning models that could explain their behavior in ways understandable to humans, without significantly sacrificing performance. It emphasized user-centric design of explanations, acknowledging that effective explanation depends on the end-user's background, goals, and trust requirements.

Parallel to this, academic and industry researchers began creating post-hoc interpretability techniques[21], such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive explanations), which allowed opaque models to be probed and interpreted after training. These tools became the cornerstones of XAI by providing localized and global explanations for model behavior. While the principles of explainable AI have long existed in the form of rule-based systems, the field of XAI re-emerged with urgency as a response to the increasing opacity of high-performing ML/DL models. The intersection of transparency, trust, and accountability continues to drive the evolution of XAI, especially in mission-critical applications.

2.3.2 What is the Evolution of XAI Approaches?

The evolution of Explainable Artificial Intelligence (XAI) approaches reflects the ongoing balance between achieving *high predictive performance* and maintaining *interpretability* for human users [303]. Over time, this evolution can be broadly categorized into three distinct waves:

2.3.2.1 Symbolic and Rule-Based Methods

The earliest approaches to AI explanation were embedded directly into the design of *symbolic systems*, often referred to as *Expert Systems*. These systems, such as *MYCIN* and *PROSPECTOR*, were built using explicit IF-THEN rule-based logic that codified expert knowledge [338]. The inherently transparent nature of these rule-based architectures allowed the system to generate explanations by tracing the exact inference path taken during decision-making. This form of explanation was intuitive and aligned with human reasoning, enabling users to verify and trust the system's conclusions. However, these symbolic methods were limited by the complexity of knowledge acquisition and their inflexibility in assimilating new data.

2.3.2.2 Post-Hoc Explanations for Black-Box Models

With the increasing dominance of *statistical machine learning [358]* and *deep learning [199]* models from the 1990s onward, the AI landscape shifted toward data-driven, high-capacity models that often functioned as black boxes. Inherently Interpretable Tree Ensemble Learning[404]. These models provided state-of-the-art predictive accuracy but lacked transparency, as their internal decision processes were hidden in complex mathematical structures. To address this, researchers developed *post-hoc explanation techniques [171, 344, 345]* designed to interpret existing black-box models without modifying them.

Among the most influential methods are *LIME* (Local Interpretable Model-agnostic Explanations) by Ribeiro et al. [310] and *SHAP* (SHapley Additive exPlanations) by Lundberg and Lee [216]. LIME explains individual predictions by approximating the black-box model locally with an interpretable surrogate, such as a linear model, allowing users to understand which features most influenced a specific decision. SHAP builds on cooperative game theory to assign each feature a contribution value (Shapley value), providing consistent and theoretically grounded explanations that can be aggregated globally or viewed locally. These methods democratized explainability by offering model-agnostic tools applicable across a wide variety of ML models and domains, significantly enhancing transparency.

2.3.2.3 Inherently Interpretable Models

More recently, there has been a growing emphasis on designing *models that are inherently interpretable by construction*, thus avoiding the need for post-hoc explanations that may be approximations or potentially misleading. These models integrate interpretability as a core design principle, ensuring that every decision-making step can be directly inspected and understood by humans.

Examples include *attention-based neural networks*, where attention weights highlight important inputs; *rule lists and decision sets*, which provide compact and human-readable logical conditions; and *generalized additive models* that model outcomes as sums of feature contributions. Rudin [319] advocates for the use of such models in high-stakes domains to improve reliability and reduce the risk of misleading explanations. These inherently interpretable models strive to achieve a balance where performance and explainability are not mutually exclusive, providing transparent decision boundaries and rationale while maintaining competitive accuracy. The field of XAI has transitioned from built-in explainability in symbolic systems,

through the era of post-hoc interpretation of black-box models, toward the design of intrinsically transparent models. This evolution is driven by the increasing complexity of AI systems and the demand for trustworthy and accountable AI in real-world applications.

2.3.3 What Are the Important Developments in XAI?

Over the past decade, Explainable Artificial Intelligence (XAI) has evolved from a theoretical necessity to a practical imperative, leading to a wide array of important developments. These advances have focused on making machine learning models more interpretable, usable, and trustworthy in real-world settings. Key developments in XAI can be categorized into five broad areas:

2.3.3.1 Model-Agnostic Explanation Tools

One of the most impactful contributions to XAI has been the creation of *model-agnostic tools* such as LIME (Local Interpretable Model-agnostic Explanations) [310] and SHAP (SHapley Additive exPlanations) [21, 216]. These methods allow practitioners to interpret black-box models such as deep neural networks and ensemble models—without modifying their internal architecture. LIME works by locally approximating the black-box model with an interpretable surrogate (e.g., linear regression), while SHAP is based on cooperative game theory to assign a consistent importance value to each feature. These tools are widely adopted due to their flexibility and applicability across domains, and have become standard components in many machine learning pipelines.

2.3.3.2 Visualization Techniques for Interpretability

A significant body of work has emerged around visualization tools that enhance human understanding of model behavior. These include:

- **Saliency maps** for convolutional neural networks (CNNs)[383], which highlight the input regions most influential for a model's decision commonly used in image classification tasks.
- **Feature importance plots [86]**, particularly in tree-based models such as random forests and gradient-boosting, which rank features by their influence on model output.

- **Partial dependence plots (PDPs) [245]** and **accumulated local effects (ALE)** plots, which illustrate how changes in a feature affect the predicted outcome.

These visualization techniques serve to bridge the gap between raw model outputs and human interpretability, offering intuitive insights that support both debugging and trust-building.

2.3.3.3 Integrated Explanation Frameworks

Several comprehensive platforms have been developed to streamline and standardize explainability efforts across organizations. Notable examples include:

- **IBM AI Explainability 360 (AIX360)**: An open-source library that provides a wide range of explainability algorithms for data preprocessing, model interpretation, and post-hoc analysis.
- **Google's What-If Tool**: A visual interface for TensorBoard that enables users to explore counterfactuals, feature effects, and fairness issues without writing code.
- **Microsoft InterpretML**: A platform offering both glass-box (interpretable by design) and black-box explainers.

These frameworks aim to operationalize XAI, enabling consistent integration of interpretability in machine learning workflows while addressing concerns like fairness, bias detection, and compliance.

2.3.3.4 Application-Specific Advances

The push for explainability has led to significant progress in specialized domains where transparency is critical. In **healthcare**, for instance, researchers have developed interpretable deep learning systems to diagnose diseases such as diabetic retinopathy using attention-based CNNs and saliency visualizations. In **finance**, techniques such as hybrid thresholding with LIME and SHAP (e.g., fraud detection systems) have been deployed to ensure regulatory compliance and reduce risk exposure. The use of domain-specific constraints and interpretable architectures has allowed these systems to meet the dual objectives of accuracy and trustworthiness.

2.3.3.5 Human-in-the-Loop Decision Making

A growing focus in XAI is on the integration of explanations into human decision-making workflows. This involves the development of *human-in-the-loop* systems, where users actively interact with AI models through explanations to guide, verify, or override decisions. According to Doshi-Velez and Kim [108], a core goal of XAI is not just to produce explanations, but to improve the *mental model* of users interacting with AI. When explanations are appropriately designed and aligned with cognitive principles, they can enhance human trust, accountability, and performance.

Collectively, these developments mark a shift from post-hoc explanation as an auxiliary feature to explanation as a core component of AI system design. By making models more interpretable, these tools and frameworks improve transparency, foster user trust, and support responsible AI deployment in high-stakes environments.

2.3.4 What Are the Obstacles That Make XAI Application Difficult?

While Explainable Artificial Intelligence (XAI) has made significant strides in increasing the transparency and accountability of AI systems, its application in real-world settings remains fraught with several key challenges. These obstacles span technical, cognitive, ethical, and practical dimensions, hindering the widespread deployment of effective and trustworthy XAI systems.

2.3.4.1 Trade-off Between Accuracy and Interpretability

A persistent challenge in XAI is the perceived *trade-off between interpretability and predictive performance*. Highly interpretable models such as decision trees, linear regression, or rule-based systems are easy to understand but often lack the representational capacity to model complex, nonlinear relationships in data [242, 416]. In contrast, high-performing models like deep neural networks or ensemble methods (e.g., random forests, gradient boosting) offer superior accuracy but are typically opaque to human users. This trade-off becomes especially critical in high-stakes domains (e.g., healthcare, legal systems) where both performance and transparency are non-negotiable. Finding a balance between model fidelity and interpretability remains an active area of research.

2.3.4.2 Subjectivity of Explanations

Another fundamental difficulty lies in the *subjective nature of explanations*. As highlighted by Miller [235], what constitutes a ”good” or ”satisfactory” explanation often varies across individuals, tasks, and domains. For instance, domain experts may require highly technical justifications grounded in domain knowledge, whereas end-users may prefer simple, high-level rationales. This variability makes it difficult to design one-size-fits-all explanation methods, necessitating adaptive and user-centric explanation strategies. Additionally, human cognitive biases can influence how explanations are perceived, potentially undermining their intended purpose.

2.3.4.3 Scalability and Real-Time Constraints

Generating explanations, particularly for complex or large-scale models, is computationally demanding. Post-hoc methods like SHAP involve repeated model evaluations for each feature or instance, which can become infeasible in real-time systems. This *lack of scalability* restricts the deployment of XAI in scenarios requiring low-latency decisions (e.g., autonomous driving, fraud detection in financial transactions). Efficient approximation methods or lightweight interpretable models are being explored, but often at the cost of explanation richness or fidelity.

2.3.4.4 Evaluation of Explanation Quality

There is no universally accepted metric to evaluate the quality of explanations . Unlike classification accuracy, which is objectively measurable, explanation quality is often assessed via human studies, including user satisfaction, trust calibration, or task performance. These evaluations are inherently subjective and vary across contexts, making it difficult to benchmark or compare different XAI methods rigorously. Some researchers have proposed proxy metrics such as explanation-fidelity, sparsity, or stability, but these are imperfect substitutes for human judgment.

2.3.4.5 Security and Manipulation Risks

Paradoxically, the transparency enabled by XAI can also introduce *security vulnerabilities*. As shown by Aivodji et al. [11], adversaries can exploit explanations to reverse-engineer model

behavior or craft adversarial inputs that manipulate outputs without detection. This raises concerns about *fairwashing*—where seemingly fair explanations are used to mask underlying discrimination and *model inversion attacks*, where sensitive training data may be reconstructed from explanations. Therefore, ensuring the robustness of the explanation and protecting the confidentiality of the model are emerging priorities in the XAI community. While, XAI is vital for building trustworthy AI systems, its adoption is impeded by trade-offs, subjectivity, scalability concerns, lack of evaluation standards, and emerging security threats. Addressing these challenges requires interdisciplinary collaboration among AI researchers, cognitive scientists, ethicists, and domain practitioners to develop robust, human-aligned, and context-sensitive XAI solutions.

2.3.5 What is the Future of XAI?

As Artificial Intelligence continues to permeate decision-critical domains, the role of Explainable AI (XAI) is expected to grow in both scope and significance. The future of XAI will be shaped not only by technological advances but also by regulatory, social, and ethical imperatives. Several key directions are poised to define the next wave of XAI research and applications.

2.3.5.1 Domain-Specific Frameworks

A promising avenue for XAI lies in the development of *domain-specific explanation frameworks*. Generic, one-size-fits-all explanation techniques often fall short when applied to high-stakes areas such as healthcare, finance, human resources, or education. In these domains, explanations must be aligned with domain-specific reasoning patterns and compliance requirements. For instance, an interpretable system in healthcare may need to justify treatment recommendations using medically relevant indicators, while in finance, regulatory audit trails and risk justifications are crucial. Tailoring XAI frameworks to the semantics and workflows of specific domains will be central to increasing adoption and trust.

2.3.5.2 Human-Centered and Interactive AI

The future of XAI is increasingly human-centered. This means moving beyond static explanations to *interactive, adaptive, and user-personalized* explanation systems. These systems will incorporate *user feedback* into the explanation loop, refining the nature and presentation

of explanations based on user expertise, preferences and cognitive needs. Human-centered AI emphasizes the *co-adaptation* of human and machine decision-making, where interpretability is treated not as a one-off output but as part of a continuous learning dialogue between a user and a system.

2.3.5.3 Causal and Counterfactual Reasoning

Traditional post-hoc methods such as LIME and SHAP rely on statistical associations between input features and model predictions. However, the next frontier of XAI will move toward *causal* and *counterfactual explanations*—answering not just *what* led to a decision, but *why* and *how it could have been different*. As proposed by Wachter et al. [376], counterfactual explanations provide insights by showing how small changes in input could alter the output. This is especially useful for decision subjects (e.g., loan applicants) who want to know what they could change to receive a different outcome. These methods support actionable transparency and empowerment.

2.3.5.4 Regulatory Compliance and Ethical AI

XAI will play a crucial role in achieving *ethical and legal compliance* in AI systems. Regulations such as the European Union's General Data Protection Regulation (GDPR) have introduced a "right to explanation" for automated decisions, compelling organizations to justify algorithmic outputs to affected individuals. As regulatory frameworks evolve, explainability will become a legal necessity in compliance-heavy sectors such as finance, insurance, law, and public administration. This trend is expected to fuel investment in robust, auditable and transparent AI pipelines.

2.3.5.5 Interpretable-by-Design Systems

A critical shift in the future of XAI is the movement toward *interpretable-by-design models*—those that are intrinsically transparent rather than explained after deployment. Rather than relying on post-hoc techniques to open black boxes, researchers are now developing architectures where interpretability is a primary design constraint. Examples include sparse linear models, decision rule lists, attention mechanisms, and generalized additive models. These inherently interpretable models aim to combine competitive predictive performance with human-understandable reasoning, making them ideal for high-stakes applications where accountability

and trust are paramount. The trajectory of XAI is moving toward more interactive, domain-aware, and ethically grounded systems. The integration of causal reasoning, personalized feedback, regulatory alignment, and interpretable architectures will define the future landscape of trustworthy and human-centric AI.

2.3.6 Domains of XAI

Explainable Artificial Intelligence (XAI) has found applications across a wide array of domains, particularly in areas where decision transparency, accountability, and user trust are critical. The demand for interpretability varies by field, depending on the consequences of incorrect predictions, regulatory requirements, and stakeholder involvement. Below are some key domains where XAI has made significant contributions:

2.3.6.1 Healthcare and Biomedical Informatics

In healthcare, AI systems assist in all of these: diagnosis, treatment planning, medical imaging, and patient risk stratification [251]. However, clinicians require clear justifications for recommendations due to the high stakes involved. XAI helps provide traceable and medically relevant explanations for decisions made by black-box models such as deep convolutional neural networks (CNNs). For instance, saliency maps and attention mechanisms highlight the regions of interest in medical scans (e.g., for detecting diabetic retinopathy), thereby enhancing clinical decision-making and fostering trust in AI-assisted diagnostics [392].

2.3.6.2 Finance and Credit Risk Assessment

Financial institutions increasingly employ machine learning models for fraud detection, credit scoring, and loan approvals. Given the heavy regulatory oversight of the sector, XAI is essential to provide justification to regulators and customers. Techniques such as SHAP and LIME are used to explain why a loan was denied or flagged as risky, allowing institutions to demonstrate fairness and avoid algorithmic discrimination. Moreover, counterfactual explanations help applicants understand what could be changed to improve future outcomes [62].

2.3.6.3 Human Resources and Resume Filtering

In recruitment systems, AI is used for resume parsing, ranking, and categorization. XAI plays a key role in mitigating biases and ensuring fairness, especially with sensitive features such

as gender, ethnicity, or age. Techniques like Local Interpretable Model-agnostic Explanations (LIME) and SHAP are applied to explain why a candidate is selected or rejected, fostering transparency in hiring pipelines. In our own work, we demonstrated the effectiveness of combining KNN classifiers with LIME explanations and fuzzy clustering for interpretable resume categorization and profiling [160] and [18].

2.3.6.4 Education and Intelligent Tutoring Systems

XAI is increasingly integrated into adaptive learning systems and intelligent tutoring environments. These systems personalize feedback and learning paths for students. Explainability helps educators and learners understand why specific recommendations or assessments are made, aiding in trust-building and pedagogical insights. For example, interpretable models can justify predicted knowledge gaps or recommend study resources based on traceable learning patterns [87] and [83].

2.3.6.5 Autonomous Systems and Robotics

In autonomous vehicles, drones, and industrial robots, real-time decisions must be interpretable to ensure safety and regulatory compliance. XAI is used to explain navigation choices, anomaly detection, or object recognition decisions. For example, explaining why an autonomous car chose a particular path or reacted to an obstacle can assist in post-incident analysis and liability assessment [355], [324], and [106].

2.3.6.6 Security and Legal Systems

XAI is relevant in surveillance, cyber intrusion detection, and legal decision support. Legal systems require AI decisions to be transparent and explainable, especially when used in sentencing, bail decisions, or case prioritization. In cybersecurity, explainable models help analysts understand threat predictions, root causes of intrusions, and the rationale behind alerts. The growing integration of AI into critical decision-making domains necessitates that models be interpretable, auditable, and aligned with human reasoning. XAI supports transparency not just as a technical feature, but as a core requirement for ethical, legal, and responsible AI deployment across domains [39] and [185].

2.3.7 Criticism of XAI

Despite the increasing adoption and enthusiasm around Explainable Artificial Intelligence (XAI), the field has attracted substantial criticism from both technical and philosophical perspectives. These critiques raise important concerns regarding the validity, usability, and ethical implications of explanations provided by AI systems.

2.3.7.1 Faithfulness and Fidelity

One of the primary concerns is whether the explanations are *faithful* to the actual reasoning of the underlying model. Post-hoc methods such as LIME and SHAP generate approximations, which may not always accurately reflect the internal logic of complex models. This can lead to misleading explanations, especially in cases where the explanation simplifies nonlinear interactions or masks confounding variables. Rudin [319] strongly advocates for using inherently interpretable models over black-box models with post-hoc explanations, arguing that approximations may obscure important decision factors and give users a false sense of transparency.

2.3.7.2 Explanation Versus Justification

Another critique centers on the distinction between *explanation* and *justification*. While explanations aim to provide insight into how a model works, justifications attempt to rationalize decisions in a way that aligns with human expectations. Critics argue that many XAI methods, especially model-agnostic ones, often generate outputs that serve more as justifications than true explanations. These may be tailored to appease users rather than reveal the actual mechanics of the decision-making process, leading to a phenomenon known as "fairwashing" [11].

2.3.7.3 Lack of Standardized Evaluation

XAI methods lack standardized metrics for assessing explanation quality. Evaluations are often context-dependent, relying on subjective human judgments or ad hoc criteria such as sparsity, stability, and consistency. This makes it difficult to benchmark methods or establish best practices. Doshi-Velez and Kim [108] emphasize the need for a rigorous science of interpretability, including frameworks for evaluating how explanations influence human understanding, trust, and performance.

2.3.7.4 Cognitive Bias and User Misinterpretation

Even when explanations are technically accurate, they may be misinterpreted by users due to cognitive biases or lack of domain knowledge. Users may over-trust or under-trust systems based on how explanations are framed—a phenomenon known as the *illusion of explanatory depth*. This can undermine the intended purpose of XAI, particularly when decisions must be reviewed or overridden by human experts. The effectiveness of an explanation depends not only on content but also on delivery, user background, and task context [51].

2.3.7.5 Ethical and Legal Concerns

Critics also point out that XAI, if poorly implemented, may be used to deflect accountability or satisfy regulatory requirements superficially. Superficial transparency - where explanations exist only for compliance rather than genuine insight - risks reducing XAI to a checkbox exercise. Furthermore, explanation mechanisms can inadvertently expose model vulnerabilities, making systems more susceptible to adversarial attacks or privacy breaches. While XAI holds promise for enhancing AI transparency, several unresolved challenges persist. Critics urge caution in over-relying on post-hoc explanations and advocate for inherently interpretable models wherever possible. Future progress in XAI will depend not just on technical innovations, but also on addressing philosophical, cognitive, and ethical dimensions of what it truly means to "explain" intelligent behavior [351] and [265].

2.4 XAI in Fraud Detection

Fraud detection is a critical area of industrial application where precision and interpretability are vital. Traditional models such as Logistic Regression and Random Forest have been used extensively. However, thresholding remains a bottleneck. Bhattacharyya et al. [53] addressed performance metrics under class imbalance.

Our prior work [301] proposed a hybrid thresholding mechanism that combines ROC and PR curves and integrates LIME and SHAP to enhance model interpretability and performance in fraud detection tasks.

2.5 XAI in Resume Classification and Recruitment

AI-based recruitment tools face challenges related to fairness, bias, and lack of transparency. Models like KNN and SVM have been applied to classify resumes based on skills and job titles. Yet, very few systems offer interpretable outputs. LIME has been used to explain decisions in candidate selection processes [302].

Our earlier chapter proposed a framework combining LIME with traditional models to improve interpretability in resume classification, thus aligning with responsible AI adoption in HR analytics.

2.6 Fuzzy Clustering and Semantic Embeddings

Fuzzy clustering algorithms, such as Gustafson-Kessel (GK) [154], extend traditional clustering methods by adapting to non-spherical clusters using covariance matrices. These are particularly effective for high-dimensional and weakly structured data like resumes.

Sentence-BERT [307] offers semantically rich embeddings suitable for textual clustering. Few studies, however, have explored combining fuzzy clustering with XAI for transparent grouping of unlabelled text data.

Our proposed model integrates GK fuzzy clustering with Sentence-BERT embeddings and LIME to generate interpretable clusters for resumes, contributing a novel direction in unsupervised explainable AI.

2.7 Summary and Research Gaps

The literature reveals substantial advances in Explainable Artificial Intelligence (XAI) and its applications across various domains. However, several key research gaps persist in the context of industrial applications.

- **Limited Integration of XAI in Unsupervised Learning:** While XAI techniques like LIME and SHAP have gained traction in supervised models, their application in unsupervised learning, particularly clustering—remains underexplored. This hinders transparency in decision-making processes where there are no ground truth labels.

- **Lack of Semantic-Aware Clustering in Resume Analytics:** Most resume categorization methods rely on traditional keyword-based or statistical approaches. Few models leverage contextual embeddings (e.g., Sentence-BERT) in combination with adaptive clustering techniques (e.g., Gustafson-Kessel) to capture deep semantic similarities among resumes.

- **Inadequate Thresholding Strategies in Fraud Detection:** Conventional thresholding methods (e.g., fixed or ROC-based) often fail to balance precision and recall effectively in imbalanced datasets. Hybrid thresholding approaches that integrate performance metrics with XAI tools are needed to improve both detection accuracy and interpretability.

- **Software Development approach of AI Systems in Industrial Applications:** Existing work often treats expert systems, recommender systems, and XAI frameworks in isolation. A holistic framework combining interpretability, adaptability, and domain-specific constraints is lacking in industrial decision-making systems.

- **Transparency and Fairness in Recruitment Automation:** While AI is increasingly used in hiring pipelines, few systems offer clear explanations for model decisions. This raises concerns about bias and fairness, making the integration of XAI essential for ethical and accountable AI recruitment tools.

2.7.1 Conclusion

These research gaps underscore the necessity for a unified, explainable, and domain-adaptable AI framework—addressed in this thesis through hybrid thresholding, interpretable clustering, and contextual resume summarization within a data-driven paradigm using Explainable AI (XAI) for industrial applications.

This chapter has reviewed the historical origins, evolution, key developments, challenges, future directions, domains of application, and criticisms of Explainable Artificial Intelligence (XAI). From its roots in rule-based Expert Systems to modern model-agnostic interpretability tools and inherently interpretable models, XAI has become an essential pillar in the development of transparent, trustworthy AI systems. Despite promising advancements, XAI continues to face challenges related to fidelity, subjectivity, scalability, and ethical deployment. However, its increasing integration into critical industrial domains, such as manufacturing, healthcare, finance, and human resources, reinforces its growing importance and practical value.

2. LITERATURE SURVEY ON EXPLAINABLE AI AND INDUSTRIAL APPLICATIONS

In the next chapter, we delve into a detailed review of the conceptual foundations underpinning data-driven approaches using Explainable AI for industrial applications, with a particular focus on the problem domains addressed in this thesis.

Chapter 3

Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations

In the previous chapter, literature related to XAI was reviewed in the context of industrial applications. In this chapter, we propose an interpretable resume categorization framework that integrates the K-Nearest Neighbors (KNN) algorithm with Local Interpretable Model-agnostic Explanations (LIME). This approach addresses the dual objectives of achieving high classification accuracy and ensuring transparency in automated recruitment systems. By applying TF-IDF vectorization and optimizing KNN parameters —along with LIME's local explanations for individual predictions —the framework fosters trust in AI-driven decision-making, which aligns with the overarching goal of this thesis: advancing data-driven approaches using Explainable AI for industrial applications [302].

3.1 Introduction

In today's digital and industrial landscape, efficient workforce management leads to automated resume processing. AI-driven resume classification systems are a prime example of crucial, trustworthy decision-making systems for industrial applications that ensure transparency and interpretability. This research leverages the Local Interpretable Model-agnostic Explanations

(LIME) framework to make complex models understandable, recognizing that resumes represent unique professional journeys. By addressing the opacity of AI algorithms, this approach supports fair, accountable, and human-centered recruitment processes in industrial systems.

LIME helps reveal the thought processes behind model decisions. By employing LIME with models like K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Logistic Regression, we can understand not just what the models predict but how and why. LIME deciphers the language of these models, providing insights into the influential terms within resumes.

This approach fosters transparency and trust by exposing biases and illuminating decision boundaries. It transforms the landscape from one dominated by faceless algorithms to an arena where humans and machines collaborate. By revealing the inner workings of algorithms, LIME makes interpretability a cornerstone of ethical and trustworthy AI.

Our chapter is structured as follows: Section 3.2 describes the Automated Resume Categorization, Section 3.3 details the Proposed Methodology, Section 3.4 presents the Results and Discussion, and Section 3.5 provides the Conclusion.

3.2 Automated Resume Categorization

Automated resume categorization explores the integration of K-nearest neighbors (K-NN) algorithms and Explainable Artificial Intelligence (XAI) techniques. Studies highlight the challenges of using machine learning for recruitment, emphasizing the need for models that balance accuracy with ethical considerations [318]. K-NN algorithms are prominent in assessing the suitability of candidates based on job descriptions, addressing biases in automated hiring [133, 228]. While K-NN is widely used in classification tasks, its application to resume categorization is less explored [90, 103]. This research fills that gap by implementing K-NN in resume analysis and enhancing categorization processes. Ethical considerations are crucial in AI recruitment systems [372]. Yu et al. reported a cascaded hybrid model for extracting resume information, which depends on the structure of the resume and is time consuming [412]. Kopparapu et al. proposed an automatic information extraction method from unstructured resumes with good precision and recall [189]. Text analytics, a rapidly growing field, transforms unstructured data into structured data for better classification and mining [129]. A Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction method trained by SVM yielded good classification results [93]. Manchanda et al. used a remote multi-class

supervised technique for dynamic text segmentation [222], while Jo et al. employed K-NN for text segmentation based on similarity features [173]. The literature underscores the need for ethical frameworks, transparency, and user trust in AI systems, particularly in recruitment [107]. Clear explanations of the model are essential to promote transparency and confidence in decision-making processes.

3.2.1 Introducing Interpretability through LIME

The reason we chose LIME (Local Interpretable Model-agnostic Explanations) for our approach is that it can offer interpretability that is vital for intricate machine learning models like Random Forest, KNN, and 3-layer MLP. LIME may be used with any classifier because to its model-agnostic nature, which also makes it possible to simplify complicated models into more understandable ones and provide detail on specific predictions. This promotes openness and confidence in the system by emphasizing important characteristics that have a big influence on classification choices, making sure that users, including HR specialists, can comprehend and verify the model's findings. Fairness in automated resume categorization is ensured and black-box models become more interpretable through the use of Explainable Artificial Intelligence (XAI) approaches such as LIME [311], [419], [78], [306]. This also ensures compliance with ethical standards.

3.2.2 Problem Definition

The literature review highlights a gap in automated resume categorization: balancing high accuracy of models like K-Nearest Neighbors (KNN) with the need for transparency. Current issues include model opacity, decision-making biases, and the need for real-world applicability. The main challenge is creating a resume categorization framework that ensures high predictive accuracy while providing clear and interpretable insights. The proposed solution integrates Local Interpretable Model-agnostic Explanations (LIME) to offer localized, understandable explanations for individual predictions, thus addressing these challenges and improving transparency in automated recruitment.

3.2.3 Problem Statement

The challenge is to develop an automated resume categorization system that balances high accuracy with transparency. The solution involves integrating Local Interpretable Model-agnostic

Explanations (LIME) to clarify the decision-making of complex models like K-Nearest Neighbors (KNN). This approach aims to ensure a trustworthy and fair recruitment system by providing clear, interpretable insights while maintaining predictive performance.

3.2.4 Mathematical formation and formulation

Let X represent resumes and Y the set of categories. Each resume x_i has a ground truth label $y_i \in Y$. The goal is to find a function $f : X \to Y$ that maps resumes to categories. In machine learning, this function is learned from a labeled dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. The K-Nearest Neighbors (KNN) algorithm computes the category of a resume x_q based on its k nearest neighbors $N_k(x_q)$:

$$f_{\text{KNN}}(x_q) = \arg\max_{y \in Y} \sum_{x_i \in N_k(x_q)} \delta(y_i = y) \tag{3.1}$$

The challenge is to balance accuracy.

$$(f_{\text{KNN}}(x_i) \approx y_i) \tag{3.2}$$

with transparency as shown in Eq (1). and Eq (2). We introduce an interpretability function

$$I : X \to R \tag{3.3}$$

that measures a model's decision interpretability it shown in Eq (3). The problem is to find a model balancing accuracy and interpretability:

$$f^* = \arg\max_{f \in \mathcal{F}} \alpha \cdot \text{Accuracy}(f, D) + (1 - \alpha) \cdot \text{Interpretability}(f, I) \tag{3.4}$$

where $\mathcal{F}$ is the space of possible models, and α controls the trade-off between accuracy and interpretability as shown in Eq (4). Integrating Local Interpretable Model-agnostic Explanations (LIME) can enhance interpretability. In this formulation guides developing a resume categorization framework achieving high accuracy and transparency. In the following section describes the methodology.

3.3 Proposed Methodology

Our proposed methodology for multiclass resume classification integrates various machine learning classifiers with advanced interpretability techniques. The framework consists of several key components, each contributing to the overall performance and transparency of the classification system.

3.3.1 Framework Overview

The proposed framework is designed to classify resumes into predefined job categories using a combination of machine learning models and interpretability methods.

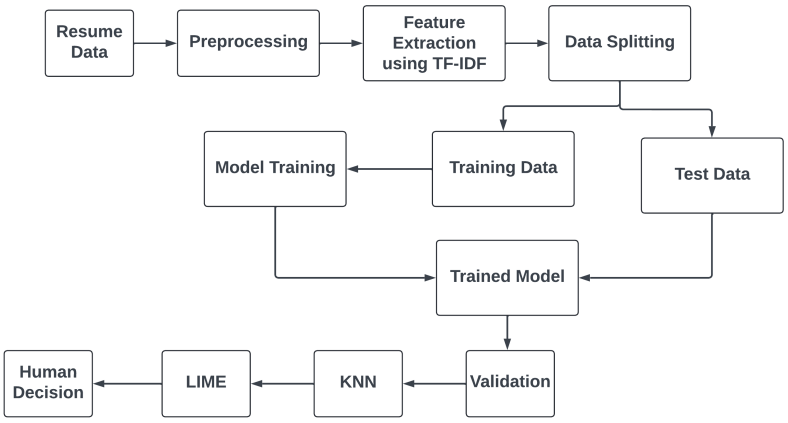


Figure 3.1: Proposed framework integrating K-Nearest Neighbors (KNN) with LIME for transparent and interpretable automated resume categorization.

Figure 3.1 depicts the workflow of our method, which consists of multiple essential elements. The data set is pre-processed at the beginning of the process to make it suitable for analysis. Feature extraction is performed after preprocessing to find pertinent attributes for categorization. The data is then divided into testing and training sets. The model is trained using the training set, and the trained model is then validated. KNN is used as the main classifier in this system, and LIME is used to improve interpretability by providing context for each prediction. Ultimately, human judgment is incorporated into the process to examine and validate the model's results. Our approach's data preprocessing is described in the below steps.

3.3.2 Data Preprocessing

Resumes and job classifications are included in the dataset, which was obtained from Kaggle[1], includes resumes and job categories. Given a dataset $\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, where x_i represents the raw resume text and y_i the corresponding job category. Imports all python libraries like pandas, numpy, pipeline, nltk, spacy, re, matplotlib, Label Encoder, TfidfVectorizer, ensemble classifiers, GridSearchCV, train_test_split, accuracy_score, classification_report, LimeTextExplainer. It is loaded into a Pandas DataFrame, where the goal is to map resumes x_i to categories y_i using a labeled dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Text preprocessing involves:

$$x_i' = \text{Clean}(x_i) \tag{3.5}$$

where $\text{Clean}(x_i)$ removes URLs, hashtags, mentions, and special characters [156]. The cleaned text is then tokenized and converted to lowercase. Stopwords are removed:

$$x_i'' = \text{Vectorize}(\text{Tokenize}(x_i')) \tag{3.6}$$

where $\text{Tokenize}(x_i')$ splits the text into tokens, and $\text{Vectorize}(x_i'')$ converts tokens into numerical features using TF-IDF. Tokenization, lemmatization and stemming are applied to further normalize and reduce dimensionality. Categorical labels are encoded as [14]:

$$y_i' = \text{Encode}(y_i) \tag{3.7}$$

The dataset is split into training and test subsets:

$$\{(X_{\text{train}}, Y_{\text{train}}), (X_{\text{test}}, Y_{\text{test}})\} = \text{Split}(\mathcal{D}) \tag{3.8}$$

The cleaned resumes are transformed into numerical features using TF-IDF vectorization. The dataset is then split into training and test subsets, denoted as $(X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}})$, with 20% of the data allocated to testing. These steps prepare the data for machine learning model training and evaluation.

[1] https://www.kaggle.com/code/avantika2001/resume-screening/input

3.3.3 Model Selection and Hyperparameter Tuning

A pipeline combining TF-IDF vectorization with k-Nearest Neighbors (kNN) classification is employed. GridSearchCV is used to optimize hyperparameters, including the number of neighbors, weighting method, and distance metric. The parameter grid for this search is as follows: This grid includes the number of neighbors for the K-Nearest Neighbors classifier, with values [3, 5, 7, 9], the weighting method ('uniform' or 'distance'), and the distance metric ('Euclidean' or 'Manhattan'). The best model is selected based on accuracy.

3.3.4 Explainability with LIME

To interpret the model's predictions, LIME (Local Interpretable Model-agnostic Explanations) is utilized. The LIME text explainer is configured to explain individual predictions by highlighting the contributions of various features. The explanation is then visualized and saved for further analysis. Using LIME (Local Interpretable Model-agnostic Explanations) to explain

Algorithm 1: LIME for K-Nearest Neighbors (kNN)

Input:
  Instance x to be explained
  Pipeline P (includes TF-IDF and kNN)
  Number of features $num\_features$ to include in the explanation
Output:
  LIME explanation object
Procedure:
1. Extract class names from the label encoder used in the pipeline
2. Create a LIME Text Explainer with the extracted class names
3. Use the LIME explainer to generate an explanation
  For the instance x, compute perturbed samples $X' = \{x'\_1, x'\_2, \ldots, x'\_N\}$
  Compute proximity weights w_i for each x'_i based on similarity to x
  Train a locally interpretable model g using weighted least squares with weights w_i
Explanation:
  Use the LIME Explainer to fit the local model g with instance x
  Specify the number of features $num\_features$ to include in the explanation.
4. Return the LIME explanation object.

K-Nearest Neighbors (kNN) classifier predictions is the method described in the algorithm. It starts by explaining an instance x and the machine learning pipeline, which consists of the kNN model and TF-IDF for text processing. It also specifies how many features the explanation should have. The process involves first collecting the class names from the pipeline's label

encoder and configuring a LIME Text Explainer with those class names. The altered samples are generated in the vicinity of the instance x, and proximity weights are calculated by comparing them to x. Next, weighted least squares are used to train a locally interpretable model g, where the weights represent the proximity of each altered sample. Ultimately, the LIME explainer chooses a certain amount of features to incorporate in the explanation and fits this local model to the original instance x. The result is a LIME explanation object that tells you which features contributed most to the model's prediction for that specific case.

3.3.5 Incorporating Alpha Terms in LIME Interpretations

As stated in Equation (4)'s Section 2. Specifically, when applying LIME to classifiers like KNN, alpha terms are utilized to control the influence of specific features on model predictions in our approach. The local model's feature weighting is adjusted via alpha terms, which affect how much each feature contributes to the prediction. For example, an alpha parameter is set to adjust the importance of specific terms while using LIME to interpret a KNN model for resume categorization. "Software engineer" is one term that is given more weight in the local model's choice if it has a high positive alpha. On the other hand, a word with a low or negative alpha has less power. This strategy contributes to improving interpretability, striking a balance between feature importance and guaranteeing clear and intelligible model predictions.

3.3.6 Model Evaluation

The performance of each trained model is evaluated using various metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the models' capabilities in classifying resumes into predefined categories. Evaluation is carried out on the $testset(X\_test\_tfidf, y\_test)$, allowing an assessment of how well the models generalize to unseen data. In order to promote transparency and interpretability in automated resume categorization, we created a K-Nearest Neighbors (KNN) classifier with Local Interpretable Model-agnostic Explanations (LIME). This method addresses any biases in recruitment while ensuring high accuracy and fairness. In the following section describes results and discussion.

3.3.7 Applications

Human resources, career counseling, educational institutions, and resume screening tools are just a few of the practical uses for the method's findings. With resume management, these apps

increase effectiveness, equity, and transparency. By adding more features and cutting-edge methods, along with continual feedback from stakeholders, future development could improve the system.

3.4 Results and Discussion

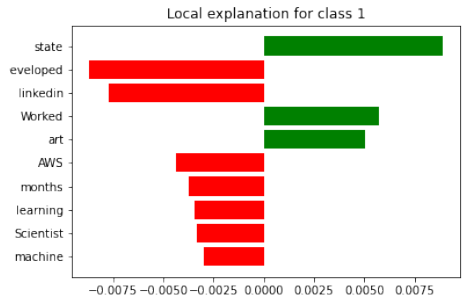
In this section, we describe the corresponding results of our method. Random Forest, Support Vector Machine (SVM), Logistic Regression, and a 3-layer Multi-Layer Perceptron (MLP) are some of the machine learning classifiers that we use in our multiclass resume categorization framework. We mainly use LIME (Local Interpretable Model-agnostic Explanations) on the KNN classifier to improve model interpretability. Using a locally interpretable approximation of the complex KNN model, LIME highlights resume terms that have a major impact on categorization decisions, hence offering insights into individual predictions.

We further apply LIME to additional models, including Random Forest, a three-layer multilayer perceptron (MLP), and Decision Tree, to provide a comprehensive understanding of individual predictions and their integration into the overall system. This interpretability-focused approach enhances the credibility and transparency of the resume classification framework. Among all classifiers, the K-Nearest-Neighbors (KNN) model demonstrates superior performance. LIME explanations for KNN are shown in Figure 3.2a, while Figures 3.2b to 3.2f present interpretability results for Random Forest, Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN), and Decision Tree (DT), respectively. The complete set of visualizations is summarized in Figure 3.2.

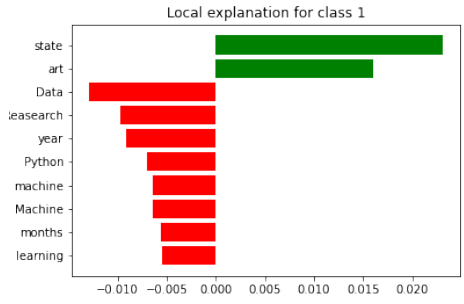
Understanding machine learning model decisions is vital, particularly in sensitive areas like resume classification. Interpretability ensures trust and comprehension for stakeholders. This research uses Local Interpretable Model-agnostic Explanations (LIME) to enhance the interpretability of K-Nearest Neighbors (KNN). The comparison between random forest, SVM, and logistic regression methods is also presented.

3.4.1 LIME Overview in Our Methodology

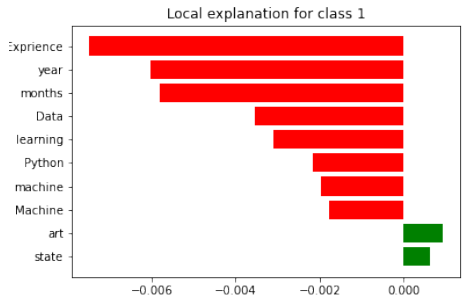
To improve the interpretability of complex models, such as KNN, Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), 3-layer Multi-Layer Perceptron (MLP), and Decision Tree, our system employs LIME (Local Interpretable Model-agnostic Explanations). By altering the feature values of a resume and using a more basic surrogate model to learn the



(a) KNN

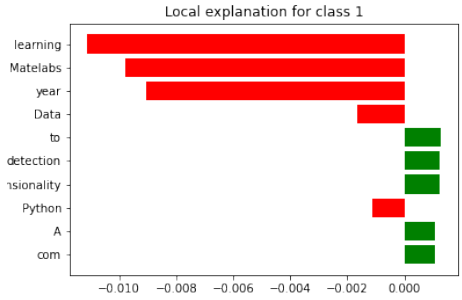


(b) Random Forest

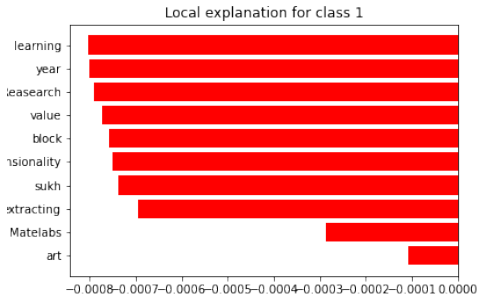


(c) SVM

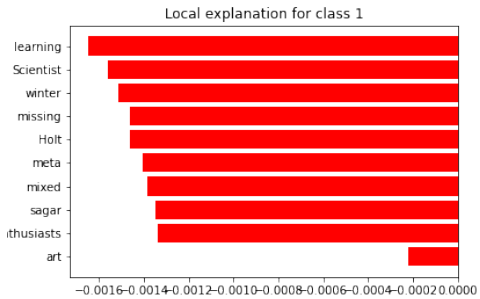
(a) LIME explanations for KNN, RF, and SVM



(d) Logistic Regression



(e) ANN



(f) Decision Tree

(b) LIME explanations for LR, ANN, and DT

Figure 3.2: Side-by-side comparison of LIME explanations for six classifiers on resume data

decision function of the more complex model, LIME offers local explanations. This method improves transparency and confidence in our automated resume categorization system by locating and emphasizing key terms in resumes that have a substantial impact on classification judgments.

3.4.2 Interpretation: Resume data

Each term in the resume data is assigned a significance score and the importance of characteristics is shown in our study as key-value pairs. Terms with good scores have a favorable influence on the KNN model's categorization, whereas terms with negative scores indicate that red color has a negative effect. Higher absolute values indicate the green color in the figure with greater relevance, while the magnitude of these scores represents the strength of their influence. By offering a clear understanding of how particular features affect the model's conclusions, LIME (Local Interpretable Model-agnostic Explanations) improves the interpretability of these scores, making it simpler to interpret and put your trust in the predictions of complex models like KNN.

3.4.3 Comprehensive Evaluation

The KNN-based LIME method is shown in figure 5.2 and table3.1 results. The significance of each feature (word or term) in the immediate vicinity of a data point determines the importance of the features in KNN. Each prediction given by the KNN classifier has a thorough explanation provided by the LIME approach. During our tests, we used the KNN classifier with the following hyperparameters: $k \in \{3,5,7,9\}$, weight options of 'distance' or 'uniform', and distance metrics such as 'Euclidean' and 'Manhattan'. With a precision of 0.995, recall of 0.997, and an F1 score of 0.995, the classifier demonstrated strong performance metrics and a stable capacity to classify resumes with a balanced trade-off between precision and recall. Positive significance scores in the context of LIME highlight qualities that are relevant in the local context by contributing positively to the predicted class, these features are shown in the figure as green. On the other hand, negative scores indicate characteristics that make the predicted class less likely; in the figure, these qualities are indicated by red. This methodology guarantees a transparent and comprehensible decision-making process within the model, offering lucid insights into the influential elements within each forecast.

In comparison, the Random Forest model achieves a slightly lower accuracy of 0.984, with a precision of 0.989, recall of 0.970, and F1-Score of 0.972. Although it shows high importance scores for certain features, its overall contribution to features is less pronounced compared to KNN. The RF model benefits from its ensemble approach but may lack the clarity offered by LIME to reveal the impact of characteristics.

The SVM and LR models both achieve an accuracy of 0.994, with precision and recall values similar to those of KNN. These models demonstrate high performance, but their feature importance is less transparent without additional interpretability tools. Logistic regression, with its coefficient-based importance, provides clear insights into feature contributions, though it performs marginally lower than KNN in recall.

The MLP model, while achieving high accuracy and F1-Score (both 0.994), is noted for its complexity and longer training times. This complexity makes it less practical for quick interpretations compared to KNN. Despite its strong performance, the extended training time of MLP emphasizes the advantage of simpler models like KNN for scenarios requiring rapid insights.

The Decision Tree model, with an accuracy of 0.974 and an F1-Score of 0.963, provides straightforward interpretability but falls short of the performance metrics of other models. Its ability to rank features based on impurity reduction is useful, yet it may not capture feature interactions as effectively as more sophisticated models.

3.4.4 Limitations and Future Work

Data biases and the computational complexity of models such as the Multi-Layer Perceptron (MLP) are among the constraints of this work. Subsequent research endeavors should focus on improving these concerns by optimizing pre-processing stages, investigating additional functionalities, and assessing impartiality and moral implications in automated resume categorization. More sophisticated methods and hybrid models may improve interpretability and performance even more. The relationship between the number of neighbors (k) and model accuracy is plotted to understand the effect of k on performance. Additionally, the importance of LIME features is visualized to provide insights into which features are most influential in the model's predictions.

3.4.5 Quantitative results of proposed method in comparison to other state-of-the-art-methods

Table 3.1: Comparison of our method with existing model [20].

S.no	Classifiers	Precision	Recall	F1-Score	Our-method	Ali-method	Mis-class
1	**Proposed**	**0.995**	**0.997**	**0.995**	**0.994**	0.972	0.006
2	RF	0.989	0.970	0.972	0.984	No	0.016
3	SVM	0.992	0.997	0.994	**0.994**	0.993	0.006
4	LR	0.992	0.997	0.994	**0.994**	0.993	0.006
5	MLP	**0.995**	**0.997**	**0.995**	**0.994**	No	0.006
6	DT	0.975	0.975	0.963	0.974	No	0.026

Note: RF = Random Forest, SVM = Support Vector Machine, LR = Logistic Regression, MLP = Multi-Layer Perceptron, DT = Decision Tree.
Mis-class refers to misclassification rate, and *Ex-Accuracy* indicates existing accuracy from [20].

Table 3.1 describes evaluation metrics in resume data. While MLP offers high performance, its complexity and longer processing times make KNN a practical choice for scenarios demanding both high efficiency and interpretability. LIME's ability to elucidate both positive and negative feature impacts in MLP enhances our understanding of model behavior, bridging the gap between high-performing yet complex models and the need for clear, actionable insights.

3.5 Conclusion

This chapter contributes to the thesis objective of developing data-driven approaches using Explainable AI for industrial applications by advancing an interpretable framework for automated resume categorization. By integrating KNN with LIME, the approach improves model transparency, interpretability, and fairness, which are key elements for AI-driven trustworthy decision making in recruitment systems. It supports ethical adoption of AI while addressing practical challenges such as bias and privacy. This work demonstrates how Explainable AI can improve operational effectiveness and stakeholder trust in industrial contexts, paving the way for more responsible and adaptive intelligent systems. In the following chapters, we will delve into more specific applications and evaluations.

Chapter 4

Leveraging LIME Explainability and Gustafson-Kessel Fuzzy Clustering for Resume Grouping and Text Summarization

In the previous chapter, we propose an interpretable resume categorization framework that integrates the K-Nearest Neighbors (KNN) algorithm with Local Interpretable Model-agnostic Explanations (LIME). Building upon this foundation of explainable AI in industrial applications, the current chapter introduces an innovative approach that combines the Gustafson-Kessel (GK) fuzzy clustering algorithm with advanced semantic embeddings generated by Sentence-BERT to improve the interpretability of unsupervised learning for resume data. By integrating LIME, the proposed framework offers transparent insights into cluster memberships, addressing a key challenge in applying unsupervised techniques within data-driven industrial environments. This method advances applicant profiling and human resource management by delivering effective and interpretable clustering results, thereby contributing to trustworthy AI-driven decision-making in industrial contexts.

4.1 Introduction

Modern business pays a lot of attention to employing the right person for their job vacancies. This task is as important as it is arduous. To attract the best talent, companies would like to

be perceived as fair and objective employers when such appointments are made. However, traditional human resource (HR) approaches in the industry are not seen to be scalable against the very large number of applications. Further, conventional HR approaches appear to be subjective and of doubtful efficacy.

Recent advances using Large Language Models (LLMs) have shown promise in automating aspects of resume analysis (e.g., ResumeFlow [424]). However, LLMs are heavyweight approaches and are prone to well-known hallucination problems, whose resolution is a non-trivial research challenge in itself [124, 280]. To avoid these complications, we adopt a lightweight strategy for resume grouping. In addition to classification accuracy, interpretability and explainability of results are crucial in recruitment contexts, since HR professionals must trust and understand the basis of automated recommendations.

In this study, we combine interpretability and explainability with computationally lean clustering methods. Specifically, we adopt the Gustafson–Kessel (GK) fuzzy clustering algorithm to group resumes based on their textual representations, and we augment the clustering process with explainable AI (XAI) techniques. This grouping aims to uncover latent semantic similarities in resumes while providing interpretable explanations for HR decision support.

The exponential rise in unstructured textual data in domains such as Human Resource Management (HRM) has created a pressing need for automated techniques capable of efficiently analyzing and categorizing text. Resume processing is one such application, where clustering can facilitate talent pooling, job matching, and candidate selection. However, most clustering approaches lack transparency, making it difficult for end-users to interpret and trust results. Our approach addresses this limitation by integrating Local Interpretable Model-Agnostic Explanations (LIME) [310] and SHapley Additive exPlanations (SHAP) [216] with GK clustering. To achieve this, we train a Random Forest surrogate model to approximate fuzzy cluster assignments, enabling the application of supervised explainability techniques in an unsupervised setting. LIME provides local explanations for individual resumes, while SHAP provides global feature-importance insights, together bridging the gap between clustering outputs and human-understandable reasoning.

4.1.1 Contributions

The primary contributions of this study are as follows:

1. **Leveraging contextual embeddings for semantic resume representation:** We employ Sentence-BERT embeddings (specifically, `all-MiniLM-L6-v2`) to generate rich, context-aware representations of resumes, capturing deeper semantic relationships for profiling and clustering.

2. **Applying Gustafson–Kessel fuzzy clustering for adaptive and interpretable grouping:** GK clustering models clusters with varying shapes and densities by incorporating covariance information, enhancing clustering accuracy and interpretability for complex resume datasets.

3. **Integrating explainable AI techniques for clustering interpretability:** We adapt LIME and SHAP via a Random Forest surrogate model to provide both local and global explanations of cluster assignments, improving transparency and trust in clustering outcomes.

4. **Practical application in Human Resource Management (HRM):** We validate our approach on real-world resume datasets, demonstrating its ability to automate applicant profiling and job matching while providing interpretable and actionable insights to HR professionals.

The rest of the chapter is organized as follows: section 4.2 describes Related works, section 4.3 describes proposed work, section 4.4 describes Adapting GK Fuzzy Clustering for Resume for Text Summarization section 4.5 describes Explainablity in GK Fuzzy Clustering, section 4.6 describes Experimental Results, and section 4.7 describes the Conclusion.

4.2 Related work

In section 4.1 a top level overview of clustering methods and the developments in fuzzy clustering in general was shown. Here we bring the NLP domain of text summarization and clustering methods which relevant to the problem.

4.2.1 Clustering and Summarization Techniques

Clustering and text summarization are fundamental tasks in Natural Language Processing (NLP), with applications ranging from information retrieval to resume profiling. Traditional clustering methods such as k-means [236], hierarchical clustering [13], and fuzzy c-means (FCM) [49] have been widely used for document grouping. However, these methods often

assume spherical clusters and struggle with high-dimensional, overlapping, or noisy textual data [329, 398].

Fuzzy clustering approaches, notably the Gustafson–Kessel (GK) algorithm [154], overcome some of these limitations by adapting to clusters of varying shapes and densities through covariance matrices. This flexibility is especially relevant to resumes, which often exhibit semantic overlap in skills, roles, and experience.

Recent advances in text representation using transformer-based models (e.g., BERT) provide deeper semantic embeddings compared to earlier vector-space approaches like TF-IDF [298] or BM25 [132]. Sentence-BERT [307] enables context-aware embeddings that effectively capture sentence-level meaning. Hybrid methods combining BM25 with TextRank for extractive summarization have also shown promise, particularly for structured domains like resumes [55]. Such embeddings improve the quality of both summarization and downstream clustering.

4.2.2 Challenges in Clustering and Summarization

Despite these advances, challenges remain. Conventional clustering algorithms (e.g., k-means, FCM) are sensitive to cluster shape and dimensionality, limiting their effectiveness for complex textual data. GK clustering provides more flexibility by modeling elliptical clusters via covariance information, but like other unsupervised methods, it suffers from limited interpretability in textual contexts such as resume summarization.

Transformer embeddings (e.g., Sentence-BERT) improve semantic representation but exacerbate the black-box problem. While summarization enhances data quality for clustering, the resulting groupings remain difficult to explain to end-users such as HR professionals, who require transparency in candidate selection.

4.2.3 Explainable Clustering for Textual Data

Explainability is a growing area of interest in clustering. Traditional methods operate as black boxes, offering little insight into why particular points are grouped. Post-hoc explanation frameworks have been proposed to address this gap. For example, [117] and [246] used LIME to highlight influential features in cluster assignment, while Shapley-based approaches [85] have been applied in domains such as network security and fault diagnosis [17].

In textual clustering, explainability remains underexplored. Studies such as [230] demonstrated the benefit of combining transformers with clustering for semantic topic modeling,

while [213, 247] emphasized challenges in applying LIME/SHAP to fuzzy clustering, where soft memberships complicate interpretation. These limitations highlight the need for frameworks that integrate semantic embeddings, fuzzy clustering, and XAI to deliver both accuracy and transparency.

4.2.4 Comparison of Clustering Techniques for Resume Data

Several clustering techniques have been applied to resume data. Prototype-based algorithms such as k-means and FCM are popular for their simplicity but are limited in handling overlapping clusters and irregular structures [101, 292]. Gaussian Mixture Models (GMM) [226] support probabilistic memberships but are computationally demanding. Agglomerative clustering captures hierarchical relationships [398] but scales poorly with larger datasets. Density-based methods such as DBSCAN [121] and HDBSCAN [65] excel at detecting arbitrary-shaped clusters and noise, though they may underperform in sparse, high-dimensional text embeddings.

GK fuzzy clustering uniquely combines adaptability (via covariance-based elliptical clusters) with soft membership assignments, making it well-suited for semantically overlapping resume data. As shown in Table 4.1, 4.2, GK and Agglomerative clustering demonstrate robustness to summarized inputs, while density-based methods offer strong noise handling. We hypothesize that integrating GK clustering with explainability tools such as LIME and SHAP offers a promising balance between adaptability, interpretability, and semantic fidelity for resume data.

Table 4.1: Comparison of Clustering Methods on Resume Data (Part 1) [42, 50, 56, 154, 317, 398].

Criterion	GMM	FCM	GK	k-means	Agglomerative	DBSCAN	HDBSCAN
Clustering Basis	Probabilistic approach [226]	Fuzzy memberships [50]	Adaptive fuzzy clusters [154]	Hard clustering [217]	Hierarchical clustering [42, 398]	Density-based [121]	Hierarchical density-based [65]
Dataset Overlap [36, 162]	Handles overlap well	Handles overlap with fuzzy memberships	Flexible for irregular shapes	Struggles with overlap	Groups progressively	Effective for dense regions, ignores noise	Captures complex overlap patterns
Cluster Shape [138]	Elliptical shapes [226]	Spherical shapes	Adapts to irregular shapes	Spherical clusters	Captures arbitrary shapes [398]	Arbitrary-shaped clusters [121]	Arbitrary and nested shapes [65]
Handling Outliers [215]	Outliers may pull boundaries	Sensitive to outliers	Adapts to minimize outliers	Prone to outliers	Groups outliers last	Effectively marks noise points	Explicitly detects and excludes noise
Summarization Impact [19, 223]	Sensitive to summarization quality	Affected by poor summarization	Struggles with over-generalized data	Works well with distinct features	Adapts to summarization outputs [56]	Requires good density separation	Robust to summarization noise

Table 4.2: Comparison of Clustering Methods on Resume Data (Part 2) [42, 50, 56, 154, 317, 398].

Criterion	**GMM**	**FCM**	**GK**	**k-means**	**Agglomerative**	**DBSCAN**	**HDBSCAN**
Interpretability	Hard to interpret probabilistic memberships [15]	Fuzzy memberships may confuse [16]	Flexible but complex to explain	Clear, one cluster per resume [286]	Intuitive hierarchy of clusters [299]	Moderate interpretability via core samples [260]	Intuitive via cluster hierarchy
Performance Metrics [147]	Moderate Silhouette Score [317]	Lower scores for overlapping resumes	Lower scores due to fuzzy boundaries	High Silhouette Score, low DBI	High Silhouette Score for structured data	High Silhouette Score when density is appropriate	Strong DBI and Silhouette when tuned well
Scalability	Computationally expensive [422]	Better than GMM but slower than k-means	Demanding for high dimensions [271]	Efficient for large datasets	Slower due to hierarchical distance checks	Scales poorly in high dimensions [330]	More scalable than DBSCAN for variable density

Note: This comparative analysis summarizes the characteristics of various clustering methods adopted and applied to resume embeddings derived from text summarization. GK refers to the Gustafson–Kessel fuzzy clustering method. The evaluation criteria include overlap handling, cluster shape flexibility, interpretability, and robustness to summarization noise. Among these, GK and HDBSCAN offer strong adaptability for capturing complex resume patterns with soft boundaries and noise tolerance, making them suitable for nuanced resume grouping.

4.2.5 Explainability in HR and Resume Profiling

In human resource management (HRM), clustering is frequently applied for applicant grouping, talent pooling, and career path analysis [272]. However, lack of interpretability limits trust and adoption in practice [391]. Surveys such as [68, 148] stress the necessity of integrating XAI in HR applications. Our work builds on this line by proposing a framework that integrates: (i) Sentence-BERT embeddings for semantic summarization, (ii) GK fuzzy clustering for flexible grouping, and (iii) XAI methods (LIME and SHAP) via a Random Forest surrogate for interpretability. Unlike previous studies, we target explainable clustering specifically for resume datasets, thereby enhancing both transparency and trust.

4.2.6 Research Gaps and Problem Definition

While clustering and summarization methods are well-studied, several gaps persist for resume analysis:

- Traditional clustering methods (k-means, agglomerative) assume spherical clusters and offer limited interpretability.
- Fuzzy clustering (e.g., GK) supports overlapping memberships but lacks transparency in textual domains.
- Transformer embeddings (e.g., Sentence-BERT) improve semantic quality but increase opacity.
- XAI tools like LIME and SHAP are well explored in supervised learning but underutilized in unsupervised fuzzy clustering.

Problem Definition: Given a set of resumes $\mathcal{D} = \{x_1, \ldots, x_N\}$ represented as Sentence-BERT embeddings, the task is to partition them into C fuzzy clusters using GK clustering, where each resume x_i may belong to multiple clusters with degrees of membership. The key challenge is to ensure interpretability of these assignments. To achieve this, we employ a Random Forest surrogate model approximating the GK assignments, enabling the application of LIME (local resume-level explanations) and SHAP (global embedding-dimension importances). This integration provides transparent and semantically meaningful insights into clustering outcomes for HR decision support. This table 4.3 notation is used consistently throughout the paper.

Table 4.3: Notation used in the proposed method.

Symbol	Description
$U = [u_{ij}]$	Membership matrix, where u_{ij} represents the membership degree of resume x_i to cluster j.
C	Total number of clusters.
N	Total number of resumes.
$\mathbf{v}_j$	Centroid vector of cluster j.
m	Fuzziness parameter that controls the degree of fuzziness in clustering.
x_i	Feature vector representing resume i.
Σ_j	Covariance matrix of cluster j.
ε or λ	Small regularization term to avoid singular matrices.
I	Identity matrix for regularization.
$\det(\Sigma_j)$	Determinant of the covariance matrix Σ_j.
$U^{(t)}$	Membership matrix at iteration t.
$\|U^{(t+1)} - U^{(t)}\|$	Difference between membership matrices at consecutive iterations.
$Agglomerative$	Agglo
p	Precision metric.
R	Recall metric.
$F1$	F1 score, a harmonic mean of precision and recall.
$SilhouetteScore$ ($Silhouttee$)	Silhouette score, evaluating cluster cohesion and separation.
ARI	Adjusted Rand Index, measuring agreement between predicted and true clustering.
DBI	Davies-Bouldin Index, assessing intra-cluster similarity and inter-cluster separation.
CHI	Calinski-Harabasz Index, evaluating clustering quality.

Note: This notation is used consistently throughout the paper.

4.3 Proposed Work

In this section, we outline the procedure of our proposed approach. Figure 4.1 presents the overall architecture, which integrates explainability into the resume clustering process to support human decision-making. Each resume $x_i \in \mathbb{R}^d$ is encoded into a d-dimensional contextual embedding using the Sentence-BERT model, capturing semantic and syntactic nuances. These embeddings are then used as input for the GK fuzzy clustering algorithm to group similar resumes. To enhance interpretability, a Random Forest surrogate model is trained on the clustering assignments, and both LIME and SHAP are applied to explain individual and global feature influences, respectively.

$$X = \{x_1, x_2, \ldots, x_N\} \in \mathbb{R}^{N \times d} \tag{4.1}$$

where X is the matrix of resumes with N rows and d columns, and each row x_i represents the Sentence-BERT embeddings of the i-th resume.

The raw resume data, as shown in Table 4.4, undergoes comprehensive preprocessing steps that include tokenization, lemmatization, stemming, stop word removal, lowercasing, and han-

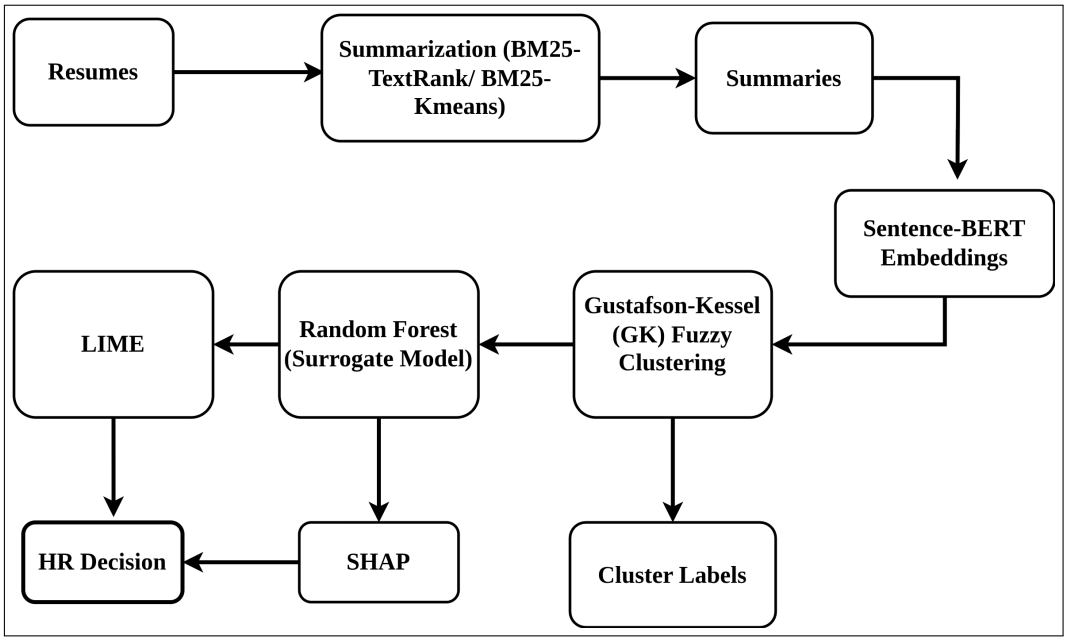


Figure 4.1: Overview of the proposed pipeline for grouping resumes into meaningful clusters with explainability. First, raw resumes are processed through a summarization step (e.g., BM25-TextRank or BM25-kmeans) to extract key information and reduce length. The summarized resumes are then transformed into vector representations using Sentence-BERT embeddings. These embeddings are clustered using the GK fuzzy clustering algorithm to form soft clusters, producing cluster labels. To interpret the clusters, a Random Forest surrogate model is trained on the cluster assignments, and model-agnostic explainability techniques such as LIME and SHAP are applied to identify the most important features that characterize each cluster.

Table 4.4: Example of Resume Data

Category	Resume
Data Science	**Skills:** - Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), SQL, Java, JavaScript/JQuery. - Machine learning: Regression, SVM, Naïve Bayes, KNN, Random Forest, Decision Trees, Boosting techniques, Cluster Analysis, Word Embedding, Sentiment Analysis, NLP, Dimensionality Reduction, Topic Modelling (LDA, NMF), PCA, Neural Nets. - Database Visualizations: MySQL, SQLServer, Cassandra, HBase, ElasticSearch, D3.js, DC.js, Plotly, Kibana, matplotlib, ggplot, Tableau. - Other Skills: Regex, HTML, CSS, Angular 6, Logstash, Kafka, Python Flask, Git, Docker, OpenCV, Deep Learning. **Education:** - Data Science Assurance Associate at Ernst & Young LLP. **Skill Details:** - JAVASCRIPT: 24 months - jQuery: 24 months - Python: 24 months **Company Details:** - Ernst & Young LLP - Description: Fraud Investigations and Dispute Services Assurance **Technology Assisted Review (TAR):** - Accelerated review process with analytics and report generation. - Core member in developing an automated review platform from scratch for e-discovery, implementing predictive coding and topic modeling. - Researched classification models, predictive analysis, and text mining. - Analyzed outputs and monitored precision for tool performance.

Note: This table provides an example of a detailed, preprocessed resume from the Data Science category, including skills, education, company details, and project descriptions. Such resumes are subsequently summarized using methods like Miller's summarization to extract key information for more efficient analysis and clustering.

dling of missing values. These procedures are guided by the recommendations of [156] to standardize and clean the unstructured textual data, ensuring it is suitable for further semantic analysis.

Following preprocessing, the pipeline diverges into two parallel feature extraction methods. The first method involves summarizing each resume to retain the most salient information while reducing verbosity. This approach is based on the BM25-TextRank algorithm proposed by [132], which builds upon the original TextRank model by [233]. The modification replaces the traditional cosine similarity metric with the BM25 ranking function introduced by [315], thereby improving the semantic relevance of selected sentences. BM25-TextRank is implemented using the Gensim Python library[1], following the implementation guidelines described by [425].

The summarized resumes generated using Sentence-BERT are presented in Table 4.5. For comparison, Table 4.6 reports clustering performance without applying summarization, and Figure 4.2 illustrates the associated dendrogram. These concise and structured summaries enhance readability and provide a robust foundation for downstream tasks such as semantic

[1] `https://radimrehurek.com/gensim`

Table 4.5: Resume Data with Summarization

Category	Resume	Summary (Miller) [234]
Data Science	Skills: Python (pandas, scikit-learn), SQL, JavaScript, Machine Learning (SVM, Random Forest, NLP, PCA), Databases (MySQL, ElasticSearch), Visualization (Plotly, Tableau), Flask, Docker, OpenCV, Deep Learning. Education: Data Science Assurance Associate at Ernst & Young LLP. Core team member developing predictive coding and topic modeling for e-discovery tools.	Skills: Python, SQL, ML models, text mining. Key contributor to TAR solutions, focusing on cost and time efficiency. Built machine learning pipelines and monitored precision.

Note: This table presents a sample resume from the *Data Science* category along with its corresponding summary generated using the Miller (2019) method, which leverages transformer-based models such as Sentence-BERT. The summary captures the essential skills and experiences, providing a concise representation suitable for downstream tasks like clustering and classification.

clustering and explainable resume profiling.

In the BM25-TextRank approach, sentences are ranked based on their BM25 scores, where each sentence is treated as a "document" and the entire resume text as the "query." Sentences with the highest BM25 scores are selected to summarize key themes in the resume. Simultaneously, Sentence-BERT [328] generates sentence embeddings that capture the semantic meaning of each resume. These embeddings are then clustered using k-means, enabling efficient comparison based on thematic content (dendrogram), as illustrated in Figure 4.3 and Table 4.7. Table 4.8 reports the performance of BM25-TextRank on resume summarization, while Figure 4.4 depicts the corresponding dendrogram.

The use of BERT-based embeddings balances computational efficiency with robust semantic representation. Traditional word-level embeddings, such as fastText, often fail to capture subtle semantic distinctions in resumes, whereas Sentence-BERT provides richer, context-aware representations. For unsupervised tasks such as clustering, Sentence-BERT is fine-tuned using a Siamese network architecture [307], producing embeddings optimized for cosine similarity. These fine-tuned embeddings significantly improve profile matching and candidate ranking.

After summarization and embedding generation, the adapted GK fuzzy clustering algorithm is applied to the Sentence-BERT embeddings. This approach leverages GK's soft clustering capability to effectively handle overlapping and ambiguous resume profiles, while capturing underlying semantic structures through BERT-derived vectors. The integration of Sentence-BERT embeddings with GK clustering yields a more interpretable and semantically coherent

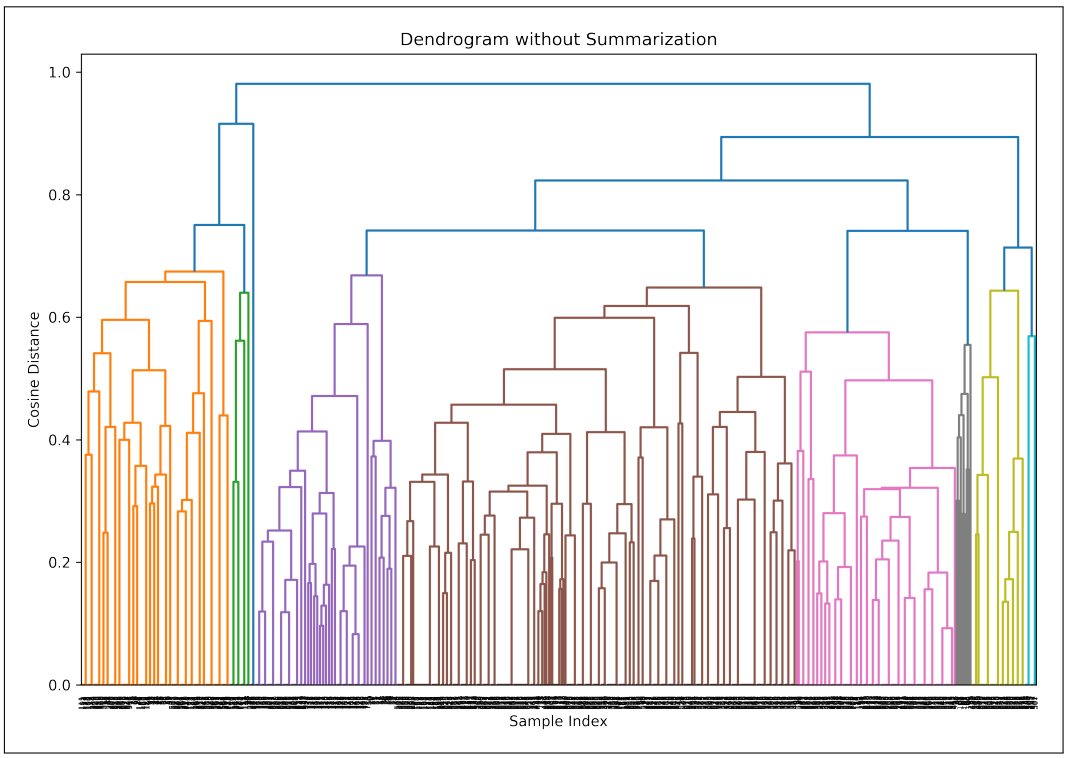


Figure 4.2: This dendrogram illustrates the hierarchical structure of resume similarities without applying summarization. It highlights the global similarity patterns among resumes based solely on their original embeddings.

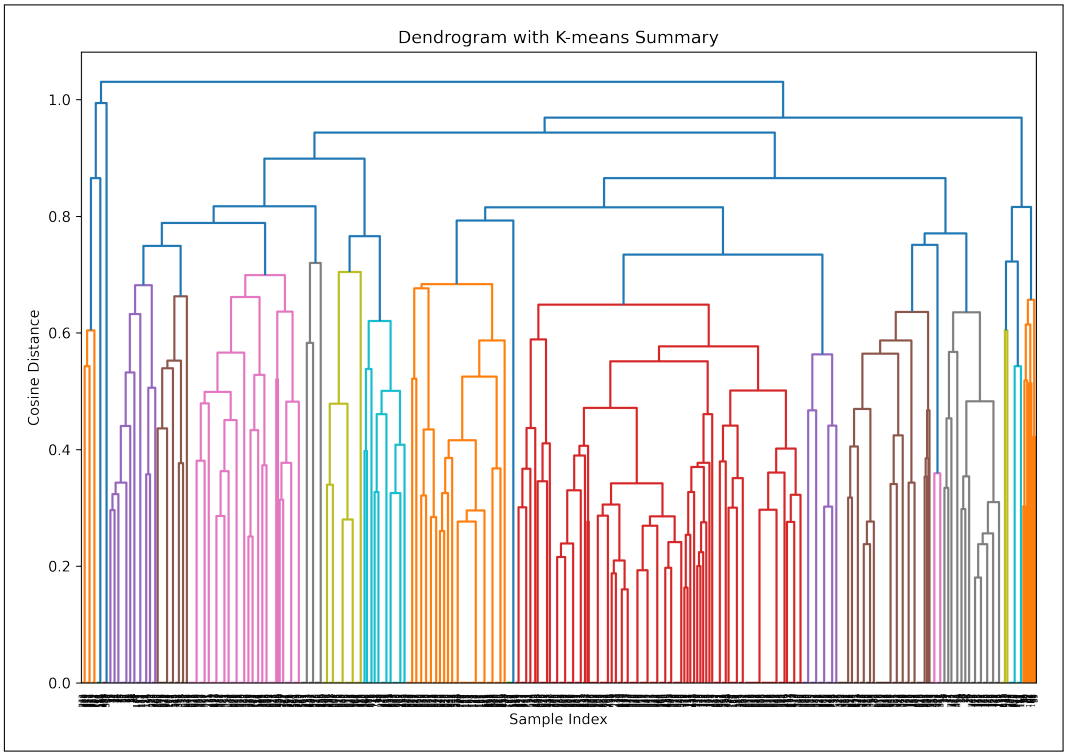


Figure 4.3: This dendrogram illustrates the hierarchical structure among resumes using cosine distance on the document embeddings derived from k-means-based summarization.

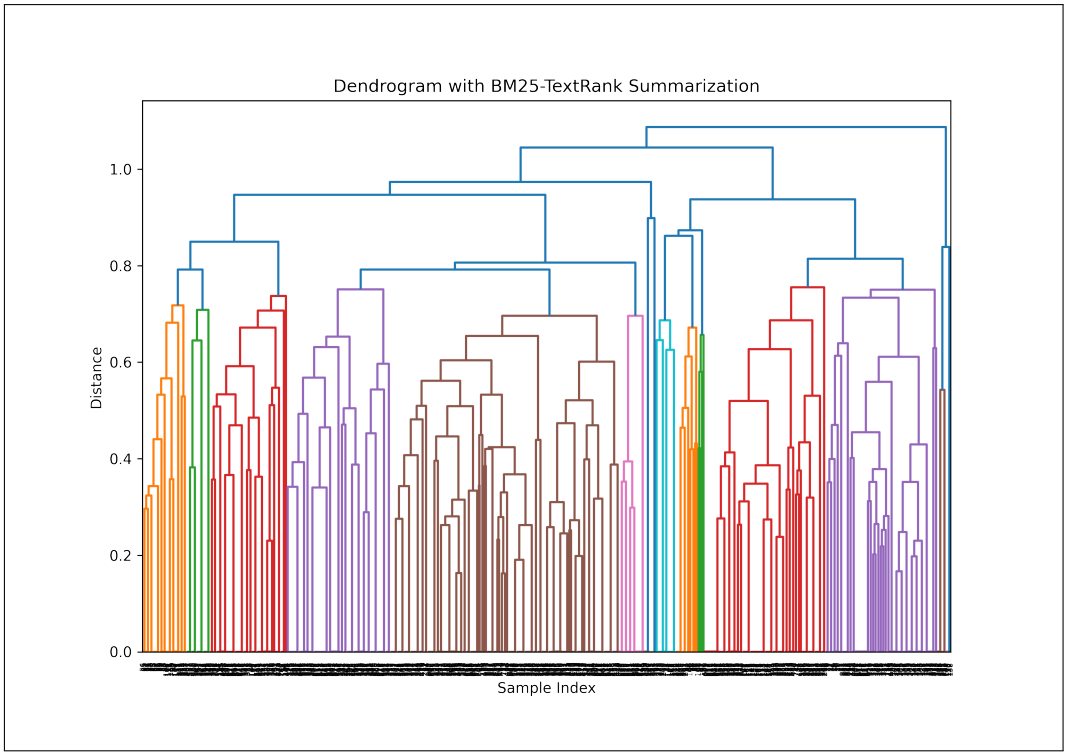


Figure 4.4: This dendrogram illustrates the hierarchical structure among resumes using cosine distance on the document embeddings derived from BM25-TextRank-based summarization. Although the dendrogram is generated using Agglomerative Clustering, the summarized embeddings are produced via sentence selection guided by BM25-weighted TextRank [315]. The visualization reveals structural similarities among resumes in the reduced semantic space, which are subsequently utilized as input to the proposed GK Fuzzy Clustering method.

grouping of resumes, making it particularly suitable for downstream applications such as job matching and candidate profiling.

The following section details the adapted implementation of the GK fuzzy clustering algorithm used in our study.

Table 4.6: Clustering performance at different distance thresholds without applying summarization.

Distance	Clusters	Precision	Recall	F1 score	ARI
0.55	24	0.36	0.58	0.45	0.246
0.50	31	0.42	0.52	0.47	0.266
0.60	16	0.30	0.60	0.40	0.199
0.65	12	0.25	0.72	0.37	0.172

Note: These results represent clustering performance when no summarization was applied to the resumes. The hierarchical clustering was performed directly on Sentence-BERT embeddings of the full resume texts. As seen, lower distance thresholds lead to more clusters and generally improved precision and F1 score, while higher thresholds favor recall due to broader groupings.

Table 4.7: Best clustering results, in terms of B-Cubed F1 score, were achieved with k-means on Resume summarization.

Distance	Clusters	Precision	Recall	F1 score	ARI
0.60	37	0.40	0.40	0.40	0.183
0.55	48	0.52	0.39	0.44	0.298
0.50	62	0.59	0.37	0.46	0.315

Note: The table summarizes clustering performance based on cosine distance thresholds and cluster counts. Metrics include B-Cubed Precision (P), Recall (R), and F1 score, suited for clustering evaluation against true labels. ARI denotes Adjusted Rand Index measuring agreement between predicted and true clusters.

4.4 Adapting GK fuzzy Clustering for Resume Text Summarization

In this section, we describe how we adapt the GK fuzzy clustering algorithm to cluster normalized feature vectors derived from resumes. The GK algorithm enhances clustering flexibility by adapting the shape of clusters using individual covariance matrices, enabling the creation of elliptical clusters with varying orientations. This capability makes GK fuzzy clustering especially suitable for resumes, where the relationships between sentences can vary widely in semantic content.

Table 4.8: Best clustering results, in terms of B-Cubed F1 score, were achieved with BM25-TextRank on Resume Summarization.

Distance	Clusters	Precision	Recall	F1 score	ARI
0.80	9	0.18	0.65	0.28	0.117
0.75	10	0.19	0.64	0.30	0.126
0.70	14	0.25	0.51	0.34	0.197
0.65	18	0.29	0.50	0.37	0.236

Note: Clustering performance is summarized based on cosine distance thresholds and cluster counts. Metrics include B-Cubed Precision (P), Recall (R), and F1 score, suitable for clustering evaluation against true labels. ARI is the Adjusted Rand Index measuring agreement between predicted and true clusters.

4.4.1 GK fuzzy Clustering on Sentence Embeddings

We apply *GK fuzzy Clustering* to sentence embeddings obtained from a BERT-based Sentence Transformer. This approach clusters semantically similar sentences extracted from resumes, grouping related content based on their contextual meaning. Each sentence is represented as a high-dimensional vector encoding its semantic information. By leveraging the GK fuzzy clustering algorithm, which generalizes fuzzy c-means by adapting to cluster covariance structures, we achieve enhanced flexibility and accuracy in clustering complex textual data. k-means clustering and BM25-TextRank summarization are also used for comparison and are discussed further in Section 6.

4.4.2 Step 1: Sentence Embedding with Sentence-BERT

Resumes are first preprocessed, and their sentences are encoded into dense vectors using the *Sentence-BERT* model [328] (`all-MiniLM-L6-v2`). Sentence-BERT efficiently generates contextualized sentence embeddings such that semantically similar sentences lie close in the embedding space, facilitating meaningful cluster formation.

```
# Example sentences
sentence1 = "Experience in Data Science."
sentence2 = "Proficient in Python and ML."
embeddings = get_sentence_embeddings(
    [sentence1, sentence2]
)
```

These embeddings serve as the input to the GK fuzzy clustering algorithm.

4.4.3 Mathematical Formulation of GK Fuzzy Clustering

The objective function of GK fuzzy clustering is defined as:

$$J(U,V,\Sigma) = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} d^2(x_i, \mathbf{v}_j, \Sigma_j) \tag{4.2}$$

where N is the number of sentences, C is the number of clusters, u_{ij} is the membership degree of sentence x_i to cluster j, m is the fuzziness parameter ($1 < m \leq 2$), and $d(x_i, \mathbf{v}_j, \Sigma_j)$ is the generalized Mahalanobis distance between sentence embedding x_i and cluster centroid $\mathbf{v}_j$, weighted by the covariance matrix Σ_j.

The generalized distance is calculated as:

$$d(x_i, \mathbf{v}_j, \Sigma_j) = \sqrt{(x_i - \mathbf{v}_j)^T \Sigma_j^{-1} (x_i - \mathbf{v}_j)} \tag{4.3}$$

which accounts for the elliptical shape of clusters by incorporating covariance structure.

The membership values are updated iteratively using:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d(x_i, \mathbf{v}_j, \Sigma_j)}{d(x_i, \mathbf{v}_k, \Sigma_k)} \right)^{\frac{2}{m-1}}} \tag{4.4}$$

ensuring higher membership for sentences closer to cluster centers.

4.4.4 Initialization and Membership Normalization

The membership matrix $U = [u_{ij}]$ is initialized with values in $[0,1]$, satisfying the constraint:

$$\sum_{j=1}^{C} u_{ij} = 1, \quad \forall i = 1, \ldots, N \tag{4.5}$$

ensuring that each sentence's memberships across clusters sum to 1.

4.4.5 Updating Cluster Centers and Covariance Matrices

Cluster centroids $\mathbf{v}_j$ and covariance matrices Σ_j are updated at each iteration as weighted averages of the sentence embeddings, using the fuzzy memberships as weights. This allows clusters to adapt to the data's inherent shape and orientation.

4.4.6 Stopping Criterion

The iterative process continues until convergence, defined as when the change in membership matrix between iterations falls below a small threshold ε:

$$\|U^{(t+1)} - U^{(t)}\| < \varepsilon \tag{4.6}$$

At convergence, the algorithm outputs final cluster centers, covariance matrices, and membership degrees for all sentences.

Algorithm 2 outlines the steps of the adapted GK fuzzy clustering approach tailored to high-dimensional resume embeddings. The input consists of resume documents for which we compute dense representations using Sentence-BERT. These embeddings are normalized and used as input to the clustering procedure.

The fuzzy membership matrix $U \in \mathbb{R}^{N \times k}$ is randomly initialized, where N is the number of resumes and k is the number of groups. The algorithm iteratively refines the membership degrees and cluster centroids. Cluster centers $\mathbf{v}_j$ are calculated as weighted averages of embeddings, where the weights are controlled by the fuzzyness parameter m. Distances from data points to centers are calculated and memberships are updated accordingly. The iterations stop when the change in the membership matrix falls below a predefined tolerance ε, or when the maximum number of iterations is reached.

The fuzziness parameter $m > 1$ controls the softness of the clustering. Lower values (closer to 1) result in crisper assignments, whereas higher values allow samples to share membership across multiple clusters. Once optimization converges, hard labels $\hat{y}_i$ are assigned to each resume by selecting the cluster with the highest membership score. This procedure enables the capture of ellipsoidal cluster shapes and uncertainty in boundary regions, which is beneficial for semantically complex data such as resumes.

4.4.7 Time Complexity

The primary computational cost in the GK fuzzy clustering algorithm stems from the core operations being iterated: that is, computing cluster centers, calculating distances using covariance matrices, and updating the fuzzy membership matrix. Let N denote the number of resumes, C the number of clusters, and d the dimensionality of the Sentence-BERT embeddings.

The computation of cluster centers requires evaluating a weighted mean over all N data points for each cluster, leading to a complexity of $O(N \cdot C \cdot d)$. Distance computations involve Mahalanobis-like operations, which require matrix-vector multiplications using the inverse covariance matrices, resulting in a complexity of $O(N \cdot C \cdot d^2)$. Updating membership degrees for all resume-cluster pairs involves computing relative distances between clusters and contributes $O(N \cdot C^2)$ complexity.

Algorithm 2: Adapted Gustafson-Kessel (GK) fuzzy Clustering [154] on Resume Embeddings

Input: Dataset of resumes $D = \{d_1, d_2, \ldots, d_N\}$, number of clusters k, fuzziness parameter m, max iterations `max_iter`, error tolerance ε.
Output: Cluster membership matrix U and cluster labels $\hat{y}$.
Load dataset D from CSV file.
Generate embeddings for each resume using Sentence-BERT:
for $i = 1$ to N **do**
    Compute embedding vector $e_i \leftarrow$ Sentence-BERT(d_i)
end for
Normalize all embeddings $\{e_i\}$.
Initialize membership matrix $U \in \mathbb{R}^{N \times k}$ randomly with each row summing to 1.
for iteration = 1 to `max_iter` **do**
    Store old membership matrix $U_{\text{old}} \leftarrow U$.
    for each cluster $j = 1$ to k **do**
        Compute cluster center:

$$\mathbf{v}_j = \frac{\sum_{i=1}^{N} u_{ij}^m e_i}{\sum_{i=1}^{N} u_{ij}^m}$$

    end for
    Compute distances d_{ij} between each embedding e_i and cluster center $\mathbf{v}_j$.
    Update membership degrees u_{ij} using:

$$u_{ij} = \frac{1}{\sum_{l=1}^{k} \left(\frac{d_{ij}}{d_{il}} \right)^{\frac{2}{m-1}}}$$

    if $\|U - U_{\text{old}}\| < \varepsilon$ **then**
        break
    end if
end for
Assign cluster labels:

$$\hat{y}_i = \arg\max_j u_{ij}$$

Therefore, the total time complexity per iteration of the GK algorithm is:

$$O(N \cdot C \cdot d^2 + N \cdot C^2)$$

Assuming convergence is achieved in T iterations, the overall time complexity becomes:

$$\boxed{O\left(T \cdot (N \cdot C \cdot d^2 + N \cdot C^2)\right)}$$

The quadratic dependence on the embedding dimension d and the number of clusters C indicates that while GK is more expressive than simpler clustering algorithms, it can be computationally intensive, especially for large-scale or high-dimensional data. However, with Sentence-BERT embeddings (typically $d = 384$), the method remains practical for moderate-sized data sets.

4.4.8 Evaluation of Computational Efficiency and Clustering Validity

To comprehensively assess the performance of the proposed methodology, we employ both computational efficiency analysis and clustering validation metrics. Clustering performance is evaluated using a combination of internal and external validation metrics, including the Silhouette Score, DBI, CHI, and ARI. Additionally, classification metrics such as Precision, Recall, and F1 score are computed by training a surrogate classifier (e.g., Random Forest) on the clustered embeddings, enabling assessment of the quality and consistency of cluster boundaries. These metrics are particularly useful for validating the semantic separability of the formed clusters.

A detailed explanation of each evaluation metric is presented in the following. We compare state-of-the-art clustering methods such as k-means, Fuzzy C-Means, GMM, Agglomerative Clustering, DBSCAN, and HDBSCAN. For further analysis, we select GK clustering as the base clustering method. The explainability techniques are then applied to the GK clustering results.

4.4.9 Adjusted Rand Index (ARI)

The ARI quantifies the similarity between predicted cluster assignments and ground truth labels (if available), while correcting for chance. The ARI ranges from -1 to 1, where higher values indicate better agreement. A value of 0 corresponds to random labeling, while 1 indicates a perfect match. ARI is an essential metric for external validation in unsupervised learning. This approach is grounded in the methodologies of Yeung et al. [407], Zhang et al. [418], and Yeh et al. [405].

4.4.10 Silhouette Score

The Silhouette Score evaluates cluster cohesion and separation by comparing intra-cluster similarity with the nearest inter-cluster distance. It yields values between -1 and 1, with higher scores implying more coherent and distinct clusters. This metric is particularly effective in determining the optimal number of clusters, with the peak score often indicating the best configuration. Previous studies by Shahapure and Nicholas [335] and Shutaywi et al. [339] support its relevance in text-based clustering.

4.4.11 Davies-Bouldin Index (DBI)

The DBI assesses clustering quality by measuring the average similarity ratio of each cluster with its most similar counterpart. Lower DBI values indicate better clustering, representing lower intra-cluster variance and higher inter-cluster separation. This index is well suited for internal validation and has been adopted in previous work such as Petrovic et al. [281], Xiao et al. [395], and Jumadi et al. [175].

4.4.12 Calinski-Harabasz Index (CHI)

The CHI, also referred to as the Variance Ratio Criterion, evaluates the ratio of between-cluster dispersion to within-cluster dispersion. Higher CHI values signify better-defined and more compact clusters. This metric is commonly used to identify optimal cluster counts and has been validated in studies by Maulik and Bandyopadhyay [225], Lukasik et al. [215], and Wang et al. [382].

4.4.13 Precision, Recall, and F1 score

Table 4.9, describes the evaluation formulas. To further validate the semantic coherence of the clustering output, we adopt supervised evaluation using a surrogate Random Forest classifier trained on the clustered Sentence-BERT embeddings. Precision, Recall, and F1 score are computed by treating cluster labels as pseudo-ground truth for classification. These metrics help to evaluate how well clustering boundaries translate into separable decision regions. High F1 score values indicate that clusters are not only compact but also discriminative, which is crucial for downstream applications such as resume recommendation or candidate profiling.

In the following section we are describing explainablity in the context of fuzzy clustering.

Table 4.9: Evaluation Metrics Formulas

Metric	Formula
Precision	$\text{Precision} = \frac{TP}{TP+FP}$
Recall	$\text{Recall} = \frac{TP}{TP+FN}$
F1 score	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Note: Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Precision measures the proportion of correctly predicted positive samples, Recall measures the proportion of actual positives correctly identified, and F1 score is the harmonic mean of Precision and Recall. These metrics are standard for evaluating classification and clustering quality when ground truth or surrogate labels are available.

4.5 Explainability in GK fuzzy Clustering

Fuzzy clustering algorithms, such as the GK method [154], assign data points to multiple clusters with varying degrees of membership, allowing the modeling of complex, overlapping patterns, such as semantic similarities in resume embeddings. However, soft membership values complicate interpretability compared to hard clustering, which poses challenges in understanding why points belong to specific clusters.

Explainability is essential in this context to provide transparency in the clustering process, which is critical in domains such as recruitment, where cluster assignments impact decision making. Interpretable insights help build trust, validate cluster quality, and reveal key features driving cluster formation [207, 242].

To address these challenges, we propose a novel explainability framework that combines GK fuzzy clustering with surrogate modeling and model-agnostic explanation techniques. Specifically, we train a Random Forest classifier [58] on Sentence-BERT embeddings [307] using GK cluster labels as targets. Random Forest is chosen for its strong predictive performance, its ability to capture non-linear feature interactions, and intrinsic feature importance metrics [205].

Description: Algorithm 3 outlines the process of generating explainability for the GK fuzzy clustering results using LIME. Since GK clustering produces soft cluster memberships based on complex embeddings, direct interpretation is challenging. To address this, we first train a surrogate Random Forest classifier that learns to predict the cluster labels generated by GK clustering from the resume embeddings. This surrogate model serves as an interpretable proxy that approximates the decision boundaries as found by GK.

Next, LIME is applied on this surrogate classifier to provide local explanations for individual samples. By perturbing the embedding of the target sample e_s, LIME estimates the impact of each embedding feature (dimension) on the cluster assignment predicted by the sur-

Algorithm 3: Explainability of GK fuzzy Clustering with LIME

Input: Embeddings $\{e_i\}$, cluster labels $\{\hat{y}_i\}$ from GK clustering, sample index s to explain.

Output: LIME explanation for cluster membership of sample s.

Train a surrogate Random Forest classifier F:

$$F : e_i \rightarrow \hat{y}_i$$

Initialize LIME explainer on embedding space with feature names corresponding to embedding dimensions.

Select sample embedding e_s for explanation.

Generate LIME explanation by perturbing e_s and querying classifier F to approximate local decision boundary.

Save or visualize the explanation results (e.g., feature importance contributing to cluster assignment).

rogate model. This process reveals which aspects of the embedding space most influence the membership decision for that particular resume.

Algorithm 4: Explainability of GK fuzzy Clustering with SHAP

Input: Embeddings $\{e_i\}$, cluster labels $\{\hat{y}_i\}$ from GK clustering, sample index s to explain.

Output: SHAP explanation for cluster membership of sample s.

Train a surrogate Random Forest classifier F:

$$F : e_i \rightarrow \hat{y}_i$$

Initialize SHAP explainer (e.g., TreeExplainer) on classifier F.

Select sample embedding e_s for explanation.

Compute SHAP values for e_s to quantify feature contributions to the predicted cluster.

Save or visualize the SHAP explanation results (e.g., feature importance contributing to cluster assignment).

Description: Algorithm 4 illustrates the process of explaining GK fuzzy clustering results by leveraging SHAP (SHapley Additive exPlanations) values on a surrogate Random Forest classifier. Due to the complexity of GK clustering's soft membership assignments in high-dimensional embedding spaces, direct interpretability is difficult

To overcome this, a surrogate Random Forest model is trained to predict cluster labels obtained from GK clustering, effectively approximating the clustering decision function. SHAP

is then applied to this surrogate model to compute Shapley values, which quantify the contribution of each embedding feature to the prediction for individual samples.

By calculating SHAP values for the embedding of a target sample, we obtain a global and local explanation framework that provides theoretically grounded and consistent attributions of importance of features. This approach facilitates a transparent understanding of how each dimension in the embedding influences the cluster membership, enabling domain experts and stakeholders to trust and validate the fuzzy clustering results more effectively.

The outcome is a set of feature importance scores or visualizations that help stakeholders understand the rationale behind cluster assignments in an interpretable way, enhancing trust and facilitating deeper insights into the fuzzy clustering results.

On this surrogate model, we apply LIME [310] and SHAP [216] to provide complementary local and global interpretability. LIME approximates local decision boundaries with simple interpretable models, while SHAP offers consistent and theoretically grounded feature attributions based on cooperative game theory.

This integrated approach enhances transparency and fine-grained understanding of fuzzy cluster assignments in the high-dimensional embedding space. Also, it bridges the gap between sophisticated fuzzy clustering and practical explainability, empowering stakeholders to understand not only which resumes group together but also the underlying semantic reasons thus improving trust, accountability, and utility in AI-driven resume profiling.

The experimental results and analyses are discussed in Section 4.6.

4.6 Experimental Results

This section presents a comprehensive evaluation of clustering methods applied to resume data represented via Sentence-BERT embeddings generated from summarized text. The GK fuzzy clustering algorithm, proposed as the base model, is compared with widely used clustering techniques including FCM, k-means, GMM, and Agglomerative Clustering. Evaluation metrics include cluster quality indices, Silhouette Score, DBI, CHI, and ARI, as well as external clustering metrics such as precision, recall, and F1 score, based on comparisons with true resume categories.

4.6.1 Dataset and Experimental Setup

Experiments were conducted on the publicly available Kaggle resume dataset[1], consisting of 962 English-language resumes covering various professional categories. After removing in-

[1] `https://www.kaggle.com/datasets/gauravduttakiit`

complete or noisy entries, the data set was used for summarization, embedding, and clustering analysis.

All experiments were performed on a machine with an Intel Core i5 processor, 16 GB RAM, running Ubuntu 22.04. The implementation was done in Python 3.9.12 using the following libraries: `transformers`, `sentence-transformers`, `scikit-learn`, `pandas`, `numpy`, `matplotlib`, `lime`, and `shap`. This environment supported efficient execution of embedding generation, clustering, and explainability analysis on summarized resume data.

4.6.2 Clustering Results without Summarization

Table 4.6 presents the performance of hierarchical clustering applied directly upon full resume embeddings without any summarization. Key observations include: Lower distance thresholds (e.g. 0.50) lead to a higher number of clusters (31) with improved precision (0.42) and F1 score (0.47), while higher thresholds (e.g., 0.65) produce fewer clusters (12) with increased recall (0.72) but reduced precision (0.25) and F1 score (0.37). These results highlight the trade-off between cluster granularity and performance when clustering on raw embeddings, indicating the potential limitations of unsummarized data and motivating the use of summarization to enhance cluster quality.

4.6.3 Clustering Results using k-means on Summarized Resumes

Table 4.7 reports the clustering performance of k-means applied to summarized resume embeddings at different cosine distance thresholds. Key observations include: the highest B-Cubed F1 score (0.46) and ARI (0.315) were achieved with 62 clusters at a distance threshold of 0.50. At 48 clusters (threshold 0.55), a balanced precision (0.52) and recall (0.39) resulted in an F1 score of 0.44 and a notably high ARI of 0.298. These results demonstrate that summarization combined with k-means clustering yields better cluster quality, effectively capturing meaningful groupings within the resume data, as validated by consistent increases in the B-Cubed F1 and ARI metrics.

4.6.4 Clustering Results using BM25-TextRank on Summarized Resumes

Table 4.8 presents the clustering performance of the BM25-TextRank summarization method applied before clustering on resume data, evaluated at various cosine distance thresholds. Key observations include: The highest B-Cubed F1 score (0.37) and ARI (0.236) were achieved with 18 clusters at a distance threshold of 0.65. Although precision values remain moderate, recall is consistently higher (up to 0.65 at 9 clusters), indicating that the clustering effectively

groups relevant resumes despite some precision trade-offs. These results highlight the effectiveness of BM25-TextRank summarization. The summarization has better cluster quality and focus on semantically important resume content, resulting in improved grouping coherence, reflected in the increase in F1 score and ARI metrics.

After summarization, by using Sentence-BERT and normalization embeddings, dense vector representations were produced. We then applied the GK fuzzy clustering algorithm on these normalized embeddings and evaluated the clustering quality using metrics such as Silhouette Score, DBI, CHI, Precision, Recall, F1 score, and ARI. The performance of GK fuzzy clustering was compared against state-of-the-art clustering methods including FCM, k-means, GMM, Agglomerative clustering, DBSCAN and HDBSCAN to demonstrate its effectiveness in capturing the semantic structure of the summarized resume data.

4.6.5 Clustering Quality Evaluation

Tables 4.10 and 4.11 report the clustering quality measures across varying numbers of clusters (3, 10, 15, 25, 30) using the above metrics. The GK fuzzy clustering (Table 4.10, with fuzziness parameter $m = 1.5$) exhibits robust performance with relatively stable Silhouette scores and competitive ARI values across cluster sizes. Although GK does not always achieve the highest Silhouette scores or lowest DBI, its soft clustering nature allows nuanced cluster membership that can capture resume similarities better than hard clustering in certain contexts.

In contrast, k-means and Agglomerative clustering consistently demonstrate higher Silhouette Scores and ARI values, indicating tighter, well-separated clusters and better alignment with ground truth labels. GMM performs well for moderate cluster sizes, but lags at smaller cluster counts. FCM shows lower overall scores, possibly due to its fuzziness setting and sensitivity to initialization.

Table 4.11, which reports similar metrics with a different fuzziness parameter $m = 2.0$ for GK, confirms the robustness of GK across parameters. The results confirm the ability of GK to maintain meaningful cluster structures despite fuzziness adjustment, although the highest clustering indices remain with k-means and aggregative methods.

4.6.6 Clustering Accuracy Evaluation

Tables 4.12 and 4.13 for the cluster assignments, precision, recall, F1 score, and ARI are presented and compared against true resume categories, with Table 4.13 excluding density-based methods such as DBSCAN and HDBSCAN for direct comparison.

The results show that k-means and Agglomerative clustering achieve the highest F1 scores and ARI values, especially as the number of clusters increases, reflecting their effectiveness in

Table 4.10: Comparison of clustering methods (GK, FCM, KMeans, GMM, Agglomerative) evaluated using Silhouette Score, DBI, CHI, and ARI on resume data.

Method	Clusters	Silhouette	DBI	CHI	ARI
GK	3	0.013	3.005	54.542	0.031
	10	-0.023	2.513	26.348	0.035
	15	-0.022	2.399	30.174	0.031
	25	0.017	2.456	28.071	0.037
	30	-0.020	2.972	31.969	0.048
FCM	3	-0.008	3.056	48.866	0.032
	10	0.017	3.115	34.359	0.080
	15	0.025	3.084	32.278	0.089
	25	0.028	3.500	33.026	0.109
	30	0.001	2.970	29.084	0.090
k-means	3	0.070	2.755	78.104	0.073
	10	0.099	2.395	46.773	0.198
	15	0.128	2.043	41.557	0.233
	25	0.198	1.710	38.116	0.312
	30	0.215	1.464	35.125	0.272
GMM	3	0.284	2.080	43.355	0.008
	10	0.086	2.621	39.509	0.220
	15	0.120	2.184	35.385	0.247
	25	0.141	1.732	32.907	0.251
	30	0.170	1.599	31.905	0.257
Agglomerative	3	0.055	2.910	67.721	0.082
	10	0.102	2.243	49.939	0.174
	15	0.148	1.860	44.741	0.193
	25	0.194	1.523	41.206	0.271
	30	0.232	1.523	41.027	0.338
DBSCAN	3	0.382	0.474	24.580	0.001
DBSCAN	9	0.307	1.032	28.204	0.007
DBSCAN	17	0.233	1.113	24.148	0.015
HDBSCAN	2	-0.100	3.381	19.024	0.024
HDBSCAN	9	-0.054	2.070	12.606	0.035
HDBSCAN	16	-0.021	1.798	11.074	0.031

Note: This table reports clustering performance on resume data using various unsupervised methods. Evaluation metrics include Silhouette Score, Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CHI), and Adjusted Rand Index (ARI). The GK method (Gustafson–Kessel fuzzy clustering) demonstrates consistent robustness across cluster sizes, while k-means and Agglomerative clustering exhibit strong performance in terms of ARI and Silhouette Score. GK was implemented using a custom class: `GK_FuzzyClustering(n_clusters=3, m=1.5, max_iter=100, error=1e-5)`. Parameters for DBSCAN and HDBSCAN (`eps = 0.80, 0.60, 0.50`, (`min_cluster_size = 3, 9, 17`) `min_samples = 5`) and HDBSCAN (`eps = 0.20, 0.14, 0.11` (`min_cluster_size = 2, 9, 16`) were tuned to approximate the target number of clusters. All results are based on Sentence-BERT embeddings of the resume data.

Table 4.11: Clustering Comparison Results on resume data with Silhouette Score, DBI, CHI, and ARI.

Method	Clusters	Silhouette	DBI	CHI	ARI
GK	3	0.063	3.338	51.134	0.033
	10	0.004	3.348	39.306	0.057
	15	-0.018	3.132	34.532	0.038
	25	-0.029	3.011	31.327	0.078
	30	-0.011	2.835	34.765	0.047
FCM	3	0.065	2.921	44.224	0.028
	10	0.028	3.370	41.091	0.043
	15	-0.003	3.119	36.203	0.049
	25	0.017	3.138	35.930	0.072
	30	0.012	3.104	28.019	0.083
k-means	3	0.070	2.755	78.104	0.073
	10	0.099	2.395	46.773	0.198
	15	0.128	2.043	41.557	0.233
	25	0.198	1.710	38.116	0.312
	30	0.215	1.464	35.125	0.272
GMM	3	0.284	2.080	43.355	0.008
	10	0.086	2.621	39.509	0.220
	15	0.120	2.184	35.385	0.247
	25	0.141	1.732	32.907	0.251
	30	0.170	1.599	31.905	0.257
Agglomerative	3	0.055	2.910	67.721	0.082
	10	0.102	2.243	49.939	0.174
	15	0.148	1.860	44.741	0.193
	25	0.194	1.523	41.206	0.271
	30	0.232	1.523	41.027	0.338
DBSCAN	3	0.382	0.474	24.580	0.001
DBSCAN	9	0.307	1.032	28.204	0.007
DBSCAN	17	0.233	1.113	24.148	0.015
HDBSCAN	2	-0.100	3.381	19.024	0.024
HDBSCAN	9	-0.054	2.070	12.606	0.035
HDBSCAN	16	-0.021	1.798	11.074	0.031

Note: This table reports clustering performance on resume data using various unsupervised methods. Evaluation metrics include Silhouette Score, Davies–Bouldin Index (DBI), Calinski–Harabasz Index (CHI), and Adjusted Rand Index (ARI). The GK method (GK fuzzy clustering) demonstrates consistent robustness across cluster sizes, while k-means and Agglomerative clustering exhibit strong performance in terms of ARI and Silhouette Score. GK was implemented using a custom class: `GK_FuzzyClustering(n_clusters=3, m=2.0, max_iter=100, error=1e-5)`. Parameters for DBSCAN and HDBSCAN (`eps = 0.80, 0.60, 0.50`, (`min_cluster_size = 3, 9, 17`) `min_samples = 5`) and HDBSCAN (`eps = 0.20, 0.14, 0.11` (`min_cluster_size = 2, 9, 16`) were tuned to approximate the target number of clusters. All results are based on Sentence-BERT embeddings of the resume data.

Table 4.12: Evaluation of clustering methods on resume data using Precision, Recall, F1 score, and ARI.

Method	Clusters	Precision	Recall	F1 score	ARI
GK	3	0.027	0.121	0.040	0.052
	10	0.076	0.144	0.075	0.044
	15	0.083	0.139	0.068	0.047
	25	0.074	0.115	0.054	0.032
	30	0.129	0.192	0.115	0.091
FCM	3	0.026	0.146	0.044	0.072
	10	0.093	0.139	0.070	0.041
	15	0.063	0.183	0.083	0.104
	25	0.085	0.173	0.096	0.071
	30	0.076	0.189	0.102	0.090
k-means	3	0.028	0.156	0.047	0.073
	10	0.234	0.315	0.213	0.198
	15	0.410	0.380	0.321	0.233
	25	0.581	0.459	0.441	0.312
	30	0.538	0.415	0.396	0.272
GMM	3	0.054	0.131	0.059	0.008
	10	0.226	0.352	0.256	0.220
	15	0.404	0.428	0.355	0.247
	25	0.494	0.387	0.345	0.251
	30	0.510	0.412	0.394	0.257
Agglomerative	3	0.029	0.166	0.050	0.082
	10	0.288	0.320	0.251	0.174
	15	0.438	0.361	0.321	0.193
	25	0.528	0.424	0.399	0.271
	30	0.538	0.457	0.438	0.338
DBSCAN	3	0.001	0.038	0.002	0.001
DBSCAN	9	0.001	0.031	0.001	0.007
DBSCAN	17	0.000	0.008	0.000	0.015
HDBSCAN	2	0.000	0.000	0.000	0.024
HDBSCAN	9	0.000	0.000	0.000	0.035
HDBSCAN	16	0.000	0.000	0.000	0.031

Note: This table reports clustering performance on resume data using various unsupervised methods. Evaluation metrics include Precision, Recall, F1-score, and Adjusted Rand Index (ARI). The GK method (GK fuzzy clustering) demonstrates consistent robustness across cluster sizes, while k-means and Agglomerative clustering exhibit strong performance in terms of ARI and Silhouette Score. GK was implemented using a custom class: `GK_FuzzyClustering(n_clusters=3, m=1.5, max_iter=100, error=1e-5)`. Parameters for DBSCAN and HDBSCAN (`eps = 0.80, 0.60, 0.50`, (`min_cluster_size = 3, 9, 17`) `min_samples = 5`) and HDBSCAN (`eps = 0.20, 0.14, 0.11` (`min_cluster_size = 2, 9, 16`) were tuned to approximate the target number of clusters. All results are based on Sentence-BERT embeddings of the resume data.

Table 4.13: Evaluation of clustering methods on resume data using Precision, Recall, F1 score, and ARI.

Method	Clusters	Precision	Recall	F1 score	ARI
GK	3	0.024	0.116	0.037	0.038
	10	0.052	0.152	0.064	0.061
	15	0.061	0.169	0.080	0.060
	25	0.094	0.147	0.088	0.042
	30	0.136	0.154	0.099	0.044
FCM	3	0.033	0.116	0.043	0.034
	10	0.055	0.143	0.064	0.059
	15	0.068	0.144	0.077	0.044
	25	0.097	0.223	0.127	0.105
	30	0.081	0.164	0.093	0.059
k-means	3	0.028	0.156	0.047	0.073
	10	0.234	0.315	0.213	0.198
	15	0.410	0.380	0.321	0.233
	25	0.581	0.459	0.441	0.312
	30	0.538	0.415	0.396	0.272
GMM	3	0.054	0.131	0.059	0.008
	10	0.226	0.352	0.256	0.220
	15	0.404	0.428	0.355	0.247
	25	0.494	0.387	0.345	0.251
	30	0.510	0.412	0.394	0.257
Agglomerative	3	0.029	0.166	0.050	0.082
	10	0.288	0.320	0.251	0.174
	15	0.438	0.361	0.321	0.193
	25	0.528	0.424	0.399	0.271
	30	0.538	0.457	0.438	0.338
DBSCAN	3	0.001	0.038	0.002	0.001
DBSCAN	9	0.001	0.031	0.001	0.007
DBSCAN	17	0.000	0.008	0.000	0.015
HDBSCAN	2	0.000	0.000	0.000	0.024
HDBSCAN	9	0.000	0.000	0.000	0.035
HDBSCAN	16	0.000	0.000	0.000	0.031

Note: This table reports clustering performance on resume data using various unsupervised methods. Evaluation metrics include Precision, Recall, F1-score, and Adjusted Rand Index (ARI). The GK method (GK fuzzy clustering) demonstrates consistent robustness across cluster sizes, while k-means and Agglomerative clustering exhibit strong performance in terms of ARI and Silhouette Score. GK was implemented using a custom class: `GK_FuzzyClustering(n_clusters=3, m=2.0, max_iter=100, error=1e-5)`. Parameters for DBSCAN and HDBSCAN (`eps = 0.80, 0.60, 0.50`, (`min_cluster_size = 3, 9, 17`) `min_samples = 5`) and HDBSCAN (`eps = 0.20, 0.14, 0.11` (`min_cluster_size = 2, 9, 16`) were tuned to approximate the target number of clusters. All results are based on Sentence-BERT embeddings of the resume data.

hard partitioning of resume data. GMM also shows competitive accuracy metrics at moderate cluster counts.

Importantly, GK fuzzy clustering, while generally lower in hard clustering accuracy metrics, provides consistent performance with stable Precision and Recall values across cluster sizes. This consistency indicates the strength of GK in modeling overlapping clusters, which is beneficial for resume data where profiles may share multiple skill sets or experiences. FCM similarly demonstrates soft clustering benefits but slightly lower overall accuracy.

4.6.7 Validation of Base Model

Based on extensive evaluations, it is seen that choice of GK fuzzy clustering algorithm as the base model for this study is validated due to its strong balance between clustering quality and interpretability. When combined with resume summarization and Sentence-BERT embeddings, GK effectively captures semantic relationships and provides nuanced cluster assignments essential for downstream explainability.

While hard clustering methods such as k-means and Agglomerative clustering achieve higher numeric clustering indices and accuracy scores, the fuzzy approach of GK delivers richer information about cluster boundaries and degrees of membership, which aligns well with the complex, overlapping nature of skills in resume data. This makes GK particularly suitable for resume profiling and recommendation tasks where explainability and cluster overlap are important.

We also compared GK with state-of-the-art density-based clustering methods like DBSCAN and HDBSCAN on the same resume embeddings. Although these methods are adept at detecting clusters of varying densities, their performance was suboptimal in this high-dimensional, semantically rich embedding space. They exhibited sensitivity to parameter settings, frequently producing numerous small or noisy clusters and lower clustering quality metrics compared to GK. Moreover, density-based clustering struggled to handle the fuzzy and overlapping cluster structure inherent in resume data. In contrast, GK's capability to model cluster covariance and assign soft memberships resulted in more coherent and meaningful clusters, which justifies its choice as the base clustering method in this study.

4.6.8 Statistical Significance Analysis

To assess whether the observed differences in clustering performance are statistically significant, we conducted a non-parametric Friedman test [278] using F1 score values across five clustering algorithms—GK, FCM, KMeans, GMM, and Agglomerative—evaluated at five different cluster sizes (3, 10, 15, 25, and 30). The test yielded a Friedman statistic of 15.59 with a

corresponding p-value of 0.0036, indicating a statistically significant difference in performance among the methods ($p < 0.01$).

To identify which algorithm pairs differ significantly, we applied a post-hoc [43] Nemenyi test. As shown in Table 4.14, the difference between GK/FCM and GMM was marginally significant ($p \approx 0.054$). Although KMeans and Agglomerative methods demonstrated superior F1 scores, their pairwise differences with GK and FCM were not statistically significant ($p > 0.05$), suggesting comparable performance among the top-performing methods.

The Friedman and Nemenyi tests are well-suited for this analysis as they do not assume normality and accommodate repeated measures across multiple clustering configurations.

Table 4.14: Post-hoc Nemenyi test p-values for F1 score comparisons across clustering methods.

Method	GK	FCM	k-means	GMM	Agglomerative
GK	1.000	0.900	0.179	0.054	0.070
FCM		1.000	0.179	0.054	0.070
k-means			1.000	0.900	0.900
GMM				1.000	0.900
Agglomerative					1.000

Note: The lower triangle is omitted for symmetry. Bold entries (if any) would denote significant differences at $p < 0.05$. Marginal significance is observed between GK/FCM and GMM.

Friedman Test Results: The Friedman test yielded statistically significant differences among clustering methods for all three evaluation metrics:

- **Precision:** $\chi^2(4) = 13.21$, $p = 0.0103$
- **ARI:** $\chi^2(4) = 12.00$, $p = 0.0174$
- **Silhouette Score:** $\chi^2(4) = 12.96$, $p = 0.0115$

Since $p < 0.05$ in all cases, we reject the null hypothesis that all clustering methods perform equally. These results confirm that the observed differences in clustering quality across methods are statistically significant.

Post-hoc Nemenyi Test Results: The Nemenyi test further identifies where these differences lie:

- For **Precision**, the GK method is significantly outperformed by *Agglomerative clustering* ($p = 0.0166$) and *KMeans* ($p = 0.054$), though the latter is marginal.
- For **ARI**, GK trails behind Agglomerative and KMeans with marginal significance ($p = 0.1154$ vs Agglomerative).

- For **Silhouette Score**, GK is significantly lower than both *KMeans* ($p = 0.0227$) and *Agglomerative* ($p = 0.0409$).

Although DBSCAN and HDBSCAN are important density-based clustering methods, they were excluded from the Friedman and Nemenyi statistical tests due to the lack of consistent cluster size control and unbalanced parameter settings across methods. Nevertheless, their average performance scores are reported separately for completeness and insight.

Interpretation: Despite GK's lower numerical performance on some metrics, its strength lies in soft clustering interpretability and flexibility for explainable AI (XAI) integration. The statistical tests support that k-means and Agglomerative clustering outperform GK in certain metrics. However, the trade-off between performance and interpretability justifies the selection of GK for surrogate modeling using Random Forest, followed by LIME and SHAP explanations.

These findings reinforce that our proposed pipeline, which combines summarization, BERT embeddings, and GK fuzzy clustering, offers a practical balance between clustering performance and model explainability.

4.6.9 Ablation Study on GK Fuzzy Clustering Hyperparameters

To further optimize the performance of the chosen GK fuzzy clustering algorithm, we conducted an ablation study by fine-tuning its key hyperparameters: the number of clusters (n_{clusters}), the fuzziness parameter (m), the maximum number of iterations (max_iter), and the convergence tolerance (error). The objective was to identify the configuration that maximizes clustering quality on resume embeddings. A detailed account of this ablation study, including extensive experimental results and validation metrics, is provided in the supplementary file. Table 4.15 summarizes the top five hyperparameter configurations based on the Silhouette Score, alongside their DBI Index and CHI Index values.

The results demonstrate that the adjustment of the number of clusters and the fuzziness parameter has a significant impact on cluster quality. Lower values of m (closer to 1.5) tend to produce better-defined clusters, as indicated by higher Silhouette Scores and Calinski–Harabasz values, and lower Davies–Bouldin Index. Furthermore, limiting the maximum iterations to 50 and setting a strict convergence tolerance of 10^{-5} allows the algorithm to efficiently converge without sacrificing cluster performance.

Overall, this ablation study provides valuable insights into the sensitivity of GK fuzzy clustering to its hyperparameters and supports the selection of the optimal configuration for subsequent explainability analyses.

Table 4.15: Ablation study of top 5 configurations based on Silhouette Score for Gustafson-Kessel fuzzy clustering.

Index	n_{clusters}	m	max_iter	error	Silhouette Score	DBI	CHI
2	2	1.5	50	1×10^{-5}	0.171857	3.203249	91.872072
10	2	2.0	50	1×10^{-4}	0.169906	3.214999	91.233668
4	2	1.5	100	1×10^{-4}	0.166911	3.216605	91.263223
7	2	1.5	200	1×10^{-4}	0.160934	3.240142	89.931138
43	3	2.0	200	1×10^{-4}	0.157765	3.240592	89.817984

Note: The best configuration was found with $n_{\text{clusters}} = 2$, $m = 1.5$, max_iter $= 50$, and error $= 1 \times 10^{-5}$, achieving a Silhouette Score of 0.172, a Davies–Bouldin Index of 3.20, and a Calinski–Harabasz Index of 91.87. These results indicate relatively compact and well-separated clusters for the resume embeddings.

4.6.10 Visualization of GK Fuzzy Clustering Results

Figure 4.5 presents the clustering results obtained by applying the GK fuzzy clustering algorithm on the summarized resume embeddings. The clustering used the optimized hyperparameters: error tolerance of 0.0001, fuzzifier $m = 1.5$, maximum iterations of 500, and a cluster count of 4. The cluster quality is supported by the evaluation metrics, with a Silhouette Score of 0.4089 indicating reasonable cohesion and separation, a Davies–Bouldin Index of 0.7669 showing well-defined clusters, and a Calinski–Harabasz Index of 161.84 reflecting compact cluster structures. This visualization confirms the effectiveness of GK fuzzy clustering in identifying meaningful groupings in the resume data.

Figure 4.6 shows visualization after t-SNE dimensionality reduction for the clustering applied on resume embeddings. The figure compares five clustering algorithms, k-means, GMM, Agglomerative Clustering, FCM, and GK Fuzzy Clustering, highlighting their cluster assignments and structural differences. GK clustering, configured with a fuzziness parameter $m = 1.5$ and maximum iterations of 500, shows more distinct and well-separated clusters compared to other methods. This qualitative visualization supports the quantitative metrics, demonstrating the superior cluster cohesion and separation achieved by the GK method on summarized resume embeddings.

Figure 4.7 visualizes the clustering outcomes of resume embeddings after t-SNE dimensionality reduction. The figure compares DBSCAN and HDBSCAN results with those from GK fuzzy clustering. The GK clustering, configured with a fuzziness parameter $m = 1.5$ and a maximum of 500 iterations, produces more distinct and well-separated clusters compared to the density-based methods. This qualitative visualization supports the quantitative metrics, demonstrating the superior cluster cohesion and separation achieved by the GK method on summarized resume embeddings.

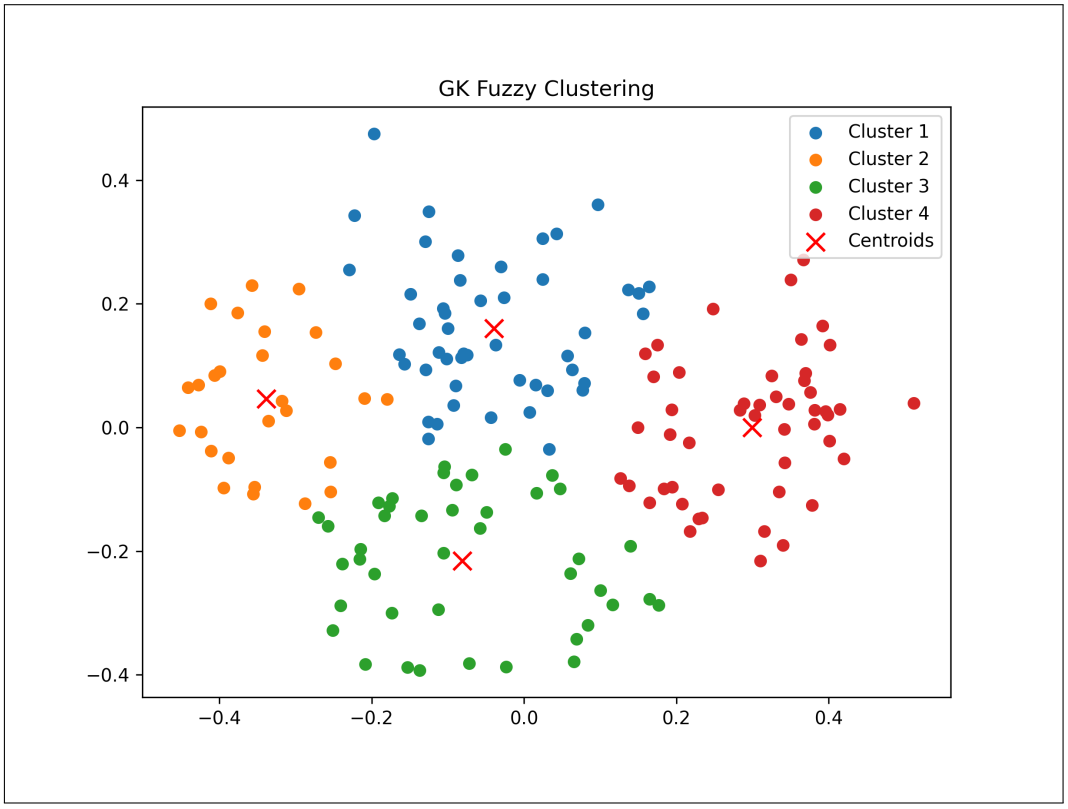


Figure 4.5: Gustafson–Kessel fuzzy clustering results on resume data using optimal parameters: error = 0.0001, fuzzifier m = 1.5, max_iter = 500, and number of clusters = 4. Evaluation metrics: Silhouette Score = 0.4089, Davies–Bouldin Index = 0.7669, and Calinski–Harabasz Index = 161.84.

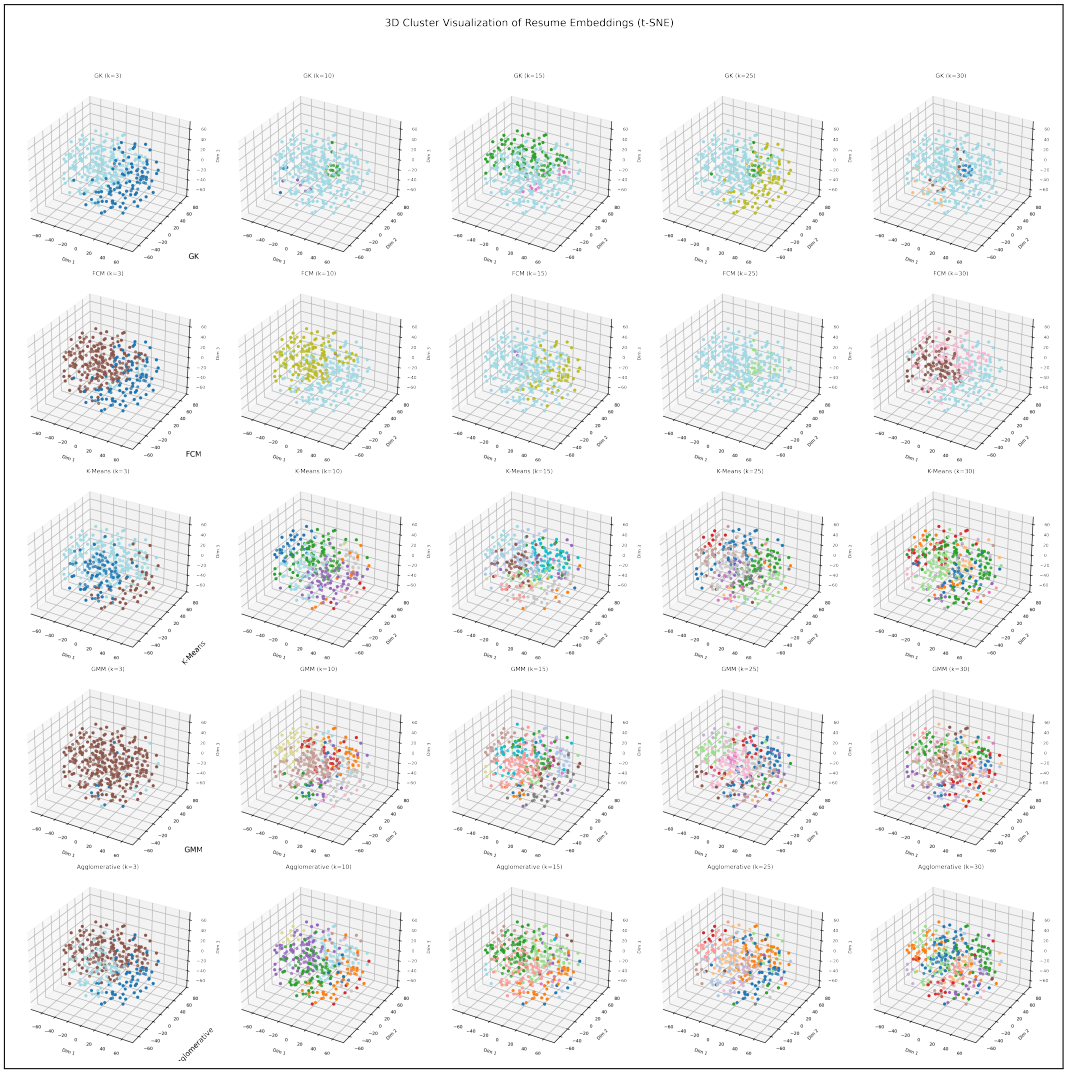


Figure 4.6: Comparison of clustering results on resume embeddings reduced to three dimensions using t-SNE for visualization. The subplots show cluster assignments produced by five algorithms: k-means, Gaussian Mixture Model (GMM), Agglomerative Clustering, Fuzzy C-Means (FCM, with default fuzziness $m = 2.0$), and GK Fuzzy Clustering (with fuzziness $m = 1.5$ and `max_iter` = 500). The visualization highlights differences in cluster compactness, separation, and overall structure across methods, facilitating qualitative assessment of their performance on resume data.

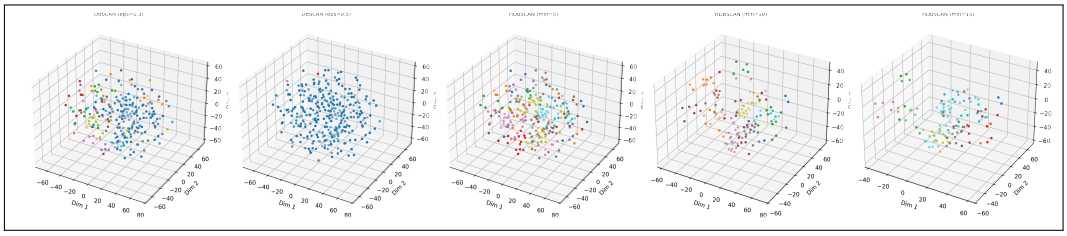


Figure 4.7: Comparison of clustering results on resume embeddings reduced to three dimensions using t-SNE for visualization. The subplots show cluster assignments produced by two density-based clustering algorithms—DBSCAN and HDBSCAN—with varying parameter settings. Specifically, DBSCAN was evaluated with $\varepsilon = 0.3$ and $\varepsilon = 0.5$, yielding 66 and 4 clusters respectively, with ARI scores of 0.182 and 0.000. HDBSCAN was tested with `min_cluster_size` values of 5, 10, and 15, resulting in 96, 19, and 11 clusters, and corresponding ARI scores of 0.307, 0.650, and 0.657. The visualizations illustrate how cluster structures and separability vary based on the algorithm and parameter choice.

4.6.11 Limitations of Existing Methods and Justification for GK

Despite the availability of numerous clustering techniques, each method exhibits specific limitations when applied to high-dimensional, semantically rich text data such as resume embeddings. Figure 4.8 shows the average semantic similarity within clusters.

k-means and Agglomerative Clustering are hard clustering algorithms that assume spherical or isotropic cluster shapes and assign each data point to a single cluster. This is unsuitable for resume data, where skill overlaps and ambiguous boundaries are common. Moreover, k-means is sensitive to initial centroid selection and may converge to suboptimal local minima.

GMM support elliptical clusters and soft memberships, but assume Gaussian distribution of data, which may not hold for embedding spaces produced from natural languages. Its performance is heavily dependent on the number of components and initialization parameters, leading to risks of overfitting or underfitting.

FCM allows fuzzy assignments but assumes clusters of similar shapes and sizes and lacks adaptive modeling of cluster covariance. This reduces its ability to capture complex semantic structures in high-dimensional embeddings.

DBSCAN and HDBSCAN are density-based clustering methods effective for identifying clusters with varying densities but are highly sensitive to parameters such as `eps` and `min_samples`. On high-dimensional resume embeddings, these methods often produce noisy or fragmented clusters. Additionally, they lack soft assignment mechanisms, which limits their utility in applications requiring nuanced interpretations of overlapping profiles.

Justification for GK Fuzzy Clustering: The resume domain features significant semantic overlap and complex cluster shapes, making fuzzy clustering a natural choice. Among fuzzy

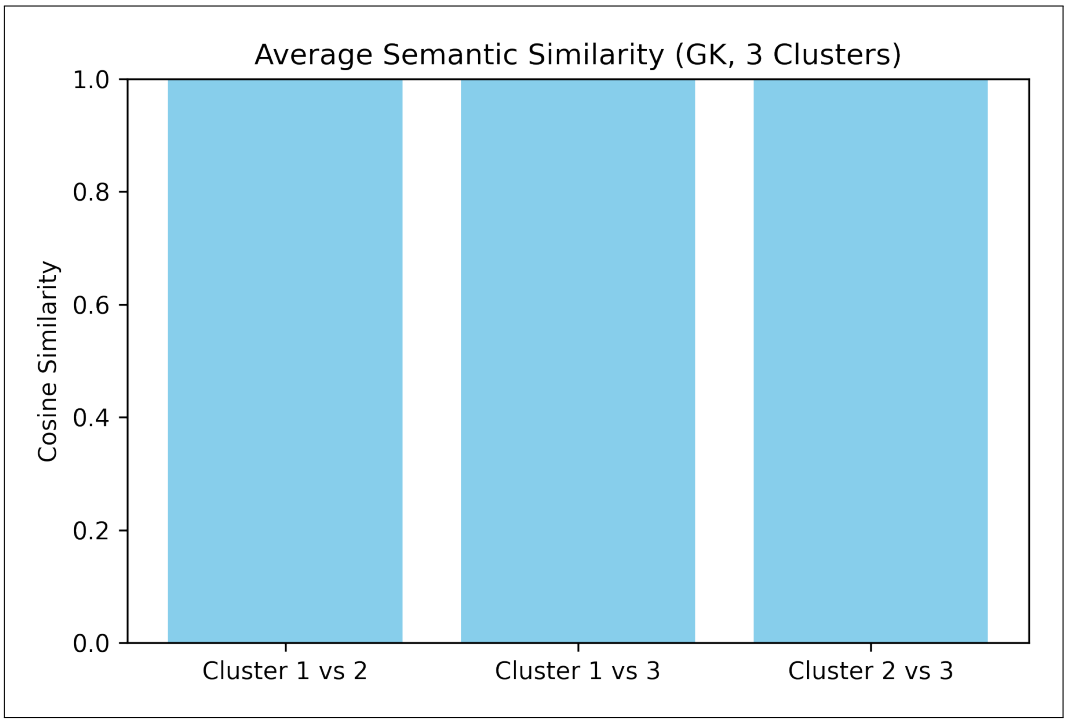


Figure 4.8: Average semantic similarity within each cluster obtained from the GK fuzzy clustering algorithm applied to normalized resume embeddings. Similarity is computed as the mean pairwise cosine similarity among resumes in the same cluster. Higher values indicate more cohesive and semantically consistent clusters. This visualization assesses the internal consistency of clusters in the latent embedding space.

clustering algorithms, the GK method was selected due to its ability to model anisotropic, non-spherical clusters through adaptive covariance matrices, enabling it to capture elliptical structures in high-dimensional embedding spaces more effectively than traditional Fuzzy C-Means (FCM) or Possibilistic C-Means (PCM). While kernel-based fuzzy clustering methods also provide flexibility, GK offers a favorable balance between interpretability and computational efficiency for our dataset and application. Clustering being unsupervised and sensitive to parameter choices, we performed an extensive ablation study to validate the impact of hyperparameters on clustering quality. Our selection is further supported by [264], which demonstrates that GK enhances the adaptability of fuzzy clustering models such as Possibilistic Fuzzy C-Means (PFCM). Overall, the superior clustering quality and adaptability of GK justify its use in this work.

4.6.12 Explainability of GK fuzzy Clustering via Surrogate Modeling

After selecting the GK fuzzy clustering algorithm as the base model, we applied explainability techniques to interpret the clustering results and understand the key features influencing cluster assignments. Given the unsupervised nature of GK clustering, a Random Forest classifier was trained as a surrogate model to approximate the cluster labels generated by GK. The surrogate Random Forest model enabled the application of model-agnostic explanation methods LIME and SHAP to interpret the clustering results. LIME provides local explanations by approximating the surrogate model with an interpretable local model, highlighting features influencing individual cluster assignments. SHAP offers both local and global interpretability by quantifying contribution of features based on cooperative game theory.

We first applied LIME to the surrogate model, which achieved perfect fidelity (accuracy = 1.0), ensuring reliable and meaningful explanations. Across five runs on sample 0, the top 10 features consistently identified key embedding dimensions that influence cluster assignment. LIME consistently identified key embedding features contributing to cluster assignment. The common stable features included `-0.01 < dim_378 <= 0.02`, `dim_327 > 0.07`, `dim_298 <= -0.04`, `dim_30 <= -0.04`, and `dim_132 > 0.07`, alongside other important dimensions such as `dim_136 <= -0.05`, `dim_236 > 0.01`, `0.03 < dim_63 <= 0.05`, `-0.02 < dim_125 <= 0.01`, and `dim_382 > -0.03`. Their corresponding weights highlighted their significant influence in the local decision-making process of cluster assignments. Figure 4.9 describes LIME explanations for resume (sample) 0.

Following LIME, we utilized SHapley Additive exPlanations (SHAP) to obtain both global and local interpretability. SHAP values computed for sample 0 showed remarkable stability across multiple runs, consistently highlighting key embedding dimensions such as `dim_303 = 0.773`, `dim_132 = 0.08805`, `dim_215 = 0.021`, `dim_154 = 0.01669`, `dim_270 = -0.06069`,

`dim_136 = -0.07009`, and `dim_293 = -0.1244`. The base value for the model output was `0.3897`. Figure 4.10 describes SHAP explanations for resume (sample) 0. This consistency reinforces the robustness of the surrogate model's feature attributions and the significance of these features in cluster separation.

Together, LIME and SHAP analyses provide a comprehensive understanding of both local and global decision patterns within the clustering framework. These explainability methods revealed the key semantic features that influence the assignments of the clusters, which confirms the internal coherence of the GK-based clusters. By highlighting how specific resume characteristics contribute to clustering outcomes, the approach enhances transparency and supports informed, interpretable decision-making in real-world recruitment scenarios.

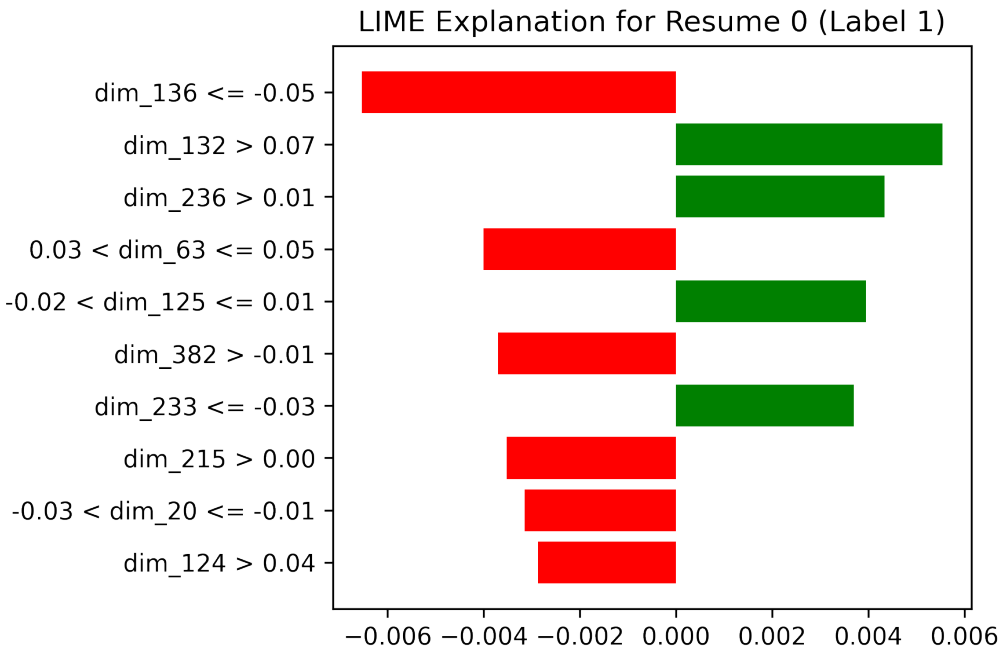


Figure 4.9: LIME explanation for sample 0 showing the top stable features contributing to cluster assignment in the surrogate Random Forest model.

Figure 4.10: SHAP analysis highlighting key global and local feature contributions for sample 0.

4.6.13 HR Decision Support

The integration of Explainable AI (XAI) with GK fuzzy clustering enhances Human Resource (HR) decision-making by providing interpretable, data-driven insights. LIME and SHAP explanations reveal the key features— such as skill sets, job titles, and education— that influence candidate grouping, offering transparency in the clustering process. This helps HR professionals understand the rationale behind automated decisions and supports informed shortlisting by verifying whether a candidate aligns with the intended job role. Additionally, the interpretability allows for bias detection, helping identify and mitigate unintended influences like location or gender-biased terms. By enabling trust and accountability through transparent explanations, the system increases user confidence in adopting AI-driven recommendations. Finally, LIME and SHAP support human-in-the-loop validation, allowing recruiters to override or validate clustering results, thereby maintaining explainability and auditability in hiring decisions. Recent studies [125, 293, 314, 399] demonstrate that explainable AI techniques improve transparency and fairness in HR processes, supporting our findings.

4.6.14 Limitations and Future Work

While the Gustafson-Kessel (GK) clustering algorithm outperforms traditional fuzzy methods such as FCM, it exhibits limitations when applied to resume data. Specifically, GK can struggle to maintain consistent cluster quality in the presence of imbalanced, high-dimensional, or noisy data. Additionally, its higher computational cost may limit scalability for large datasets.

Future work may address these challenges by exploring hybrid clustering frameworks that combine GK with more efficient or noise-resilient methods. For instance, integrating GK's capacity for modeling adaptive shapes with the scalability of k-means or the probabilistic strengths of GMM could yield more robust and efficient clustering outcomes. Enhancing these frameworks with explainability tools like SHAP or LIME would further improve transparency, which is critical for high-stakes domains like recruitment.

Graph-based clustering methods also represent a promising direction, as they are well-suited for to capture complex semantic relationships inherent in resume data. These methods could improve both cluster cohesion and interpretability in datasets characterized by rich interconnections among candidate profiles.

Additionally, incorporating bias and fairness auditing mechanisms is essential to detect and mitigate the influence of protected attributes such as gender or location. Empirical validation through user studies with HR professionals can further assess the system's usability and decision support capabilities.

Lastly, future research could explore the use of Large Language Models (LLMs), such as

GPT-4o, to generate natural language explanations at the cluster level. Although promising, this approach requires addressing challenges related to hallucination and computational cost. In general, these directions aim to enhance the robustness, fairness, and real-world applicability of explainable clustering systems in HR and related fields.

4.7 Conclusion

This chapter presents a novel explainable AI framework for clustering and summarizing resumes, specifically designed for data-driven industrial applications. Here we see that, Gustafson-Kessel (GK) fuzzy clustering algorithm in conjunction with embedding-based summarization and Local Interpretable Model-agnostic Explanations (LIME), effectively addresses challenges inherent in resume data, such as high dimensionality, semantic complexity, and weak feature correlations. Extensive evaluation on resume datasets using metrics like Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Silhouette Score, Precision, Recall, and F1-Score demonstrates that the GK algorithm consistently outperforms traditional clustering techniques, including K-means, Gaussian Mixture Models, Agglomerative Clustering, and Fuzzy C-Means. The superior precision and recall highlight its effectiveness in scenarios where accuracy and interpretability are crucial.

A significant contribution of this work is the integration of LIME and SHAP to enhance the transparency and interpretability of clustering outcomes. LIME offers actionable insights into feature weightage (importance) driving cluster assignments, thereby increasing trust and accountability in AI-driven decision-making processes relevant to recruitment and talent management within industrial environments. Additionally, embedding-based summarization improves clustering quality by capturing the semantic essence of resumes while reducing computational complexity, making the approach scalable for large datasets.

Overall, this methodology establishes a robust and explainable framework for automated resume analysis that advances the state-of-the-art in industrial AI applications. It has strong potential for deployment in recruitment, candidate profiling, and talent analytics. Future research may extend this approach to dynamic real-time data, explore more sophisticated summarization methods, and improve scalability for broader industrial applications.

Chapter 5

Hybrid Thresholding for Enhanced Performance and Interpretability in Fraud Detection: Integrating LIME and SHAP for Trustworthy AI Based Decision Making

In the previous chapter, we introduces an innovative approach that combines the Gustafson-Kessel (GK) fuzzy clustering algorithm with advanced semantic embeddings generated by Sentence-BERT to improve the interpretability of unsupervised learning for resume data. This work integrates Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to develop a hybrid thresholding method aimed at improving both interpretability and model performance in fraud detection.

By averaging the AUC-ROC and PR-AUC scores, the approach derives a balanced decision threshold that accounts for both class imbalance and performance stability, common challenges in data-driven industrial AI deployments. The model is evaluated using key metrics such as precision, recall, F1-score, MCC, and PR-AUC across varying thresholds to assess interpretability-performance trade-offs. This methodology ensures that AI systems used in industrial decision making, such as fraud detection, are not only accurate but also transparent and trustworthy. Thus, this contribution directly supports the objectives of this thesis by offering a reliable and explainable solution tailored for critical applications in data-driven industrial environments [301].

5.1 Introduction

Artificial Intelligence (AI) has become a key enabler in decision-making across critical sectors, such as banking, finance, healthcare, and cybersecurity. However, as AI models become more complex, their lack of transparency, often described as the "black-box" problem—raises significant concerns related to trust, fairness, and security. This challenge is especially acute in high-risk domains, where understanding the rationale behind AI-driven decisions is essential [114]. Despite their strong predictive performance, AI models often face skepticism from users, financial institutions, and regulatory bodies, particularly in customer-sensitive applications where clear explanations are mandatory.

To address these concerns, the field of Explainable AI (XAI) has emerged, developing methods to enhance transparency and accountability in AI systems [5, 29]. Among the most widely used XAI techniques are LIME[310] and SHAP[216]. Here we might say that LIME (Local Interpretable Model-agnostic Explanation) approximates complex models locally with simpler, interpretable models, whereas SHAP (SHapley Additive explanations) draws on cooperative game theory principles to attribute feature importance. However, LIME can be sensitive to input perturbations, and SHAP's computational complexity of SHAP may limit its scalability, especially for large datasets typical of financial applications [134].

In high-stake scenarios, such as the detection of credit card fraud, ensuring interpretability, fairness, and security is paramount. Although XAI techniques have advanced fraud detection, there remains a research gap in integrating these methods within a hybrid thresholding framework that balances performance and trustworthiness. In this study, we propose a novel hybrid thresholding approach that combines XGBoost classifiers with normalized LIME and SHAP explanations. This method enhances both the interpretability and predictive accuracy. We utilized precision-recall (PR) and AUC-ROC curves to comprehensively evaluate the model's ability to balance precision and recall, ensuring robust performance on imbalanced datasets [258]. By integrating these XAI techniques, our approach promotes transparency and fairness in decision-making, addressing critical requirements for trust and security in high-dimensional, security-sensitive financial domains.

5.1.1 Motivation

Although prior research has explored various fraud detection techniques and interpretability methods, there remains a gap in unifying robust thresholding strategies with explainable AI to address both performance and transparency in imbalanced datasets. This work is motivated by the need to bridge this gap by proposing a hybrid thresholding approach integrated with normalized LIME and SHAP explanations to, enhance the reliability and interpretability of

credit card fraud detection systems.

In recent research a comparison between LIME and SHAP [10] is presented but here we propose a different approach. To the best of our knowledge, the integration of LIME and SHAP normalization within a hybrid thresholding framework, coupled with the use of PR and ROC curves for performance evaluation, has not been previously explored in the literature, positioning our approach as a novel contribution.

5.1.2 Contributions

In this chapter presents the following key contributions:

1. Application of the synthetic minority over-sampling technique (SMOTE) with Principal Component Analysis (PCA) to address class imbalance and reduce dimensionality in credit card fraud detection.

2. Comparative evaluation of classifiers, with XGBoost identified as the most effective baseline model.

3. Integration of LIME and SHAP using min-max normalization and a hybrid contribution score ($\alpha = 0.5$) to enhance model interpretability.

4. Proposal of a hybrid thresholding method, averaging AUC-ROC and PR-AUC, to improve decision boundaries in imbalanced settings.

5. Empirical validation on a real-world dataset, demonstrated the improved performance and trustworthiness of the proposed framework.

The structure of the chapter is organized as follows: Section 5.2 outlines related work on Interpretability, Section 5.3 describes the problem formulation; Section 5.4 describes the methodology; Section 5.5 describes the Experimental Results, Section 5.6 describes the results and discussion; and Section 5.7 provides the conclusion.

5.2 Related Work on Model Interpretability

The increasing demand for interpretable machine learning models has driven the development of several XAI techniques. These methods are essential for improving the transparency of complex models, especially in high-stakes applications such as fraud detection [235]. Notable XAI techniques include LIME [311] and SHAP as in [216] and, [306], which help explain black-box models by providing local and global interpretability. LIME approximates complex

models using simpler, interpretable models to highlight feature contributions to individual predictions, as in [311]. However, LIME's reliance of LIME on local perturbations can lead to inconsistent explanations [25]. By contrast, SHAP builds on Shapley values from game theory, ensuring fair and consistent attribution of feature importance as in [216, 336]. Despite its advantages, SHAP can be computationally intensive, particularly for large datasets as in [98, 257]. To mitigate the shortcomings of individual methods, hybrid approaches have been explored in literature that combine multiple XAI techniques. For example, an approach [345] combines feature importance scores from various XAI methods to improve the robustness of the explanation, where as shown by [30] a framework is presented for aggregating the explanations to improve interpretability. Such approaches are particularly valuable in domains such as fraud detection, where both accuracy and transparency are paramount [29, 385, 413].

Recent studies have also focused on improving the scalability of XAI methods for large-scale applications[378]. Techniques for applying SHAP to large financial data sets and efficient SHAP algorithms to large models have been investigated [134]. Similarly, [263] presented some methods for optimizing LIME, improving both computational efficiency and stability.

5.2.1 Classification Methods, Data Imbalance and Interpretability

Before we present our hybrid threshold approach, we briefly review classification by unsupervised methods and also look at data imbalance issues as this is a key aspect of credit card fraud data sets. Further, more these methods, we also comment on their interpretability.

Here, we cite classification techniques that have often been employed in unsupervised settings and frequently incorporate dimensionality reduction methods such as Principal Component Analysis (PCA). Wu et al. [393] compared the following methods: isolation forest, extreme boosting based outlier detection (XGBOD), autoencoders, k-nearest neighbors (KNN), one-class support vector machine (OCSVM), one-class anomaly detection (OCAN), and copula-based outlier detector (COPOD). Although effective for outlier detection, these methods often rely on computationally intensive architectures and may lack interpretability, which is critical in credit card fraud detection.

A recent survey on machine learning techniques for credit card fraud detection [32], highlighted the importance of addressing data imbalance while maintaining model transparency. Although AUC-ROC is a standard metric for model evaluation [57], it may inadequately represent minority class performance in highly skewed datasets. Consequently, several studies [57, 69, 232, 258, 313, 323] advocate the use of precision-recall AUC (PR-AUC) as a more appropriate metric, given its emphasis on the minority class.

In this chapter, we adopt a hybrid threshold balancing approach that computes two balancing thresholds, one from the AUC-ROC curve by maximizing the difference between TPR (true

positive rate) and FPR (false positive rate), and another from the PR-AUC curve by maximizing the F1-score. The final decision threshold was determined by averaging these two values. Here we balance sensitivity and precision, making it especially suitable for imbalanced classification tasks, such as fraud detection. Existing classification models on imbalanced datasets [209, 420], have difficulty with classification accuracy and are inadequate for explainability. However, we wish to test the hybrid thresholding balancing method with LIME and SHAP explanations so that the integrated framework enhances both model performance and transparency, offering a robust solution for real-world fraud detection scenarios. To address these limitations, the proposed hybrid threshold balancing approach works within a supervised learning context, integrating LIME and SHAP normalization to improve interpretability and scalability in imbalanced datasets. We employ XGBoost as the base classifier, which is known for its effectiveness on structured and imbalanced data, as in credit card fraud detection. By leveraging well-established thresholding strategies from the AUC-ROC and PR-AUC analyses, we tailor a threshold balancing technique that supports both high performance and interpretability. Our approach utilizes both the AUC-ROC and PR-AUC for a more comprehensive evaluation, addressing the shortcomings of relying solely on a single metric. The hybrid thresholding strategy ensures a balanced contribution from both the local (LIME) and global (SHAP) explanation methods, leading to a more reliable, interpretable, and practical classification framework, particularly in domains where minority class accuracy is critical.

5.3 Problem Formulation

This chapter aims to develop a predictive model that classifies financial transactions as either fraudulent or legitimate using supervised learning. The dataset consists of the transactional and behavioral features associated with each transaction. Formally, given a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each x_i is a feature vector representing a transaction, and $y_i \in \{0, 1\}$ is the corresponding label (1 for fraud, 0 for legitimate), the goal is to learn a classification function $f : \mathbb{R}^m \to \{0, 1\}$ that predicts the label for any new, unseen transaction.

One of the key challenges in financial fraud detection is class imbalance, in which fraudulent transactions are far less frequent than legitimate transactions. This imbalance often leads to models favoring the majority class (legitimate transactions), hindering the accurate detection of fraud. To mitigate this, we optimized the classification threshold, seeking a balance between precision and recall by leveraging both the AUC-ROC and PR-AUC curves and, incorporating explainable AI (XAI) techniques. Specifically, we applied LIME and SHAP to enhance the transparency and trustworthiness of the model. These techniques provide valuable insights into the contribution of each feature to the model's predictions, which is particularly

critical in high-stakes domains, such as fraud detection, where decisions must be explainable and compliant with regulatory standards.

5.3.1 Model Interpretability with LIME and SHAP

1. LIME for Local Explanation: LIME explains a model's predicted outcome by approximating the behavior of a complex model locally around a specific instance x_i. LIME fits an interpretable surrogate model, which is typically linear, to simulate the model's decision-making process. The surrogate linear model for LIME [310] is defined as follows:

$$\hat{f}(x_i) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \tag{5.1}$$

Here, $\hat{f}(x_i)$ represents the approximated prediction for instance x_i, β_0 is the intercept, and β_j reflects the influence of feature x_{ij} on the model's decision as shown in eq. 5.1 [311]. This approximation allows instance-level insights into the features that contribute the most to the prediction, making LIME effective for understanding individual decisions.

2. SHAP for local and global explanation: SHAP [216] provides a more comprehensive approach to explain model predictions by quantifying the contribution of each feature to the model's output, drawing on principles from cooperative game theory. The Shapley value for feature x_j is defined as:

$$\phi_j = \sum_{S \subseteq \mathcal{N} \setminus \{j\}} \frac{|S|!(|\mathcal{N}| - |S| - 1)!}{|\mathcal{N}|!} [f(S \cup \{j\}) - f(S)] \tag{5.2}$$

Here, ϕ_j is the average marginal contribution of x_j across all possible subsets S of the features, and $\mathcal{N}$ is the full set of features, as shown in eq. 5.2. SHAP differs from LIME in that it provides both local and global interpretability, enabling the identification of feature importance at both the individual and dataset-wide levels. This makes SHAP particularly effective for understanding how different features consistently affect a model's output.

5.3.2 Integration for Interpretability

In our framework, we integrate LIME and SHAP to provide complementary insights: LIME offers localized and, instantaneous explanations for individual predictions. Simultaneously, SHAP captures the global feature importance patterns across the entire dataset. This combination of techniques enhances the interpretability of the model, which is crucial for both debugging the model and ensuring its robustness and trustworthiness, particularly in high-stakes domains, such as fraud detection.

5.3.3 Hybrid Threshold for Balanced Classification

The next step of our approach is to define a balanced classification threshold for fraud detection. Traditional thresholding methods often fail to adequately address the challenges posed by class imbalance, leading to poor performance in detecting the instances of minority classes. By incorporating both the perspectives of the AUC-ROC and PR-AUC curves, we aim to derive a hybrid threshold that ensures balanced precision, recall, and F1 score. This threshold was computed using the following formulation, as shown in eq. 5.3.

$$\begin{aligned} \text{Balanced Threshold} = \frac{1}{2}\big[&\arg\max_{\theta \in \text{ROC}} (\text{TPR} - \text{FPR}) \\ &+ \arg\max_{\theta \in \text{PR}} (\text{F1-score}) \big] \end{aligned} \tag{5.3}$$

This hybrid threshold may be computed quite simply by balancing the insights from both AUC-ROC and PR-AUC analyses, without relying on any optimization objective. It aims to maintain a trade-off between sensitivity and precision, leading to more reliable binary decisions in imbalanced data set settings. The final threshold is then applied to the predicted probabilities of the classification model.

In the following section, we give more details on the methodology of our model, including the implementation of the hybrid threshold approach and the integration of LIME and SHAP.

5.4 Methodology

We now present our approach to credit card fraud detection, including steps for data preprocessing, feature engineering and selection, model training, application of explainable AI techniques, normalization, hybrid thresholding, and model evaluation. The framework in Figure 5.1 illustrates our methodology for detecting fraudulent behavior of individuals using credit card data. Details of these techniques are provided in the following sections. Unlike anomaly detection approaches, we focus on detecting fraudulent transactions using classification-based methodology.

5.4.1 Dataset and Preprocessing

The dataset used in this study is a publicly available *Credit Card Fraud Detection* dataset from the Kaggle[1]. It comprises 284,807 transactions, of which only 492 are labeled as fraudulent, showing that the dataset is highly imbalanced with fraud cases accounting for only 0.172% of the total. The target variable `class` is binary, where 1 indicates a fraudulent transaction and

[1] `https://www.kaggle.com/mlg-ulb/creditcardfraud`

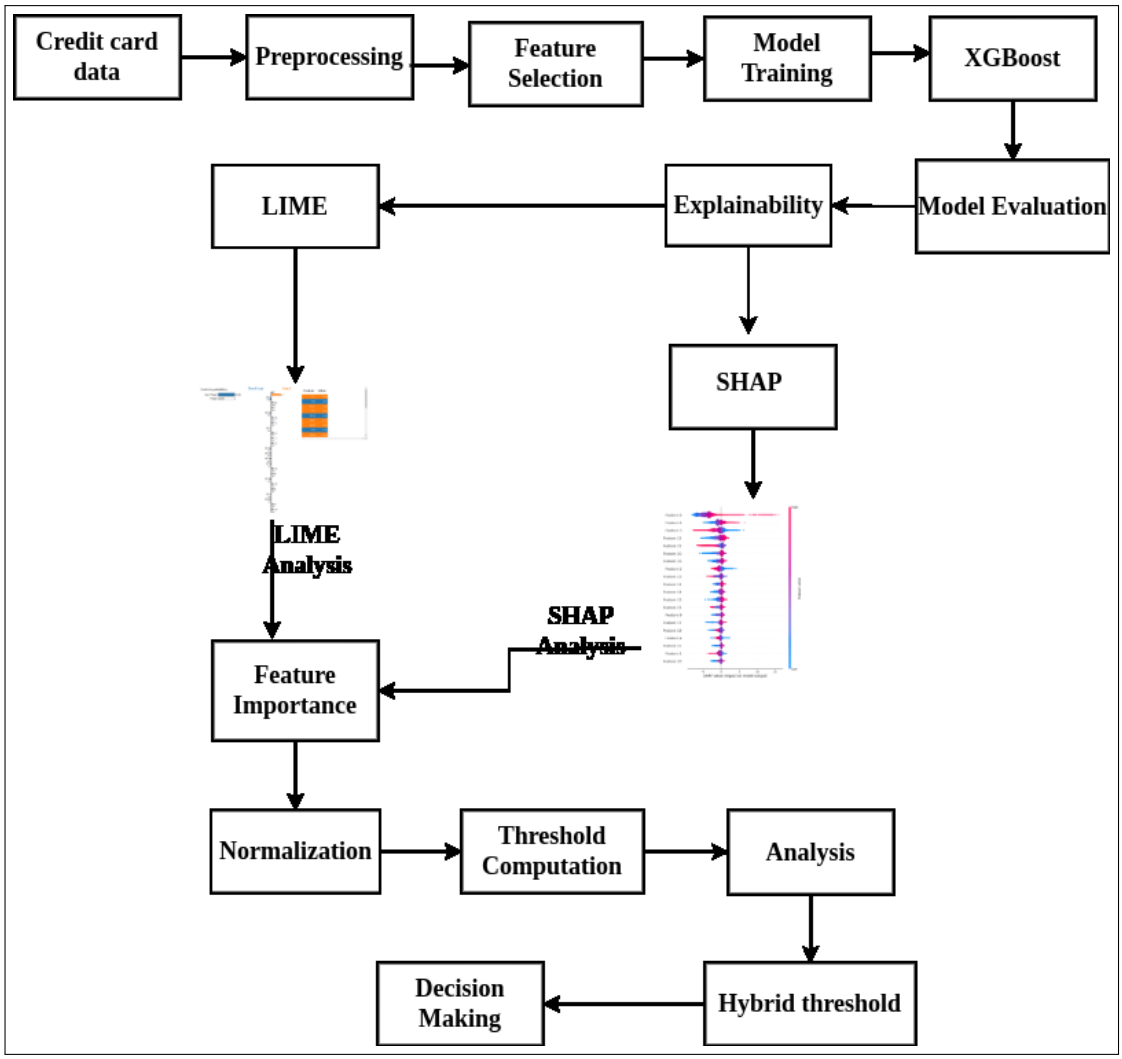


Figure 5.1: Proposed framework for explainable hybrid thresholding in credit card fraud detection.

0 indicates a legitimate transaction. To preserve the class distribution, an 80%-20% stratified split was used for training and testing. All random processes were seeded using `random_state = 42` to ensure reproducibility.

5.4.2 Handling Class Imbalance with SMOTE

To address class imbalance, we applied a synthetic minority over-sampling technique (SMOTE) [73] to the training data. SMOTE generates new synthetic samples for the minority class by interpolating existing minority class instances. The implementation us *imblearn.over_sampling.SMOTE* with default settings and `random_state = 42`.

5.4.3 Feature Scaling and Dimensionality Reduction

All features were normalized using the `StandardScaler` method to obtain a zero mean and unit variance. Subsequently, Principal Component Analysis (PCA) was applied to reduce dimensionality and computational complexity. The number of components was selected such that 99% of the variance was retained. This step also mitigates multicollinearity among the features.

5.4.4 Model Training with XGBoost

The classifier selected for this experiment is Extreme Gradient Boosting (XGBoost) [74], a tree-based ensemble method that performs well on structured data. The model was trained using PCA-transformed and SMOTE-balanced training data. The parameters for XGBoost were set as follows.

- `use_label_encoder = False`
- `eval_metric = 'logloss'`
- `random_state = 42`

The model outputs the class probabilities for each test instance, which are used in threshold optimization.

5.4.5 Threshold Balancing

To address class imbalance in fraud detection, we propose a hybrid threshold- balancing strategy instead of relying on a default threshold of 0.5. This approach integrates both the AUC-ROC and PR-AUC perspectives to ensure balanced performance. First, we identify the optimal

AUC-ROC threshold by maximizing the difference between the true positive rate (TPR) and the false positive rate (FPR), as shown in eq. 5.4.

$$\text{Threshold}_{\text{ROC}} = \arg\max_{t}(\text{TPR}(t) - \text{FPR}(t)) \tag{5.4}$$

Next, we computed the optimal threshold from the PR curve by calculating the F1 score for each threshold value using precision and recall pairs, selecting the threshold that maximizes the F1 score as shown in eq. 5.5

$$\text{Threshold}_{\text{PR}} = \arg\max_{t}(\text{F1-score}(t)) \tag{5.5}$$

Finally, the hybrid threshold was derived by averaging the two, as shown in eq. 5.6

$$\text{Hybrid Threshold} = \frac{\text{Threshold}_{\text{ROC}} + \text{Threshold}_{\text{PR}}}{2} \tag{5.6}$$

This hybrid method balances sensitivity and precision, leading to improved classification boundaries in imbalanced scenarios, while preserving interpretability.

To assess the effectiveness of our proposed model, we report all the evaluation results in the Results and Discussion section. The model demonstrated strong performance across key metrics, particularly under the imbalanced nature of fraud- detection data. The application of the hybrid threshold significantly improved the classification performance. Additionally, the confusion matrix indicates a high true positive rate and a reduction in false negatives, which are essential for minimizing undetected fraudulent transactions. These findings validated the reliability and robustness of the proposed approach.

5.4.6 Explainability

We employ LIME and SHAP to improve the interpretability of our credit card fraud- detection model. LIME provides instance-level explanations by approximating the local decision boundary, whereas SHAP quantifies both the local and global feature contributions using Shapley values. Their outputs we normalized and combined using a hybrid thresholding strategy to enhance model transparency, classification robustness, and compliance with trustworthy AI principles.

5.4.7 Feature Importance and Explainability

LIME and SHAP we applied to quantify the influence of the features on the model predictions. LIME identifies features that contribute most to individual decisions [310], whereas SHAP

provides consistent-additive attributions across instances and the dataset [357]. Figure 5.3 illustrates both types of explanation. Normalized importance scores from both methods we fused using hybrid thresholding to improve interpretability and classification performance.

5.4.8 Describing the LIME Process

LIME perturbs the input features and fits the local surrogate model to approximate the decision function of the original model. It produces normalized feature importance scores per instance to interpret fraudulent predictions. Algorithm 5 details the initialization and execution. The first instance was used in the absence of fraud. The top 20 features that influence each decision are reported to support transaction-level analysis.

Algorithm 5: Adopted LIME Algorithm for Fraud Detection [310]

Require: Preprocessed training data X_{train}, test data X_{test}, trained model f
Apply PCA to X_{train} and X_{test}
define feature names: $\text{PC}_1, \text{PC}_2, \ldots, \text{PC}_n$
Initialize LIME explainer:
    explainer $\leftarrow$ LimeTabularExplainer($X_{\text{train\_pca}}$, feature_names, class_names=$\{$Non-Fraud, Fraud$\}$, mode=classification, discretize_continuous=False)
Select $x \in X_{\text{test\_pca}}$
if fraud instances exist in X_{test} **then**
    $x \leftarrow$ first fraud instance
else
    $x \leftarrow X_{\text{test\_pca}}[0]$
end if
Generate explanation:
    $e \leftarrow$ explainer.explain_instance$(x, f.\text{predict_proba}, num_features = n)$
Visualize or save explanation
return the top k influential features with weights

5.4.9 Describing the SHAP Process

Algorithm 6 outlines the process for ensuring feature consistency [216]. SHAP values we computed using *TreeExplainer* on the trained XGBoost model to, quantify each feature's contribution to the prediction. Here, we must mention that *TreeExplainer* in the context of SHAP is a fast and exact method for estimating SHAP values for tree models and ensembles of trees, under several different possible assumptions about feature dependence.

Algorithm 6: Adopted SHAP Algorithm [216]

Require: X_{test} (test set), *xgb_model* (trained XGBoost model)
Ensure: *shap_values* (feature importance for each instance)
Initialize SHAP explainer:
shap_explainer ← `TreeExplainer`(*xgb_model*)
Compute SHAP values:
shap_values ← *shap_explainer*.`shap_values`(X_{test})
if `isinstance`(*shap_values*, list) **then**
shap_values ← *shap_values*[1] {For multiclass outputs}
end if
return *shap_values*

5.4.9.1 Normalization and Feature Contribution Fusion

To integrate LIME and SHAP contributions equally in fraud detection, we apply min-max normalization to scale their feature importance scores to the $[0,1]$ range. LIME provides local explanations with smaller magnitudes, whereas SHAP captures global contributions that can vary widely in scale. Without normalization, one method can dominate the fusion, leading to biased interpretations. Because both explanations share identical feature dimensions $(113726, 17)$, normalization ensures a balanced influence. The normalized scores are then combined using a hybrid thresholding strategy, that combines, both interpretability and classification performance. This step reinforces the model transparency and trustworthiness in the detection process.

5.4.10 Hybrid Threshold Balancing Method

This method enhances fraud detection by integrating the LIME and SHAP explanations. It combines the normalized feature contributions to compute a hybrid score for each instance. This score is then used to determine an optimal classification threshold based on ROC and Precision-Recall analysis. The method improves model interpretability while maintaining robust predictive performance. The complete process is presented in Algorithm 7.

Note: Let L_{ij} and S_{ij} represent the raw LIME and SHAP contributions for feature j of instance i. These we normalized using min-max normalization to obtain $\tilde{L}_{ij}$ and $\tilde{S}_{ij}$ respectively. The equal-weight combined contribution is computed as $C_{ij} = \frac{\tilde{L}_{ij}+\tilde{S}_{ij}}{2}$, while the hybrid contribution is given by

$$H_{ij} = \alpha \cdot \tilde{L}_{ij} + (1-\alpha) \cdot \tilde{S}_{ij}, \quad \alpha \in [0,1] \tag{5.7}$$

The suitability score for each instance was $C_i = \sum_j C_{ij}$. The optimal decision threshold is ob-

Algorithm 7: Hybrid Thresholding with LIME and SHAP

Require: $lime\_explanations$, $shap\_values$, X_{test}, α, $num\_features$
Ensure: Predictions and evaluation metrics
Normalize LIME and SHAP contributions:
for each instance i in X_{test} **do**
$\tilde{L}_{ij} = \frac{L_{ij} - \min(L)}{\max(L) - \min(L)}$, $\tilde{S}_{ij} = \frac{S_{ij} - \min(S)}{\max(S) - \min(S)}$
end for
Compute Hybrid Contributions:
for each instance i, feature j **do**
$C_{ij} = \frac{\tilde{L}_{ij} + \tilde{S}_{ij}}{2}$ // Equal-weight average
$H_{ij} = \alpha \cdot \tilde{L}_{ij} + (1 - \alpha) \cdot \tilde{S}_{ij}$ // Weighted hybrid
end for
Threshold Estimation:
Compute ROC and PR curves
$threshold\_roc \leftarrow \arg\max(TPR - FPR)$
$threshold\_pr \leftarrow \arg\max(\text{F1-score})$
$optimal\_threshold \leftarrow \frac{threshold\_roc + threshold\_pr}{2}$
Compute Suitability Score:
for each instance i **do**
$C_i = \sum_j C_{ij}$
end for
Predict using Hybrid Threshold:
for each i **do**
$$\hat{y}_i = \begin{cases} 1, & \text{if } C_i \geq optimal\_threshold \\ 0, & \text{otherwise} \end{cases}$$
end for
Evaluate Metrics: Accuracy, Precision, Recall, F1, AUC, PR-AUC, MCC (Matthews Correlation Coefficient), Sensitivity, Specificity
return $\hat{y}$, evaluation metrics

tained by averaging the thresholds that maximize $TPR - FPR$ (from the AUC-ROC curve) and F1-score (from the precision-recall curve) PR-AUC. Predictions $\hat{y}_i$ we made by comparing C_i with the optimal threshold. Instances where $C_i \geq optimal\_threshold$ we classified as fraudulent. Performance-evaluated metrics are reported in the results and discussion section.

5.4.11 Computational Complexity Analysis

The overall complexity of the proposed method is dominated by threshold optimization, which has a time complexity of $O(m \times n)$, where n is the number of instances and m is the number of threshold values. If $m \ll n$, the complexity approximates to $O(n)$; otherwise, it scales linearly with both n and m. For Random Forest, the training complexity is $O(T \times d \times \log(n))$, where T is the number of trees and d is the average tree depth. The computing explanations for each instance have complexity $O(n \times p)$ for LIME and SHAP, where p is the number of features. Thresholding methods iterate over k threshold values, contributing to the complexity of $O(n \times k)$. Thus, the total complexity of the method was: $O(T \times d \times \log(n)) + O(n \times p) + O(n \times k)$. This demonstrates that the computational cost depends on the number of trees, dataset size, feature dimensionality, and threshold evaluations.

5.5 Experimental Results

The following dataset was used in the evaluation:

5.5.1 Dataset Description

This chapter employs the publicly available Credit Card Fraud Detection dataset, originally provided by researchers from the Université Libre de Bruxelles (ULB) and hosted on Kaggle [1]. The dataset contains anonymized features, including 'Time', 'Amount', and 28 principal components resulting from a PCA transformation of the original input variables. A major challenge of this dataset is its severe class imbalance, with fraudulent transactions comprising approximately only 0.17% of the total. To address this, the Synthetic Minority Oversampling Technique SMOTE) was applied to balance the class distribution, thereby improving the effectiveness of the learning algorithms. All data usage complies with the licensing and access policies of the host platform.

[1] `https://www.kaggle.com/mlg-ulb/creditcardfraud`

5.5.2 Simulation Environment and Equipment

All experiments were conducted using Python 3.9.12 within a Jupyter Notebook environment on a system equipped with an Intel Core i5 processor and 16 GB RAM. The simulation environment utilized a range of libraries and tools, as detailed below:

- **Data Processing:** `Pandas`, `Numpy`, and `scikit-learn` for data manipulation and preprocessing.
- **Modeling Framework:** XGBoost classifier with parameters `use_label_encoder=False` and
  `eval_metric='logloss_'`, integrated with
  `scikit-learn` for model development.
- **Explainability:** `LIME` and `SHAP` for local and global model interpretability, including the normalization of feature importance values.
- **Hybrid Thresholding:** Optimal threshold is determined as the average of the values derived from the ROC and Precision-Recall curves.
- **Threshold Analysis:** Evaluation across thresholds ranging from 0.1 to 0.9 in increments of 0.1.
- **Validation Strategy:** 5-fold cross-validation was performed `StratifiedKFold` with a fixed random state of 42 to ensure reproducibility.

This setup ensures a robust and reproducible computational environment for the development, training, and evaluation of machine learning models that address the challenges of imbalanced credit card fraud detection.

Table 5.1: Evaluation Metrics for Credit Card Fraud Detection

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall (Sensitivity)	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
F1-Score	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
MCC	$\frac{(TP \cdot TN)-(FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Note: TP = true positives, TN = true negatives, FP = false positives, FN = false negatives. These formulas are standard evaluation metrics used to assess the performance of classification models in imbalanced datasets such as credit card fraud detection.

In this evaluation, the following terms are used: **True Positives (TP)** refers to correctly predicted fraud cases, **True Negatives (TN)** refers to correctly predicted non-fraud cases, **False Positives (FP)** refers to non-fraud cases incorrectly classified as fraud, and **False Negatives (FN)** refers to fraud cases incorrectly classified as non-fraud.

Additionally, we considered the **True Positive Rate (TPR)**, which measures the proportion of actual fraud cases correctly identified, and the **False Positive Rate (FPR)**, which measures the proportion of non-fraud cases misclassified as fraud, to assess the model performance. For evaluation, we trained an XGBoost model using a hybrid threshold method, as detailed in Tables 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, and 5.7. To enhance interpretability, we applied **LIME** and **SHAP** to analyze feature importance.

Table 5.2: Comparison of Baseline Methods on SMOTE

Model	**Train Time (s)**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**	**MCC**
Random Forest	54.71	0.9998	0.9996	1.0000	0.9998	1.0000	0.9996
ANN	52.20	0.9997	0.9908	1.0000	0.9997	1.0000	0.9994
Logistic Regression	3.24	0.9809	0.9908	0.9708	0.9807	0.9975	0.9619
SVM	3720	0.9960	0.9976	0.9945	0.9960	0.9997	0.9921
KNN	0.02	0.9990	0.9981	1.0000	0.9990	0.9997	0.9981
Decision Tree	38.30	0.9985	0.9980	0.9990	0.9985	0.9985	0.9969
Gradient Boosting	474.47	0.9877	0.9934	0.9819	0.9876	0.9993	0.9754
Naive Bayes	0.16	0.9241	0.9733	0.8722	0.9200	0.9751	0.8529
XGBoost	44.65	0.9998	0.9996	1.0000	0.9998	1.0000	0.9996
Ours (Hybrid)	**46.87**	**0.9999**	**0.9999**	**0.9999**	**1.0000**	**1.0000**	**0.9998**

Note: All models were evaluated using SMOTE-balanced data. PCA was not applied.

Table 5.3: Comparison of Baseline Methods on PCA

Random Forest	61.43	0.9995	0.9487	0.7551	0.8409	0.9486	0.8462
ANN	4478.33	0.9995	0.8804	0.8265	0.8526	0.9777	0.8528
Logistic Regression	5.40	0.9752	0.0601	0.9184	0.1129	0.9716	0.2315
SVM	1960.95	0.9968	0.3217	0.7551	0.4512	0.9732	0.4917
KNN	0.08	0.9995	0.9186	0.8061	0.8587	0.9437	0.8603
Decision Tree	25.01	0.9990	0.7041	0.7041	0.7041	0.8518	0.7036
Gradient Boosting	271.04	0.9988	0.8372	0.3673	0.5106	0.5341	0.5541
Naive Bayes	0.32	0.9783	0.0637	0.8469	0.1184	0.9597	0.2287
XGBoost	463.68	0.9996	0.9195	0.8163	0.8649	0.9729	0.8662
Ours (Hybrid)	**44.86**	**0.9994**	**0.8235**	**0.8571**	**0.8400**	**0.9729**	**0.8339**

Note: All models were evaluated using data transformed using PCA, with 95% explained variance. SMOTE was applied prior to PCA on the training set to address the class imbalance.

Figure 5.2 illustrates the LIME output for a specific prediction, showing each feature's

Table 5.4: Comparison of Baseline Methods on SMOTE with PCA

Methods	Accuracy	Precision	Recall	F1 Score	AUC-ROC	PR-AUC	Specificity	MCC
Random Forest	0.9998	0.9997	0.9999	0.9998	1.0000	1.0000	0.9997	0.9997
ANN	0.9996	0.9992	1.0000	0.9996	1.0000	1.0000	0.9992	0.9992
Logistic Regression	0.9772	0.9881	0.9661	0.9770	0.9966	0.9969	0.9883	0.9547
SVM	0.9953	0.9970	0.9936	0.9953	0.9997	0.9996	0.9970	0.9906
KNN	0.9989	0.9978	1.0000	0.9989	0.9996	0.9992	0.9978	0.9978
Decision Tree	0.9978	0.9971	0.9986	0.9978	0.9978	0.9964	0.9971	0.9956
Gradient Boosting	0.9782	0.9922	0.9640	0.9779	0.9985	0.9984	0.9924	0.9568
Naive Bayes	0.9392	0.9825	0.8944	0.9364	0.9823	0.9846	0.9841	0.8820
XGBoost	0.9997	0.9994	1.0000	0.9997	1.0000	1.0000	0.9994	0.9993
Ours (Hybrid)	**0.9998**	**0.9997**	**0.9999**	**0.9998**	**1.0000**	**1.0000**	**0.9997**	**0.9996**

Note: Our hybrid threshold method demonstrates exceptional performance with a cross-validation accuracy of 0.9998 ± 0.0001, along with near-perfect precision, recall, F1 score, AUC-ROC, and MCC values. These results indicate that the model effectively handles the severe class imbalance in the dataset while maintaining a strong predictive power across all relevant evaluation metrics.

Table 5.5: Wilcoxon Signed-Rank Test Comparing Models with XGBoost Across 8 Metrics (Accuracy, Precision, Recall, F1 Score, AUC, PR-AUC, Specificity, MCC)

Model	p-value	Statistically Different ($\alpha = 0.05$)
Random Forest	0.31731	No
ANN	0.04520	Yes
Logistic Regression	0.00781	Yes
SVM	0.00781	Yes
KNN	0.01563	Yes
Decision Tree	0.00781	Yes
Gradient Boosting	0.00781	Yes
Naive Bayes	0.00781	Yes
XGBoost	0.06837	No
Hybrid (Proposed)		No*

Note: This table presents the p-values from the Wilcoxon signed-rank test comparing each model to XGBoost across the eight evaluation metrics. Models with p-values less than 0.05 are considered statistically different in performance. Random Forest and XGBoost did not differ significantly from each other. The Hybrid model is marked with a star (*) because it is an extension of XGBoost; therefore, a statistical comparison is not applicable.

Table 5.6: Confusion Matrix of the classifiers

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positives)	FN (False Negatives)
Actual Negative	FP (False Positives)	TN (True Negatives)

Note: The confusion matrix illustrates classification performance. TP: correctly predicted positives, FN: actual positives incorrectly predicted as negatives, FP: actual negatives incorrectly predicted as positives, TN: correctly predicted negatives.

Table 5.7: Confusion Matrix of the classifiers after applying SMOTE + PCA

Model	TN	FP	FN	TP
Random Forest	56847	16	3	56860
ANN	56816	47	0	56863
Logistic Regression	56200	663	1926	54937
SVM	56693	170	366	56497
Knn	56736	127	0	56863
Decision Tree	56696	167	81	56782
Gradient Boosting	56430	433	2045	54818
Naive Bayes	55958	905	6007	50856
XGBoost	56827	36	1	56862
Ours (Hybrid)	**56844**	**19**	**3**	**56860**

Note: All classifiers were trained and evaluated after applying SMOTE and PCA. The **Ours (Hybrid)** model incorporates a hybrid thresholding technique applied to XGBoost, which yields superior balance and performance in handling class imbalance compared with other baseline models. TP: true positives, TN: true negatives, FP: false positives, FN: false negatives.

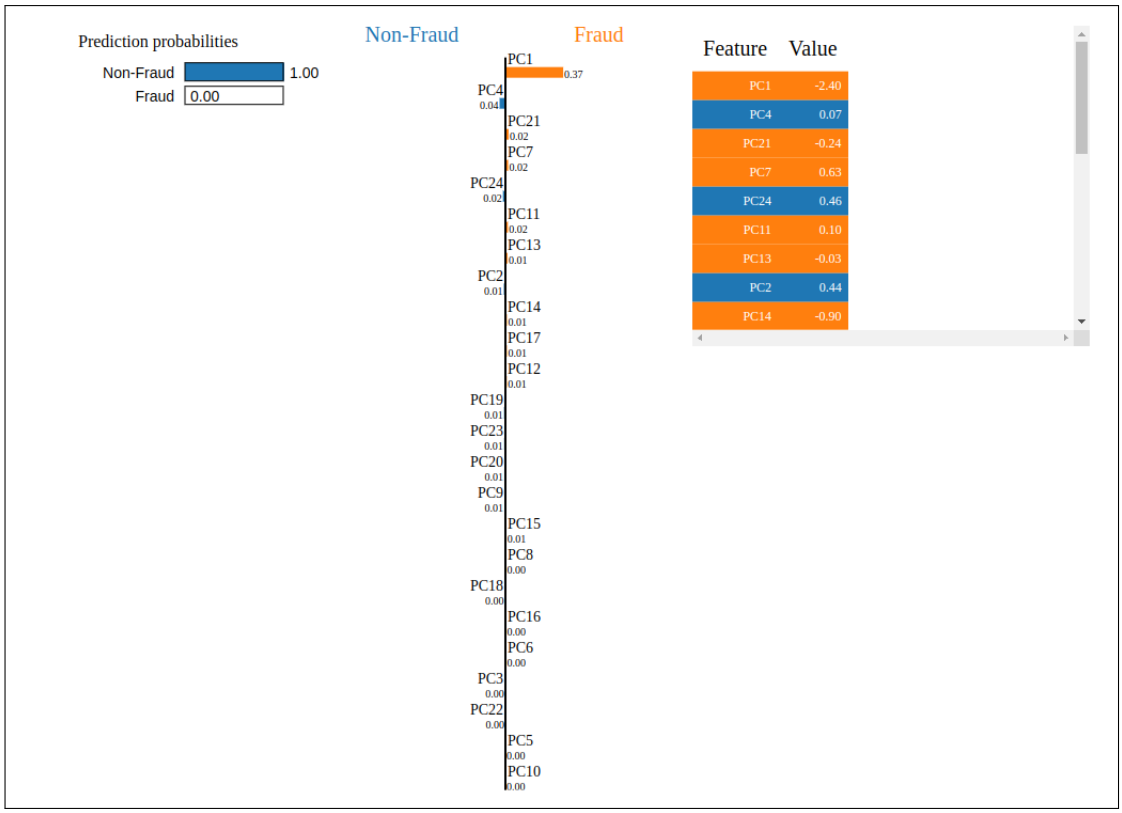


Figure 5.2: Example of one instance LIME Analysis

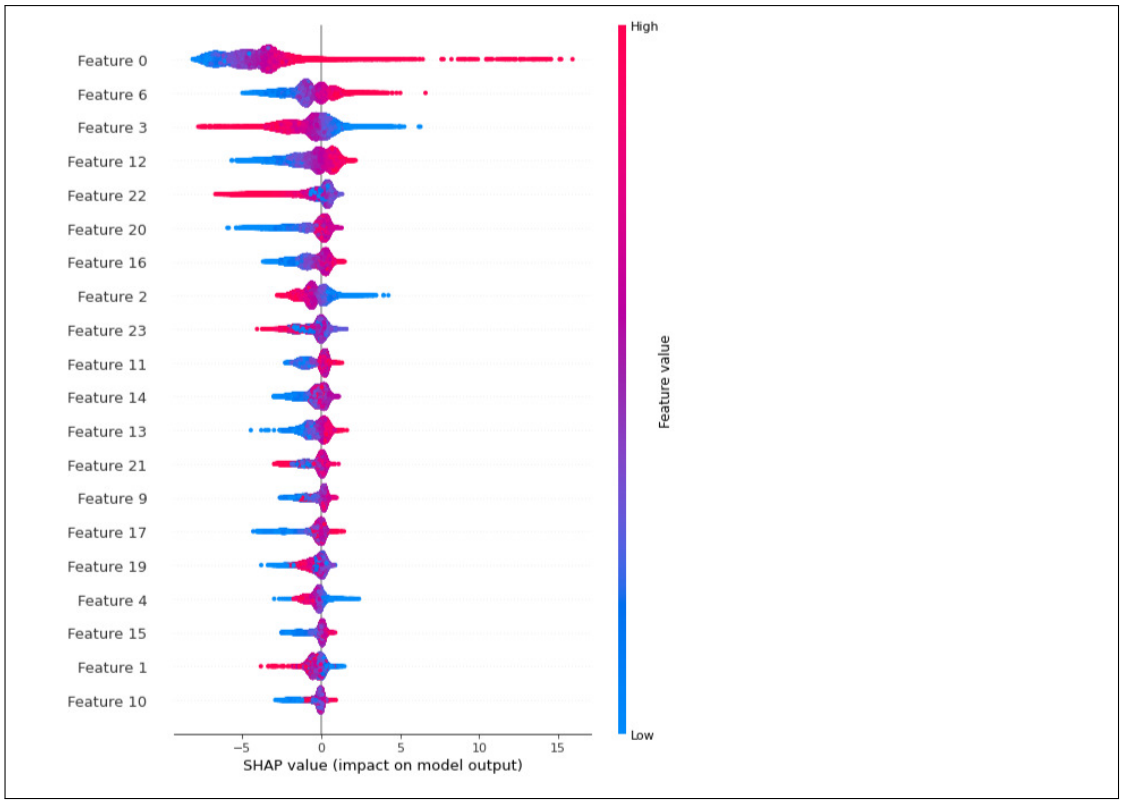


Figure 5.3: Example of SHAP Analysis

contribution to the machine learning model's decision. The bar lengths represent the magnitude of each feature's influence, providing insights into the effect of each individual feature on the prediction. LIME generates a simplified, interpretable model that locally approximates the behavior of a complex model around a given instance, thereby enhancing transparency and interpretability at the local level. This visualization highlights the key factors that drive the model's predictions.

Figure 5.3 illustrates the SHAP output, which shows the impact of individual features on the model predictions. Each point represents a feature's contribution to a particular prediction, with the color indicating the feature value (e.g., high values in pink and low values in blue). The horizontal position reflects whether the feature value increases or decreases the prediction. The SHAP values provide a unified measure of feature importance, clearly illustrating how each feature positively or negatively influences the output. This visualization improves the interpretability and transparency of the machine learning model used to detect fraudulent behavior in credit card transactions. Figure 5.4 shows the PR-AUC curve illustrating the performance of the classifier on data from unbalanced credit card fraud. Figure 5.5 shows the AUC-ROC curve illustrating the performance of the classifier using data from unbalanced credit card fraud. Note that XGBOOST should be used as the baseline for comparison.

Table 5.8: Performance of the Proposed Hybrid Threshold Balancing Method on Credit Card Fraud Detection Data After LIME and SHAP-Based Feature Normalization

Method	Accuracy	Precision	Recall	F1 Score	AUC-ROC	PR-AUC	Specificity	MCC
Proposed	**0.9109**	**0.9017**	**0.9224**	**0.9119**	**0.9703**	**0.9695**	**0.8994**	**0.8221**
ROC	**0.9110**	0.9040	0.9197	0.9118	0.9703	0.9695	0.9024	0.8222
PR	**0.9109**	0.8993	0.9254	0.9121	0.9703	0.9695	0.8963	0.8221

Note: The hybrid threshold balancing method uses a contribution score calculated by combining the LIME and SHAP explanations with a weighting factor alpha set to 0.5. Here, PR refers to the area under the precision-recall curve (PR-AUC), and AUC refers to the area under the Receiver Operating Characteristic curve (AUC-ROC). The hybrid AUC metric was computed as the average of the PR-AUC and AUC-ROC values.

5.5.3 Hybrid Threshold Balancing and Interpretability in Fraud Detection

Table 5.8 compares three threshold balancing methods for credit card fraud detection: the proposed hybrid, AUC-ROC based, and PR-AUC based thresholds. The Hybrid Threshold was computed as the average of the optimal thresholds derived from the AUC-ROC criterion (maximizing TPR- FPR) and the PR-AUC criterion (maximizing F1-score). This balancing approach considers both sensitivity and precision, resulting in more reliable classification performance under class imbalance. As shown in Equation 5.7, the hybrid contribution score quantifies the

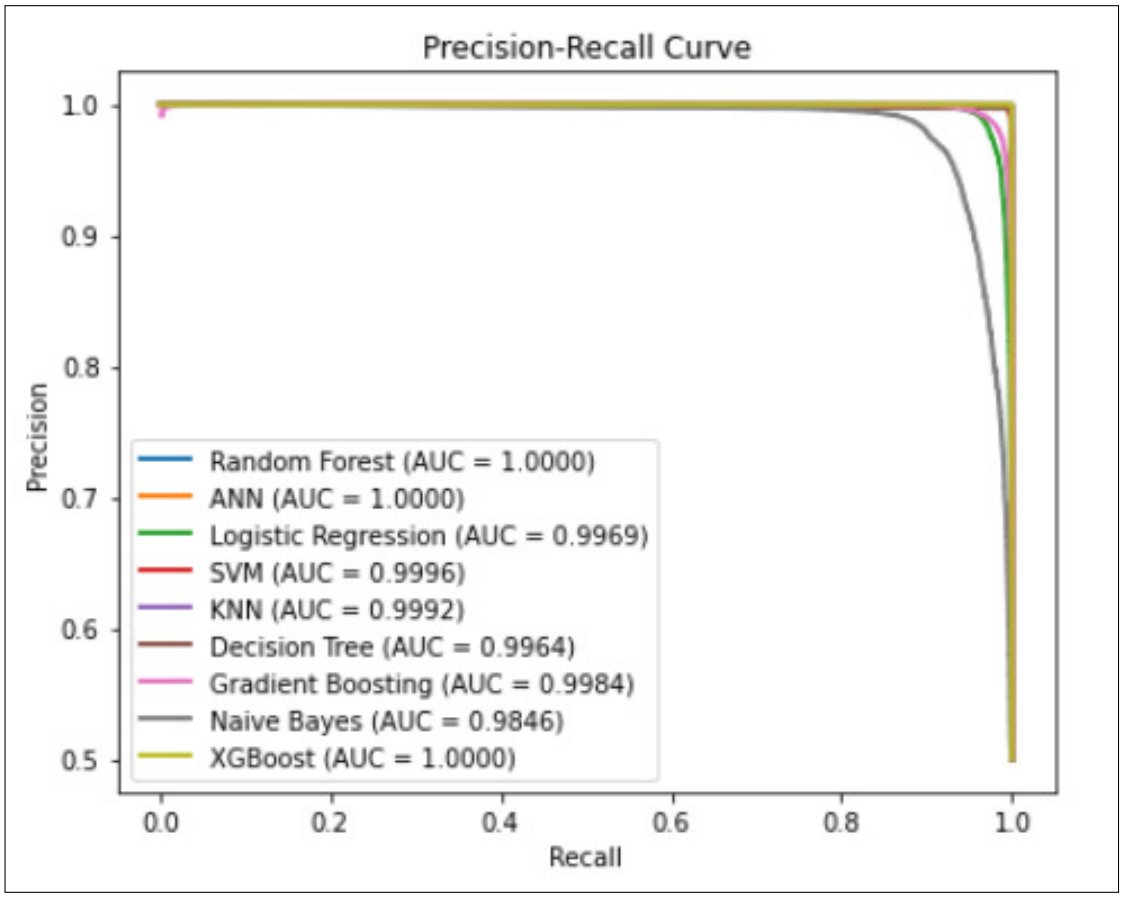


Figure 5.4: PR-AUC curve illustrating classifier performance on imbalanced credit card fraud data

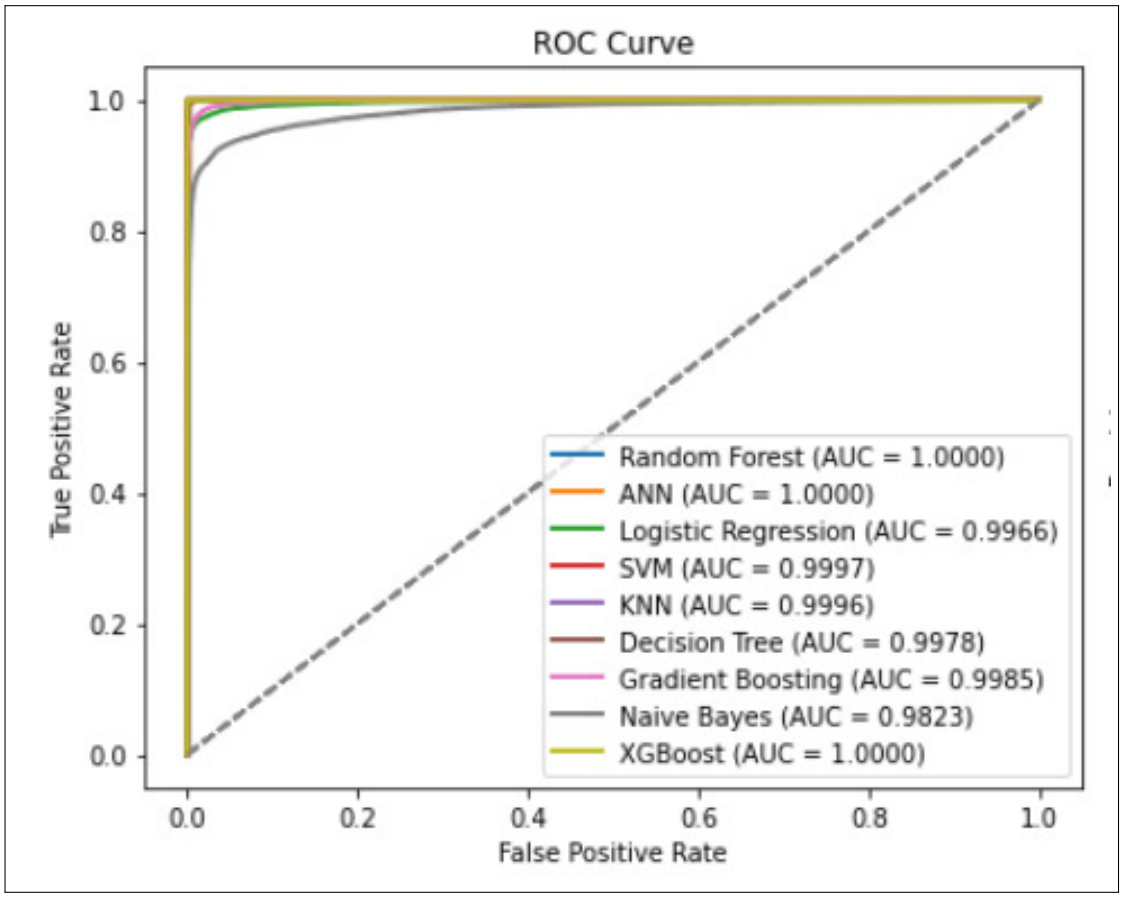


Figure 5.5: AUC-ROC curve illustrating classifier performance on imbalanced credit card fraud data

combined influence of the two criteria.

This formulation ensures a balanced trade-off between precision and recall, which is critical in imbalanced settings such as fraud detection.

Performance Analysis: As shown in Table 5.8, all three methods demonstrated strong performance across Accuracy, F1 Score, AUC-ROC, and PR-AUC. The Hybrid Threshold slightly outperforms the others in the F1 Score (0.9119), indicating a better balance between false positives and false negatives. It also maintained a high specificity (0.8994) and achieved a strong Matthews Correlation Coefficient (MCC), ensuring reliable prediction capability.

Precision-Recall Trade-off: The PR-AUC based method yielded the highest recall (0.9254) but with lower precision (0.8993), whereas the AUC-ROC-based method showed better precision (0.9040) but slightly reduced recall (0.9197). The Hybrid Threshold achieved a balanced outcome with a precision of 0.9017 and recall of 0.9224, combining the strengths of both strategies.

Interpretability and Static Nature of Threshold: This Hybrid Threshold is a *static* decision boundary derived from the evaluation metrics on the validation set. It is neither adaptive nor dynamic, nor does it rely on anomaly detection or unsupervised learning paradigms. Instead, it provides a consistent and interpretable threshold that enhances decision boundary transparency and trust in the model's output, which is essential in critical applications such as fraud detection.

5.5.4 Hybrid Contribution Score and Its Evaluation

As shown in eq. 5.7, the hybrid contribution score, where $\alpha \in [0,1]$ controls the relative contributions of the LIME and SHAP. Table 5.9 presents the evaluation of model performance across various α values.

The results show that $\alpha = 0.5$ yields the most balanced performance, achieving the highest AUC-ROC (0.9710), MCC (0.8233), and F1 score (0.9127). This suggests that equal weighting between LIME and SHAP produces more reliable and interpretable feature attribution. Incorporating this hybrid contribution score into our Hybrid Threshold method computed as the average of thresholds optimized using AUC-ROC and PR-AUC criteria further enhances model performance. This improves recall by effectively identifying fraudulent cases while maintaining high precision, thereby minimizing false positives. This makes the approach particularly suitable for imbalanced datasets such as credit card fraud detection, where reliable and interpretable decision boundaries are essential.

Table 5.9: Performance of Hybrid Thresholding at Different Alpha (α) Values

Metric	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
Accuracy	0.9694	0.9656	0.9571	0.9115	0.7785	0.6695	0.5099
Precision	0.9596	0.9553	0.9455	0.9004	0.7571	0.6408	0.5051
Recall (Sensitivity)	0.9799	0.9770	0.9701	0.9254	0.8202	0.7714	0.9796
Specificity	0.9588	0.9543	0.9441	0.8976	0.7368	0.5675	0.0402
F1 Score	0.9697	0.9660	0.9576	0.9127	0.7874	0.7001	0.6665
AUC-ROC	0.9953	0.9943	0.9920	0.9710	0.8626	0.7424	0.5980
MCC	0.9389	0.9315	0.9145	0.8233	0.5590	0.3462	0.0578
PR-AUC	0.9948	0.9937	0.9912	0.9703	0.8618	0.7367	0.5912

Note: Parameter α represents the weighting factor used in the hybrid thresholding method, balancing the contributions from the LIME and SHAP explanations. Specifically, the hybrid contribution score is computed as a weighted sum where α controls the influence of LIME, and $(1-\alpha)$ controls the influence of SHAP.

5.5.5 Hybrid Threshold Method: LIME and SHAP Evaluation Metrics

This section presents the evaluation metrics for the Hybrid Threshold Method, focusing on LIME and SHAP's performance:

LIME Fidelity (first 10 instances): The average fidelity of LIME is 0.1709, with a range from 0.0020 to 0.6848 across individual instances. This indicates varied accuracy in LIME's local explanations.

LIME Stability: LIME demonstrates high stability with a cosine similarity of 0.9964, suggesting that its explanations are consistent across similar instances.

LIME Sparsity: On average, LIME uses 24 non-zero features, suggesting moderate sparsity in the explanations.

SHAP Stability: SHAP exhibits exceptional stability, with a cosine similarity of 0.9998, signifying almost identical explanations across similar instances.

SHAP Sparsity: Similar to LIME, SHAP explanations also use an average of 24 non-zero features, reflecting consistency in sparsity between both methods.

SHAP Contrastivity (instance 0): For instance 0, SHAP achieves a contrastivity proxy of 0.8719, indicating a strong contrast between the predicted and baseline classes, enhancing the interpretability of the explanation.

5.5.6 Statistical Analysis

To assess the significance of the model performance differences, we conducted a Wilcoxon Signed-Rank Test, comparing all baseline classifiers against XGBoost across the metrics reported in Table 5.5. The test results indicate statistically significant differences (p ¡ 0.05),

confirming that XGBoost consistently outperformed the other classifiers. Furthermore, one-way ANOVA was applied to evaluate the proposed Hybrid Threshold method in the context of interpretability. The resulting p-values were extremely small (e.g., 8.54×10^{-7}), well below the conventional significance level of 0.05. This allows us to reject the null hypothesis, indicating that the means of the performance metrics differ significantly across the models. These findings underscore the impact of the model and threshold selection on overall performance and highlight the statistical robustness of the proposed Hybrid Threshold approach.

5.6 Results and Discussion

This section presents a comprehensive evaluation of the proposed Hybrid Threshold method integrated with the XGBoost classifier for credit card fraud detection. The performance of the model was assessed using several standard metrics: accuracy, precision, recall, F1 score, AUC-ROC, PR-AUC, specificity, sensitivity, and Matthews Correlation Coefficient (MCC). The Hybrid Threshold, computed as the average of the optimal thresholds derived from the AUC-ROC and precision-recall (PR-AUC) curves, was designed to achieve a balanced trade-off between precision and recall, which is critical in class-imbalanced settings such as fraud detection.

The experimental results demonstrate the effectiveness of the proposed approach. Specifically, the Hybrid Threshold method combined with XGBoost achieved high values for all evaluation metrics: accuracy (0.9109), precision (0.9017), recall (0.9224), F1 score (0.9119), AUC-ROC (0.9703), PR-AUC (0.9695), specificity (0.8994), sensitivity (0.9119), and MCC (0.8221). These results indicate that the model not only correctly identifies fraudulent transactions but also minimizes false positives and false negatives, enhancing the overall reliability and robustness of the detection system.

Compared to traditional thresholding techniques such as fixed thresholds (e.g., 0.5), the proposed Hybrid Threshold method yields significantly improved performance by adapting to the underlying data distribution. Traditional methods often fail to capture the optimal decision boundary in imbalanced datasets, leading to degraded recall or precision. In contrast, the hybrid approach leverages statistical characteristics from both the AUC-ROC and PR-AUC spaces to define a more effective threshold, which enhances classification stability and fairness.

Furthermore, interpretability was enhanced using the LIME and SHAP techniques. LIME offered instance-level explanations, allowing for localized interpretation of predictions, whereas SHAP provided global feature attribution, highlighting the most influential variables across the model. This integration of explainable AI tools ensures transparency in decision-making, increases user trust, and facilitates model validation by domain experts which is an essential

requirement in sensitive applications such as fraud detection.

By mitigating bias through careful threshold calibration, normalizing decision boundaries, and incorporating interpretability mechanisms, the proposed method contributes to a more transparent and fair fraud detection system. The model demonstrates strong resilience against adversarial variations and is well-suited for deployment in real-world financial applications where both performance and accountability are paramount.

5.7 Conclusion

This chapter builds on the complementarity of LIME and SHAP methods for model interpretability, which we combined using the Hybrid Threshold Method. We demonstrated that this approach is a robust solution for threshold selection in fraud detection, showing substantial improvements in accuracy, F1-score, AUC-ROC, and PR-AUC. The hybrid threshold method, which integrates precision recall curves of AUC-ROC and PR-AUC, outperforms traditional thresholding strategies, particularly by prioritizing the F1 score. This not only enhances predictive power but also ensures the reliability of decision-making processes. The method mitigates the impact of class imbalance, a critical challenge in fraud detection, by detecting fraudulent transactions with minimal false negatives while maintaining high precision.

In the context of data-driven industrial applications, where AI-driven decision-making is paramount, future work should focus on advancing the interpretation of deep learning models for fraud detection within industrial systems. Specifically, incorporating explainable AI (XAI) methods such as counterfactual explanations and attention-based mechanisms will further enhance model transparency, enabling stakeholders to better understand the underlying patterns of fraud. Additionally, integrating real-time adaptive thresholding techniques using reinforcement learning could dynamically adjust detection strategies in response to evolving fraud tactics, ensuring ongoing system adaptability. Multimodal fusion techniques that combine transaction data with behavioral analysis have significant potential to improve detection accuracy.

These advancements will push the boundaries of AI-driven fraud detection systems, ensuring robustness, fairness, and security in real-world applications. Ultimately, they will contribute to the progression towards fully transparent, adaptive, and reliable AI solutions within industrial environments, aligning with the overarching goal of this thesis to achieve trustworthiness and interpretability in data-driven AI models. In the following chapters, we will delve into more specific studies and applications.

Chapter 6

Evolution of AI-Driven Decision Making With Decision Support Systems, Expert Systems, Recommender Systems, and XAI

In the previous chapter, we explored a hybrid thresholding method for fraud detection that integrates Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) to enhance both interpretability and model performance. Building upon this foundation, the current chapter broadens the scope by examining the historical evolution and foundational components of AI-driven decision-making systems. Specifically, we explore four key paradigms. These are: Decision Support Systems (DSS), Expert Systems (ES), Recommender Systems (RS), and Explainable AI (XAI). This exploration aligns with the broader objectives of this thesis by tracing how these systems contribute to intelligent, transparent, and reliable decision-making in industrial contexts with SDP (Software Defect Prediction as Case study).

The chapter begins with an in-depth discussion of DSS, which offers interactive, data-driven tools to assist human decision-makers. It then examines ES, which emulate domain expertise to provide reasoning-based outputs, and RS, which support personalized recommendations through user behavior modeling. Finally, the chapter highlights the emergence and growing importance of XAI, which seeks to enhance the interpretability and accountability of complex AI models.

Through comparative analysis, we identify the strengths, limitations, and synergies among these systems, with a particular focus on their applicability to industrial applications. This

analysis serves as a foundation for designing next-generation, explainable decision-making architectures suitable for high-stakes industrial domains.

6.1 Introduction

DSS is a pioneer in the evolution of AI-based decision-making systems. The decision support systems are considered tools for high-level managerial decision-making. These are different from expert systems where the objective is to emulate a human expert and even perhaps manage with the human out of the loop. In decision support systems on the other hand human decision-making is to be augmented by a DSS. This approach is needed when humans are to take decisions but the matter requires a lot of data processing that is not possible or feasible in a short time without support from automation. If the critical nature of the decision becomes very important, a detailed explanation of the reasoning behind the decision is also very much needed. The modern approach to this is through the "explainable AI" or XAI as it is now known.

Artificial Intelligence (AI) systems have been developed to make decisions based on data that has been collected to emulate expert knowledge and reasoning skills. Many businesses now have adopted a data-driven approach to operational decision-making in recent years. Data may improve decisions, but getting the most from them requires an appropriate process. However, it is mistakenly believed that the process must emulate a human being. The phrase "data-driven" suggests that data is filtered and summarised for a human to review before being sent for processing.

In recent times with greater automation in decision-making, it is seen that to fully realize the value of data, businesses must integrate Artificial Intelligence (AI) into their workflows. Even then, the question remains whether we can get humans out of the way. In other words, is it possible to move away from data-driven workflows towards AI-driven workflows? this is an important question.

A data-driven workflow focuses on the stages of processing the data, while an AI-driven workflow attempts to build a learning model based on the preprocessed data. Processing data enables more accurate judgments and well-informed decision making by reducing complex information into a more manageable and relevant form. Jarvis, Amazon Alexa, and Hypersonix's AI-based intelligent assistants demonstrate the process of extracting data insights and putting them into action are explained in the paper [81]. AI assistants have some degree of decision-making capabilities. An AI assistant, for example, might employ decision-making algorithms to choose the optimal action based on the given assumptions. Also, decision-making systems might employ conversational interfaces to talk to users and convey their suggestions

in a more user-friendly manner. This system provides decision-makers with both prescriptive and predictive insights to help them make better data-driven decisions faster, as explained in the paper[266].

By automating data analysis, delivering user-friendly interfaces, utilizing AI algorithms, and providing contextual insights and recommendations, augmented analytics and intelligence engines provide advanced decision-making. These developments enable decision-makers to make data-driven decisions to improve corporate business and give them a competitive advantage. The study of the interaction between marketing channels and consumer decision-making has a long history in the DSS literature are explained in these papers [373],[119]. These studies have mostly focused on how customers evaluate, use, and adopt channels, as well as how the buying environment (e.g., risk, trust, and uncertainty) influences the adoption process. Although these studies help us understand how consumers make decisions on multiple channels, there is limited literature on how distinct channel features and channel choices across many channels can influence decision-making. In this work, we provide an overview of the tools for high-level decision-making and the evolution of these tools from DSS to XAI.

The rest of the chapter is organized as follows: Section 6.2 describes the individual decision systems in detail. Section 6.3 describes the expert systems. Section 6.4 describes the recommender systems. Section 6.5 describes the Explainable AI. Section 6.6 describes the case study of defect prediction, and we conclude in section 6.7.

6.2 Decision Support System (DSS)

[285] paper explains a Decision Support System (DSS) is a computer-based system that assists in making decisions. [224] paper explains DSS integrates data, models, and analytical tools to provide information and insights that aid decision-making processes. These systems focus on providing support and facilitating the decision-making process rather than emulating human expertise.

6.2.1 Components of a Decision Support System

Figure 6.1 shows the DSS Architecture proposed in these papers [284], [352].

- **Data Management**: The data management component of a decision support system handles the storage, organization, and retrieval of relevant data for analysis and decision-making processes.

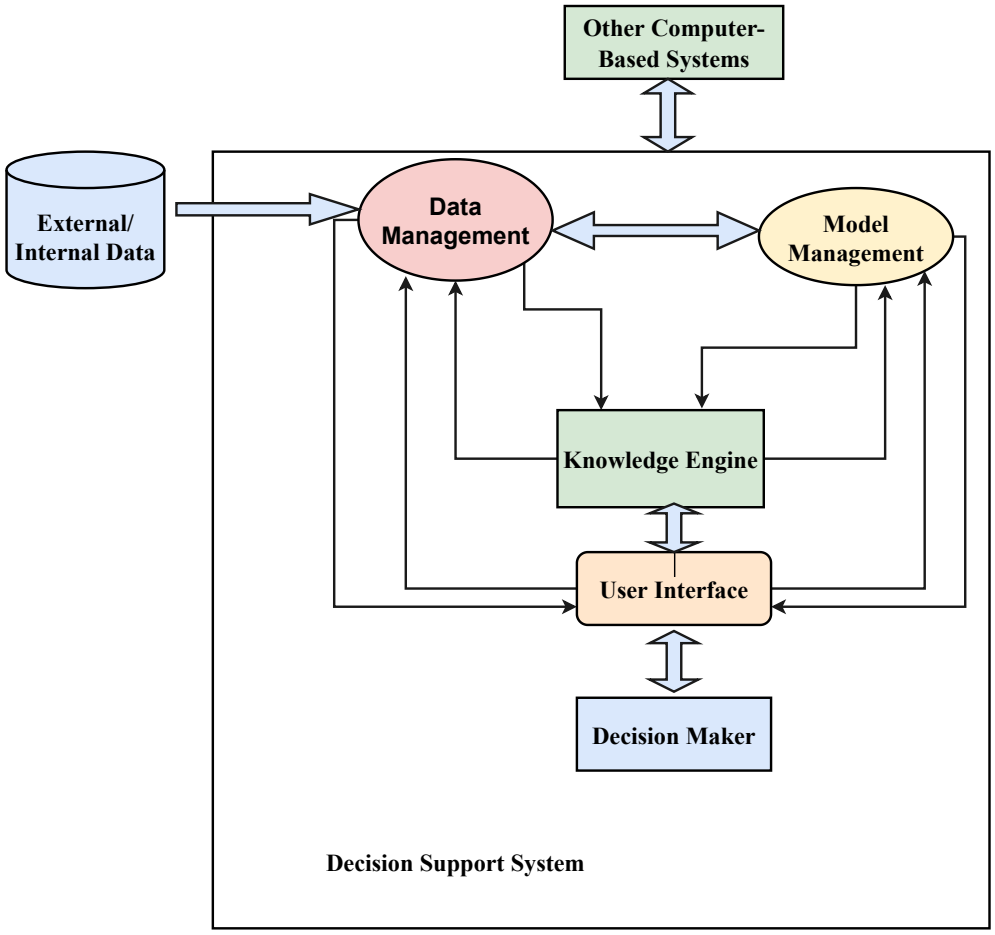


Figure 6.1: Decision Support Systems Architecture

- **Model Management**: The model management component focuses on maintaining and updating the models used in decision support, ensuring they are accurate, up-to-date, and aligned with the objectives of the system.
- **Knowledge engine**: The knowledge engine component incorporates expert knowledge, rules, algorithms, and inference mechanisms to facilitate intelligent reasoning and generate insights to support decision-making.
- **User interface**: The user interface component provides a user-friendly platform or interface through which users can interact with the decision support system, access information, input preferences, and receive results or recommendations.
- **Decision maker**: The decision-maker component utilizes the insights, recommenda-

tions, and information provided by the decision support system to aid decision-makers in evaluating alternatives, understanding implications, and ultimately making informed decisions.

Types of DSS	What it does	How it works
Data-Driven	Based on external and internal databases, it makes decisions.	Data mining (DM) and Machine Learning (ML) techniques are used to predict patterns and make decisions.
Model-Driven	It is utilized to meet a predetermined set of user needs.	Models are used to evaluate various scenarios and ensure they meet user needs.
Knowledge-Driven	Continuously manages updated information through knowledge management.	Information is provided to the user, incorporating the company's business procedures and knowledge base.
Communication-Driven	Tools enable communication among multiple individuals working on the same task.	It is utilized to increase the system's efficiency and efficacy by facilitating communication.
Document-Driven	Documents are used as the primary means of retrieving data in this type of information management system.	Users can search company websites or databases for policies and procedures using documents.

Table 6.1: Types of DSS

Table 6.1 elaborates on the types of DSS and its applications in different domains.

6.2.2 Limitations of Decision Support System

- **Dependence on Data Quality:** DSS heavily relies on the availability and accuracy of relevant data.
- **Sensitivity to Assumptions:** DSS often utilize models and algorithms based on assumptions, which can impact the accuracy of outcomes.
- **Lack of Human Intuition:** DSS may not fully capture qualitative factors and human intuition that are important in the decision-making process.
- **Narrow Focus:** DSS may be designed for specific tasks or domains, limiting their effectiveness in broader decision-making contexts.

- **Over-Reliance on Technology:** There is a risk of over-reliance on DSS outputs without critical evaluation and independent decision-making.
- **Resistance to Change:** Adoption of DSS may face resistance from individuals accustomed to traditional decision-making approaches.
- **Cost and Complexity:** Developing and maintaining DSS can be resource-intensive, requiring technical expertise and financial investment.
- **Ethical Considerations:** DSS may encounter ethical challenges, such as fairness, bias, and privacy concerns that need to be addressed.

In order to address the challenges faced by Decision Support Systems (DSS) and enhance their component functionalities, Expert Systems (ES) have been introduced. Both DSS and ES operate within similar domains and share similar working principles, thereby allowing for a seamless transition from DSS to ES. The intersection of DSS and ES is described in Table 6.2, and further details on the Expert System are presented in Section 6.4.

Domain name	**DSS**	**ES**
Corporate Financial Planning	Loan repayment schedule, depreciation, and break-even analysis explain in the paper [120].	ES built for currency exchange in international business transactions explained in the paper [259].
Market Analysis	Forecasting, sales analysis, and promotion analysis explained in the paper [120].	ES used for control of client relationships, market research, and product planning explained in the paper [241].
Health care	DSS advises health workers and patients on specific problems explained in the paper [253].	ES used for heart failure telemonitoring system explain in the paper [334].
Transportation	DSS provides information during emergency response, recovery effectiveness, and alerts the responders explained in the paper [408].	ES used for autonomous cars and DARPA challenges explain in the paper [33].
Real-Estate Investments	Financing alternatives, cash flows, and impact on taxes explained in the paper [120].	ES built for selection and evaluation of floor finishing materials explained in the paper [218].

Table 6.2: Intersection of DSS and ES

Based on the challenges and metrics of these domains, the evolution goes from Decision Support Systems (DSS) to Expert Systems (ES). The following section describes Expert Systems.

6.3 Expert Systems (ES)

[127] paper explained the Expert System is a computer system that mimics a human in decision-making. It uses a knowledge base, which contains a set of rules and facts, along with an inference engine to reason and make decisions. Expert systems are built by capturing the knowledge and expertise of human experts and encoding it into a computer program. They excel at providing specialized knowledge and advice within a specific domain, unlike DSS which can be applied to various tasks.

6.3.1 Components of Expert System

Fig 6.2 shows a general ES architecture with various connected components explained in these papers [273], [306].

- **Inference Engine**: This component searches the knowledge base for new information and infers new information using various methods. There are two approaches used in the inference engine: Forward chaining and Backward chaining.

- **Knowledge Acquisition and Learning Module**: The ES's information comes from a variety of places, including textbooks, human experts, reports, presentations, movies, and the internet. This module enables the system to gather additional information about the problem and store it in the knowledge base.

- **User Interface**: It allows users to communicate with the ES through an interface, which aids in the system's knowledge production and querying for the task at hand.

- **Explanation Module**: It allows the user to interrogate the ES about how and why it came to the result.

- **Special Interface**: This module is used to carry out a task such as dealing with ambiguous information. It deals with ambiguous data and information.

- **Case History**: It is used to extend the knowledge base of a learning module by storing files created by an inference engine utilizing a dynamic database.

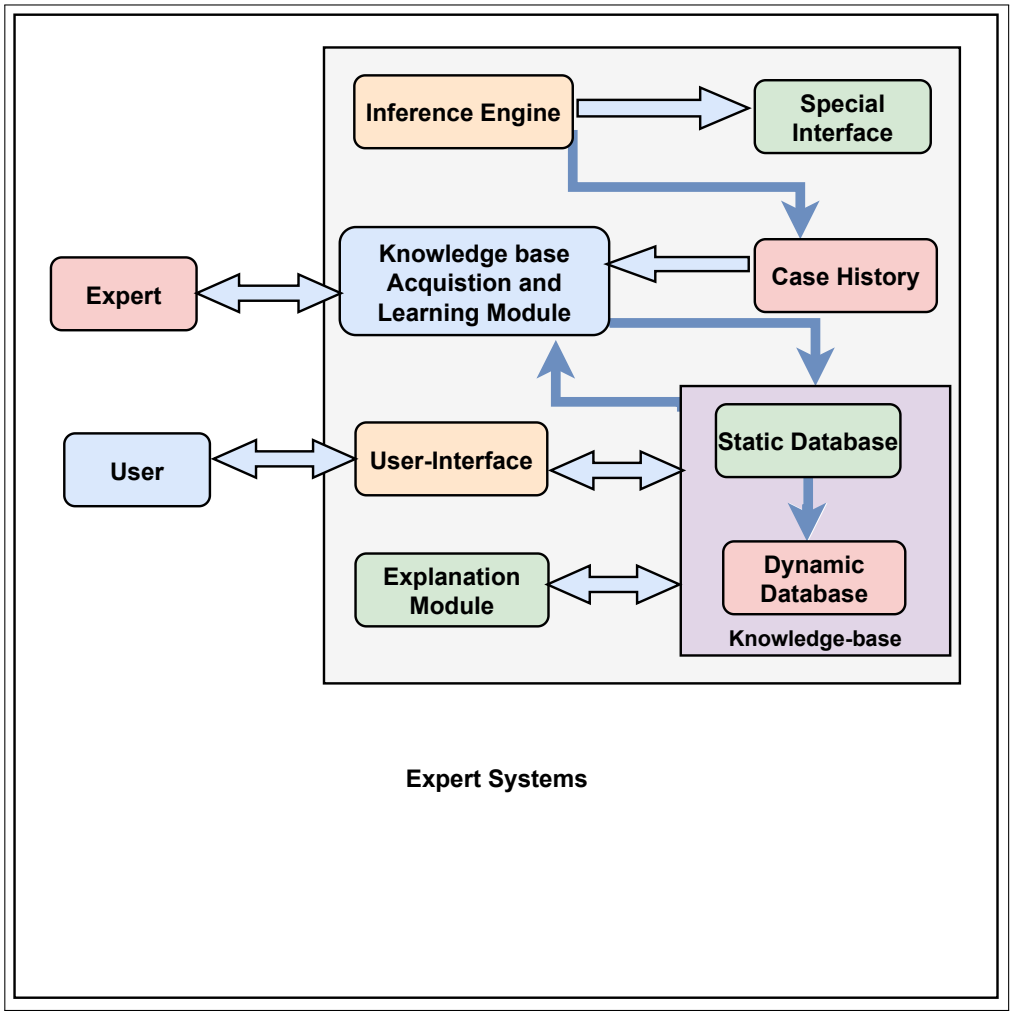


Figure 6.2: The general ES architecture explained in these papers [273], [306].

- **Knowledge Base**: An ES's knowledge base is made up of information about the problem. A static and dynamic database serves as the memory. Rules and facts make up static knowledge. Dynamic knowledge is made up of detailed material and a series of queries posted to a human expert on the subject. In practice, all of the correct answers are stored in a dynamic database.

According to the paper [256], based on the various parameters and components like knowledge base, knowledge representation, adaptability, maintenance, inference engine, and explanation module, Expert Systems can be divided into five categories: *Rule-based ES*, *Frame-based ES*, *Fuzzy-based ES*, *Neural ES*, and *Neuro-fuzzy ES*.

As expert systems developed into decision support systems, they shifted from simulating

human competence to offering data-driven analytical tools and insights for making decisions. In order to increase their capabilities and adaptability, decision support systems have included advanced analytics and AI approaches, making them more versatile and efficient in supporting decision-makers in multiple domains. Expert and decision support system integration may improve the effectiveness of both computerized systems. The following passage describes the conceptual view of ES and DSS.

6.3.2 Conceptual View of DSS and ES

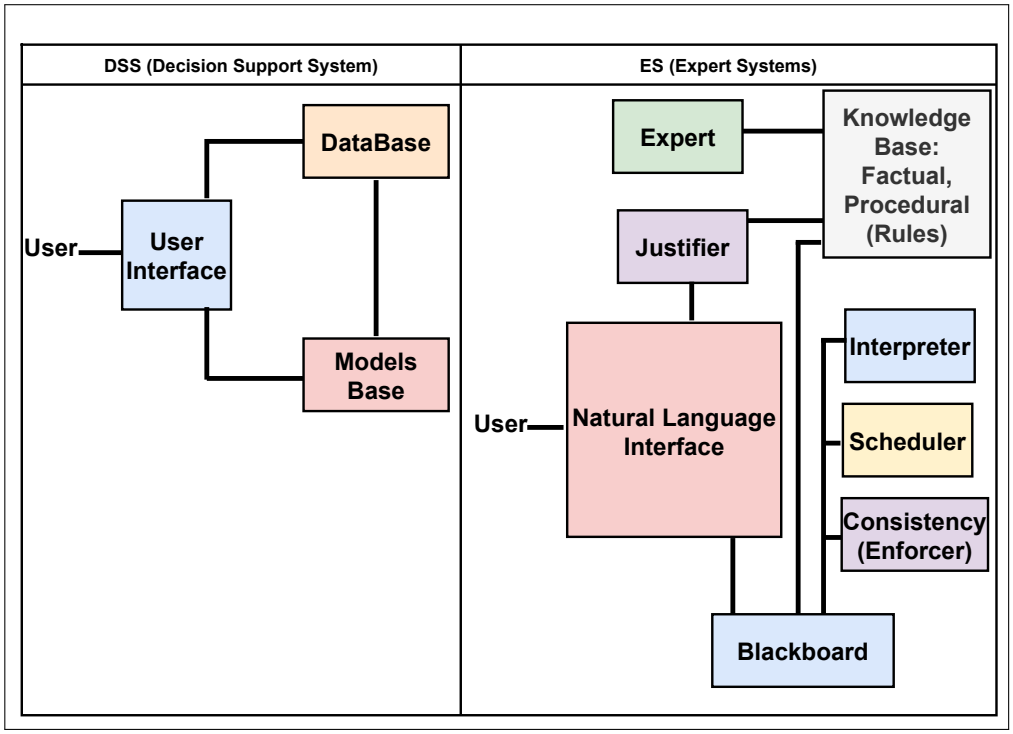


Figure 6.3: Conceptual View of DSS and ES explained in the paper [371].

Fig 6.3 describes the conceptual view of the ES and DSS. The Database, Model-base, and User-Interface are all contained within the DSS. The database houses the firm's "factual" industry's knowledge (internal or external data). The Model-base is a collection of management science, statistical, economic, and industry models. The model base/database is linked to the User-Interface, which displays the results to the user. ES contains the Expert, Knowledge-base, and Justifier. Interpreter, Scheduler, Consistency (Enforcer), and Blackboard are the components of the Natural Language Interface. The knowledge base comprises both factual and procedural information on the field in question. Interpreter, Scheduler, and Consistency (Enforcer)

are tools for controlling how a user's conversation with the system is processed. Choosing an agenda for processing, executing the chosen agenda by applying the relevant procedural knowledge with the help of an Expert, and attempting to maintain consistency in the representation of the developing solution are all reasons for which these systems are constructed. The devices used to record intermediate judgments that the ES manipulates are known as black-board devices. The Justifier is a key aspect of ES since it gives the user a justification for the system's behavior. The language processor is employed to allow the user to engage in problem-solving dialogue with the ES. As mentioned above, the features are based on idealized ES. The following Table 6.3 describes the differences between ES and DSS.

Attributes	**DSS**	**ES**
Objective	Assist human	Replacing a human by replicating (mimicking) him or her.
Who is the Decision maker?	The human	The system
The primary focus	Making Decisions	Expertise Exchange
Direction of Query	The machine is questioned by a human	A human is questioned by the machine.
Clients	Group users or Single user	Single user
Manipulation	Numerical types	Symbolic Types
Area of Problem	Integrated, Wide, Complex	Confined domain
Data-Base	Factual and Heuristic Knowledge	Procedural and Factual Knowledge
Group composition	Personal, Institutional, and Group	Personal and Group

Table 6.3: Differences between ES and DSS

Expert systems have many advantages, but they also have certain challenges. Here we describe a few challenges in the following way:

- **Acquisition of Knowledge:** It can be a challenging and time-consuming procedure to get knowledge from domain experts. Experts may possess implicit knowledge that is difficult to express, and it can be difficult to convert this knowledge into a formal knowledge representation.

- **Knowledge Representation** Representing acquired knowledge in a way that the expert system can successfully use it. The selected representation should be able to convey the intricacies and complexity of the domain while being understandable and manageable for the system.

- **Maintenance of Knowledge:** As domain knowledge develops, expert systems need to be regularly updated and maintained. To keep the system current and correct, regular changes are required. It can be difficult and resource-intensive to manage these upgrades and incorporate new information into the system.
- **Expertise and Limited Scope:** Expert systems are frequently created to address particular areas of concern and may not have the adaptability needed to handle novel or unusual circumstances. They do well in their areas of expertise but may fail when dealing with issues that fall outside of those areas.
- **Lack of Transparency and Explanation:** Expert systems frequently base their choices on complex rules and inference techniques, which can make it difficult to communicate their logic to end users. Lack of transparency can cause user mistrust and prevent the system from being used more widely.
- **User Adoption and Acceptance:** Expert system adoption and trust among end users, including domain experts, can be difficult to achieve. Expert systems may have difficulty being accepted and used effectively due to human issues of resistance to change, a lack of familiarity, and skepticism regarding machine-based decision making.

Many expert systems applications fail simply because management failed to do a sober, false assessment of their problem's eligibility for an expert system due to momentum, enthusiasm, competitive pressure, or need. Even projects that are viable could fail due to expensive expenses, long development delays, insufficient validation, or the realization that the finished product is useless. Companies are not ready to encourage full-time employees to train and upgrade the skills for ES explained in the paper [84]. So the evolution occurs from ES to Recommender Systems (RS) due to failing of ES, DSS components getting merging with RS.

6.4 Recommender Systems (RS)

The concept of the Recommender System was introduced in the early 1990s to help human users identify (suggest) the most useful and interesting products. Collaborative filtering was the first terminology used in the field of recommender systems. The term Collaborative filtering was introduced in the paper [143]. In 1992 Xerox PARC's Tapestry developed the project using the name "Recommender Systems". Later, Resnick and Varian used the term "Recommender Systems" in 1997 explained in the paper [309]. When large data sets became accessible in 2006, and the Netflix Prize explained in the paper [40] was announced for improved prediction accuracy, research in recommendation systems skyrocketed. The prize enabled the first summer school to be held the same year, as well as the inaugural ACM Recommender Systems

conference – RecSys – in 2007. The concept behind recommender systems is that they forecast recommendations based on user preferences, allowing them to select items from a large number of web applications.

Recommender systems are designed to predict a user's interests and propose things that are likely to interest them. These are among the most advanced machine learning algorithms used by online retailers to boost sales. Companies that use recommender systems aim to boost sales by providing more tailored offers and a better customer experience. The recommendation usually speeds up searches and makes it simpler for consumers to find the information they're interested in, as well as surprising them with offers they wouldn't have looked for. In Fig. 6.4,

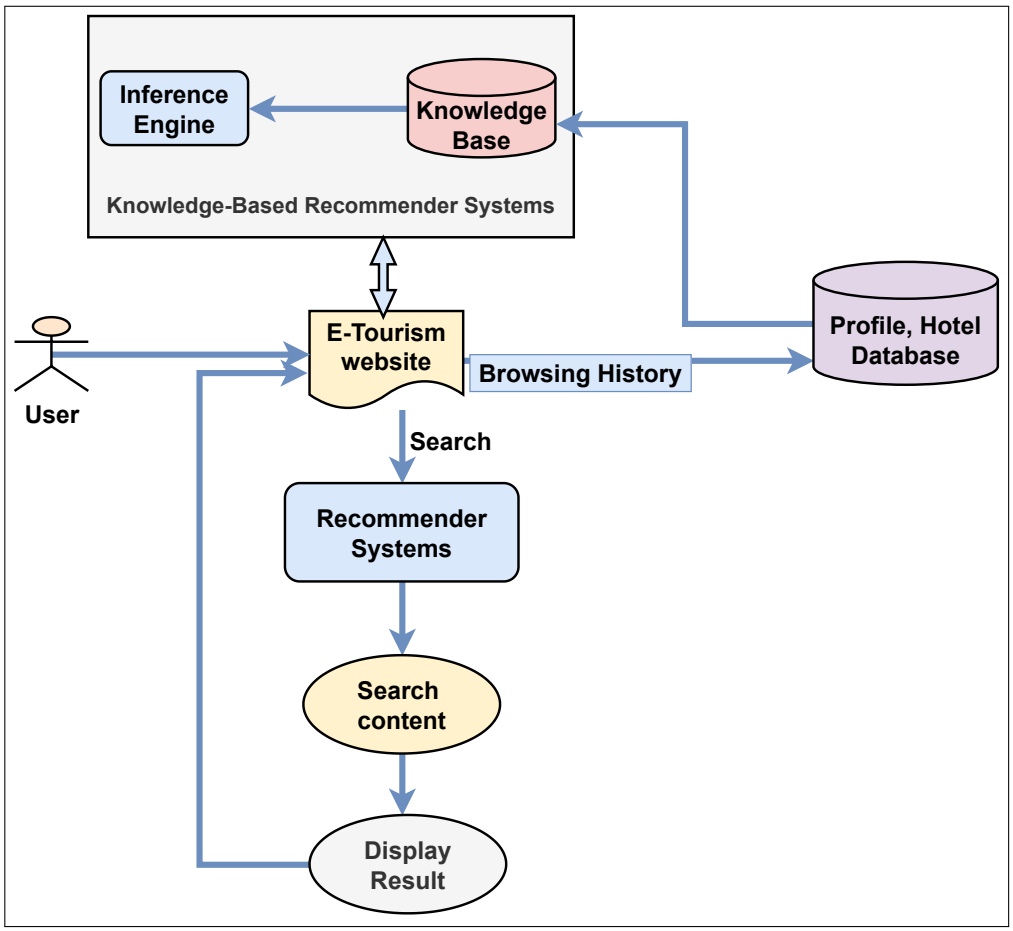


Figure 6.4: General framework of Recommender Systems explained in the paper [294].

the general framework of Recommender Systems (RS) is presented. In the following, we discuss the general framework of RS.

For example, a user plans to visit a tourist place and searches E-tourism websites according

to his interest. The Recommender System filters the data and recommends options to the user based on his preferences and displays the results. If the user browses the hotels' profiles in the database, the system communicates with the knowledge base and the inference engine of the knowledge-based recommender system to provide specific information to the user.

Recommender systems work with two kinds of information, described below:

- **Characteristics Information:** Information about items (keywords, categories, etc.) and users (preferences, profiles, etc.).
- **User-item Interactions:** Information such as ratings, number of purchases, likes, etc. Interactions can be defined as either explicit or implicit.
  - **Explicit:** When a person expresses either positive or negative interest in an item, such as by rating it or writing a review.
  - **Implicit:** When a user's interest is inferred from their actions, such as browsing or purchasing an item.

  The better the outcomes, the more interactions per user and item.

Based on these, RS can be classified into seven types: *Content-Based RS*, *Collaborative RS*, *Demographic RS*, *Hybrid-Based RS*, *Knowledge-Based RS*, *Context-Aware RS*, and *Session-Based RS*.

Here Table 6.4 describes the Types of RS and their working process.

6.4.1 Limitations and challenges of RS

RS has some limitations like lack of data, changing user preferences, unpredictable items, and changing data. To build the RS it is most important to understand the limitations.

- **Cold start problem:** It occurs when a new user is added to the site or new items are added to the systems. In the first scenario, we didn't know what to recommend to the new user because we didn't know what he was interested in. He had never given any goods a rating before. In the second scenario, we can promote a new item to others even if no one has rated it as good or bad by consumers, which is explained in these papers [337], [170].
- **Sparsity:** Sparsity problems occur when the user watches movies or buys products and did not rate those movies or products. It occurs for a large user-product matrix. Rating-product matrix recommender systems recommend products to others, these papers explain them [337], [170].

Table 6.4: Types of RS

Types of RS	What it does	How it works
Content-Based	Used to describe the item and taste of the user profile explained in the paper [94].	Information provided by the user, either explicitly (ratings) or implicitly (search results, clicking on a link). A user profile is created using that information and is then utilized to offer recommendations to the user.
Collaborative-Based	Used to capture and analyze user activity in the form of feedback, ratings, preferences, and actions are explained in the paper [94].	The method of removing items that a user might enjoy based on the reactions of other users. It works by sifting through a large group of people to locale a smaller group of users with similar likes to a specific user.
Demographic-Based	Used to employ user profile information such as age, gender, demographic area, education, interests, and their opinion about rating things to locate common users who have comparable ratings are explained in the paper [169].	Based on demographic attributes. Demographic RS distinguishes the users based on their attributes and recommends the movies by utilizing their demographic data are explained in the paper [275].
Hybrid-Based	A combination of both CBRS and CFRS that reduces the issues and challenges of the applications are explained in the paper [184].	Using CBRS or CFRS alone does not provide high performance and does not address the concerns and challenges effectively.
Knowledge-Based	It is a form of RS that is built on explicit knowledge of the item selection, user preferences, and suggestion criteria.	Based on specific knowledge of the item assortment, user preferences, and criteria for making recommendations are explained in the paper [61].
Context-Aware	To provide individualized services, it uses sensing and analysis of the user context to increase the accuracy of the recommendations.	The user context is sensed and analyzed to deliver customized services. Sensors can be used to gather contextual data, which helps improve the accuracy of the recommendations are explained in the paper [7].
Session-Based	To predict the next given item (product) or a sequence of previous items (products) consumed in the session.	For accurate and timely recommendations, it records both a user's short-term preferences and dynamic changes from one session to the next are explained in the paper [380].

- **Scalability:** Scalability is a metric that measures a system's capacity to run efficiently with high performance while expanding its data are explained in the paper [198].

- **Privacy:** In RS, privacy is one of the most crucial factors. When RS provides suggestions based on a user's preferences, we need to know certain information about the user's preferences. Users must be aware of what data is required in order to recommend the best products to them, as well as how they will be used, is explained in the paper [218].

- **Shilling Attacks:** It happens when a malevolent person tampers with the system's design and assigns fraudulent ratings to certain things, either to boost or decrease popularity, are explained in these papers [184], [170].

Several challenges in the ES are listed below: Cold start problem, sparsity, scalability, over-specialization, diversity, novelty, serendipity, privacy, shilling attacks, and gray sheep. Here, the quality of an item or product is a major issue in the RS. In general, the quality of items is evaluated by the RS applications using ML are explained in the paper [283], DL techniques and algorithmic approaches are better works in some cases. Thus, evaluation factors (metrics) are user-centric approaches and take into consideration factors that impact the user's satisfaction and motivate (creating interest) them to make the decision related to recommendation (purchase, listen, watch) are explained in the paper [327]. In this scenario, a good RS should be transparent in one approach to explain why items (products) are recommended to the active user. Moving beyond accuracy measures to create and evaluate explainable RS is a promising, yet difficult, search area. It is challenging to evaluate RS results because most RS is based on ML and DL techniques that are opaque by nature. Fortunately, XAI are explained in the paper [5], a new branch of research, proposes ways for making AI systems' results more human-friendly. We recommend XAI then that the results of this field be used, adapted, and projected upon RS to make it more transparent and trustworthy. In the next section, we explore the emerging XAI technology.

6.4.2 Evaluation of Recommender System

There are various types of RS we should evaluate based on their performance are explained in the paper [8]. Mainly RS metrics are MAE (Mean Absolute Error), RMSE (Root Mean Absolute Error), HR (Hit Rate), ARHR (Average Reciprocal Hit Rate), CHR (Cumulative Hit Rate), RHR (Rating Hit Rate), Novelity, Diversity, and coverage. The following section describes the XAI.

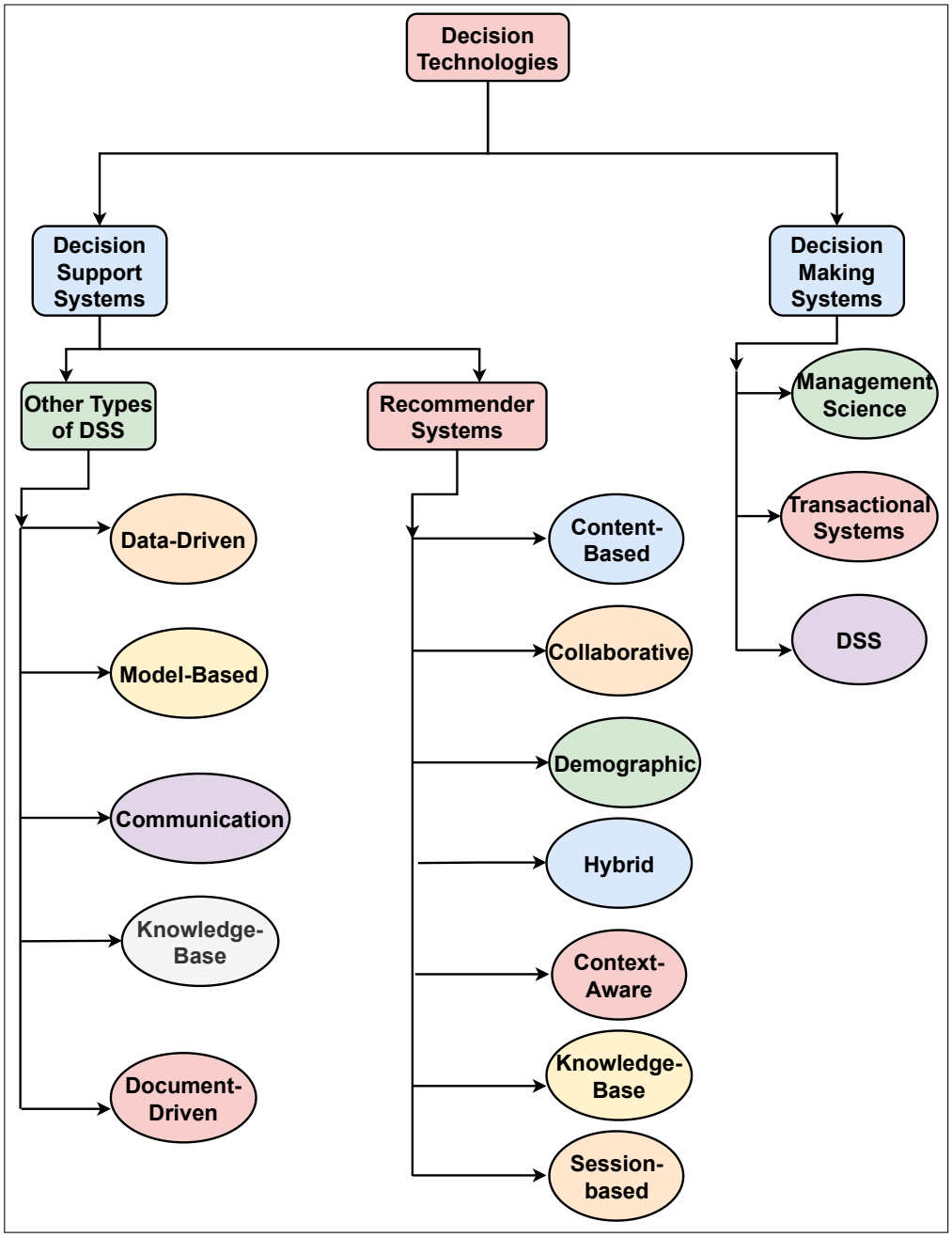


Figure 6.5: Flow chart of Decision Support Systems and RS

6.4.3 How DSS are now merging with RS

Fig 6.5 describes the DSS merging with RS. Decision Support Systems (DSS) were developed as a complement to computerized choice-making systems. For circumstances when human input was required, since the decision situation was not highly structured (programmable), therefore, required human judgment and intuition are explained in the paper [40]. The computer's job was to give information in the form of databases or models to aid in decision-making explained in the paper [23]. Because of the abundance of information and the speed with which decisions must be made, a decision scenario might become unstructured as a result of the new information economy. Users require computer assistance not just due to the complexity of a particular decision (the original DSS notion), but also due to a lack of processing speed. This category of decision support includes recommendation systems, which analyze options. Typical applications of recommendation systems include suggesting information to a decision-maker (filtering systems) and ranking films, restaurants, books, and other items. In some ways, these are minor decisions, yet they add up. However, we do not want to limit recommendation systems to low- or medium-importance issues for individuals or organizations. As shown below, systems can be imagined to assist users in making big decisions, such as purchasing equipment for companies and purchasing home or automobiles for individuals. To produce their recommendations, recommendation systems may employ all or any information from any source, as well as a variety of processing techniques. On the other hand, recommender systems (limited definition) fulfill the same function as recommendation systems, but with the addition of human recommenders. Recommendants and recipients may not know one other, they may or may not collaborate overtly with the recipients of the information explained in the paper [309].

6.5 Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) is an emerging research topic that aims to give end-users intelligible AI outcomes are explained in the paper [5]. Explainable AI (XAI) also known as Interpretable or Understandable AI, aims to solve the black-box problem in AI. XAI approaches, in particular, aim to develop ML/DL techniques to give intelligible, trustworthy, and explainable rationales for black-box model decisions are explained in the papers [151], [110], [22]. XAI is sometimes also called a combination of Interpretability and Explainability. Interpretability describes the extent to which cause and effect can be observed within the system and Information representation for experts. Explainability describes the extent to which the internal mechanics of a machine or Deep Learning (DL) system can be explained in human terms and from An expert's point of view for decision-makers. XAI clarifies the logic for the

decision-making process, identifies the strengths and flaws of the method, and forecasts how the system will perform in the future are explained in the paper [295]. Building on the architectural foundations presented in Chapter 3, this section explores the role of XAI in decision support applications. We focus on evaluation metrics, practical implementations, and domain-specific impacts, particularly in scenarios involving Recommender Systems and Expert Systems.

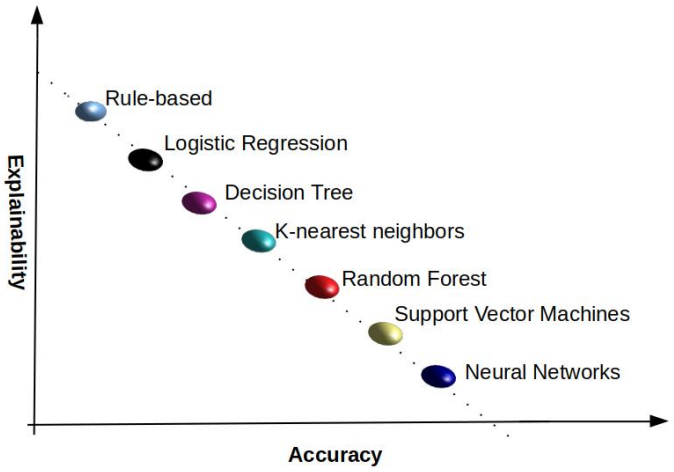


Figure 6.6: Explainability are explained in the paper [29].

Fig 6.6 describes various algorithms present in the Explainability. There are two dimensions to machine learning (ML) models as we can see on the graph here, one is accuracy and the other one is explainability. The loss function can be used as an indicator of model accuracy. And that works very well but loss function or model accuracy doesn't explain how fair or responsible our ML model is you cannot trust the ML model prediction just based on the accuracy of the model. We need a second dimension and that is where explainability comes into picture, which is explained in the paper [29]. Explainability is important to establish trust between humans and ML systems. It is also important or becoming important because of increasing regulatory compliance around the right to explanation. In 2018, the GDPR Act was enacted and we started talking about the privacy of data, but there is a less popular article in GDPR that talks about the right to explanation. Now, the right to explanation is only defined in principle for GDPR because it is precautionary, but there are countries like Singapore that

not only define the right to explanation in principle but also use or have developed a framework for enterprise companies to adopt explainability into ML systems. The interesting part is accuracy and explainability as you can see on the graph here are negatively correlated. What it means that in the top left you will see ML models like rule-based and logistic regression, which have low to moderate accuracy, but have high, Intrinsic explainability built in, and on the other hand, on the bottom right, you will see complex models like support vector machines (SVM) and neural networks which have very high accuracy which is desired. But are almost black-box and are not explainable so with that understanding.

6.5.1 Explainability Methods:

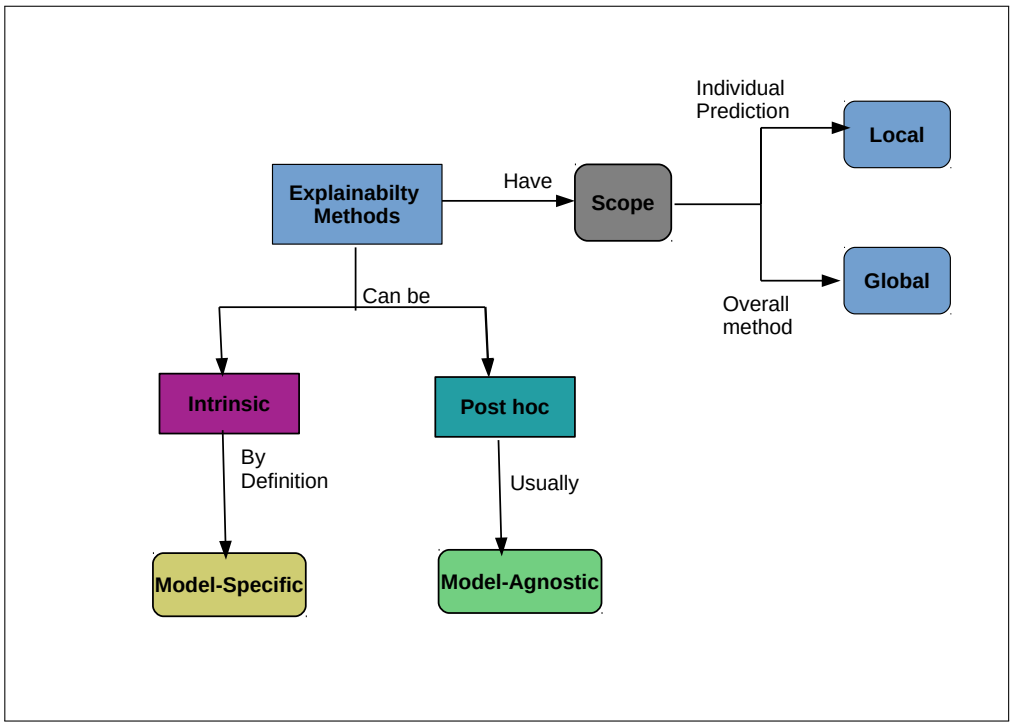


Figure 6.7: Explainability Methods are explained in the paper [5].

Fig 6.7 describes the various methods present in Explainability. Explainability methods in one way can be categorized according to the scope. We can have a local scope or a global scope. Local scope means something that we were trying to do come with an explanation, for example a specific credit card transaction. To identify why it was declined, and why it was tagged as fraudulent that is local scope but maybe data scientists and ML engineers are something interested in learning overall across multiple examples, which particular features contribute the most to the model prediction score. And that is where we take into account the

global scope. And we usually use the averages across the training or the validation examples. The other dimension that can be used to categorize explainability methods, is method could be intrinsic. What it means is that by nature or the way the information is represented. Inside the models that can be used to explain or decode the cause and effect observed in the system. Now this type of method is model specific because they take into account the information representation on the other hand we have methods where we don't care about explainability until at the time of the model training and after the model training, we try to come up with a point of view to explain the ML reasoning behind the predictions made up that model and these models are called post hoc. And it can be categorized into model-agnostic method are explained in the paper [5].

6.5.2 Several definitions of XAI

D. Gunning defines XAI as "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners are explained in the paper [151]." Dwivedi et al. defines XAI as "Explainability is the ability to explain the reasoning behind a particular decision, classification, or forecast are explained in the paper [114]."
Sicular et al. define XAI as "Explainable AI is a set of capabilities that describes a model, highlights its strengths and weaknesses, predicts its likely behavior, and identifies any potential biases. It can articulate the decisions of a descriptive, predictive or prescriptive model to enable accuracy, fairness, accountability, stability, and transparency in algorithmic decision-making are explained in the paper [340]."

6.5.3 Various aspects of XAI

For a better understanding of XAI is familiar with a few concepts like understandability, comprehensibility, traceability, and auditability. XAI terms are more focused on Explainability, Transparency, and Interpretability. In the below, we explain each term in detail.

- **Traceability:** The capacity to track all processes from raw material purchase to manufacture, consumption, and disposal to determine "when and where the product was made by whom" is known as traceability.
- **Comprehensibility:** When we used the ML/DL models, comprehensibility defines the ability to learn the learning algorithms in a human-understandable manner. It is used for the evaluation of model complexity.

- **Auditability:** Auditability here is referring to the ability of an auditor to successfully obtain results in the financial sector. Here XAI is used to audit the company's financial reporting.
- **Understandability:** It is the feature of a model to be better understood by humans. How the model and internal architecture work, according to their algorithms. It is further divided into two types, such as model understandability and human understandability.
- **Explainability:** Explainability's aim is that an ML/DL model and its output can be explained in such a manner that "makes sense" to a human being at the acceptance level. Explainability is a powerful tool for detecting model faults and data biases, resulting in greater trust among all users. It can aid in the verification of predictions, the improvement of models, and the discovery of new insights into the topic at hand. Understanding what the model does and why it makes its predictions makes detecting biases in the model or dataset easier.
- **Interpretability:** Doshi-Velez et al. and Kim et al. defined that "the ability to explain or show anything to a human in a way that they can understand." The assumption is that the more interpretable an ML/DL system is, the easier it is to distinguish cause-and-effect links within its inputs and outputs are explained in the paper [197].
- **Transparency:** A model is said to be transparent in it can be understood on its own. A model can have various levels of understandability and transparency. It can be divided into three types; simulatable models, decomposable models, and algorithmically transparent models are explained in the paper [207].

In the above-mentioned definitions, In XAI understandability is the most important concept, this concept is strongly connected to transparency and interpretability. Here Transparency refers to a model's ability to stand on its own and human understanding. Here Understandability refers to the degree to which a human can understand a model's choice. In the same way that understandability is connected to comprehensibility, comprehensibility relies on the audience's ability to understand the knowledge of the model. The following passage describes the XAI framework in detail.

We show how an XAI system is set up to allow users to select a piece of matter that speaks to communicate within a time frame in Fig 6.8 Users can then choose from a menu of questions to ask the entity about the current time point. The conversation (dialogue) manager guides the system's response, first gathering pertinent data using the reasoner and then providing English responses using the natural language generator (NLG). NLG uses data from databases and

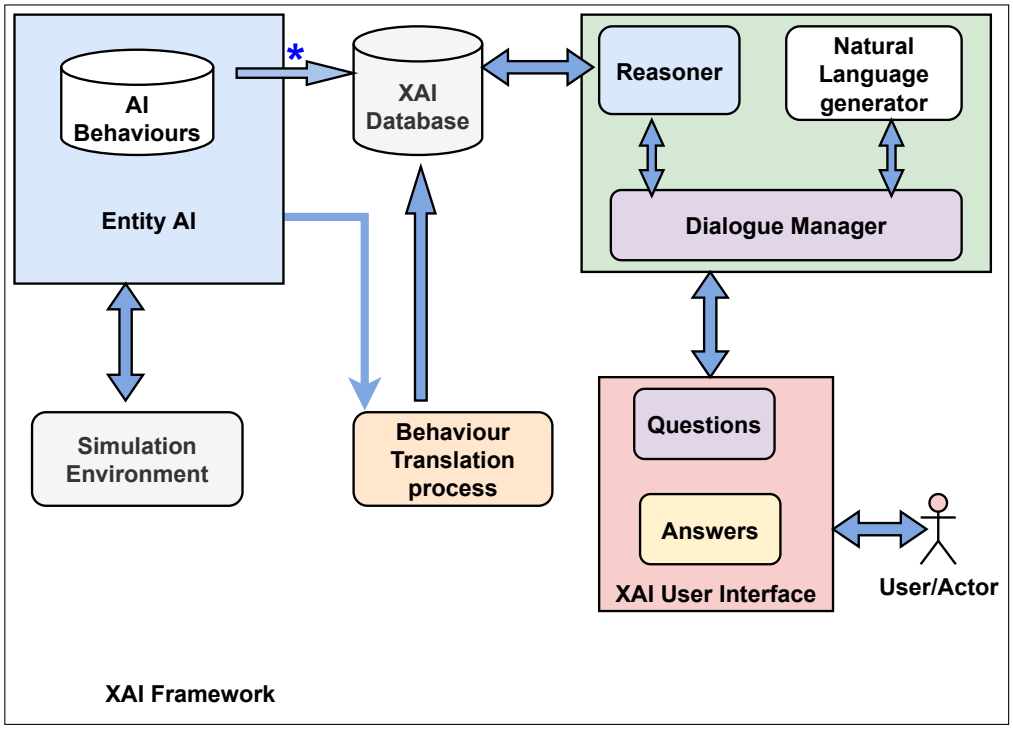


Figure 6.8: XAI frame work, Here The marker '*' represents the scenarios and logs are explained in these papers [88], [306].

human interaction to generate natural language templates are explained in the paper [88]. There are two types of ML algorithms used in XAI: white box and black box ML algorithms.

6.5.4 Characteristics of XAI

There are several characteristics of XAI we define some of them below.

- **Trust:** The expectation that the corresponding predictive algorithms for explanations should have certain performance.
- **Fidelity:** Expectation that the explanation and the predictive model algorithms align with one another.
- **Domain Sense:** The explanation should be sense in the domain of application and to the user of the system.
- **Parsimony:** The explanation should be as simple as possible.
- **Consistency:** The explanation should be consistent across different models and across different runs of the model.

- **Generalizability:** The explanation should be generalizable.

6.5.5 How XAI is Overcoming issues of RS

There are two main reasons XAI is overcoming issues of RS are following below. i) To explain why a recommendation is produced from the user's perspective. This helps to believe the users to accept the results. As a result, it contributes to the system's increased credibility and satisfaction, resulting in more consumer loyalty.
ii) Interpretability helps developers understand and debug the system from the perspective of modeling. This increases its scrutability and effectiveness are explained in the paper [327].
Due to these reasons, it is difficult to achieve explainability. Based on ML/DL, deep neural network algorithms are used in RS. To improve the performance, prediction, quality and transparency of the items using XAI.

6.5.6 Domains of Intersection of RS and XAI

Here Table 6.5 describes the intersection RS and XAI.

6.5.7 Business benefits

XAI is used mainly to strengthen stakeholder confidence, improve their performance, make better use of AI, and to further develop. The greater the confidence in the AI, the more quickly and extensively it may be used. Your company will also be better positioned to stimulate innovation and stay ahead of the competition in terms of developing and implementing next-generation capabilities. There are Eight main benefits of XAI are described below.

- **Model performance:** It's utilized to figure out where the model's possible flaws are. Easier it is to conduct models with a better grasp of what they do and why they occasionally fail. For example, DeepMinds was able to improve and optimize AlphaGo's (the machine that famously beat the world's best Go player) model and create success. Nowadays the system was working like a black box. For that reason, the Explainability tool is used for spotting the errors in the model and to improve the performance of the model.

- **Decision making:** Machine learning applications in business are primarily used to make automated decisions. However, we frequently want to employ models for analytical purposes. For example, utilizing data on location, operating hours, weather, time of year, products carried, outlet size, and so on, you could train a model to forecast shop sales throughout a major retail chain. The program would allow you to forecast sales

Domain Name	RS	XAI
Corporate Financial Planning	The RS is used to assist consumers in locating better products, financial plans, and other related information are explained in the paper [279].	Money laundering, transaction data analysis are explained in the paper [374].
Market Analysis	RS that can learn consumers' purchasing patterns and predict their purchasing orders are explained in the paper [200].	Predicting stock market crises are explained in the paper [262].
Healthcare	RS used to recommend drugs, predict health status, provide healthcare services, and suggest healthcare professionals are explained in the paper [368].	Detecting diseases using blood images are explained in the paper [231].
Transportation	RS are used to recommend routes, estimate trip costs, consider weather conditions, and factor in driving variables are explained in the paper[359].	Intrusion detection and security measures in smart cities are explained in the paper [288].
Real-Estate Investments	RS used to recommend land, prices, and sizes are explained in the paper [423].	Predicting residential market values are explained in the paper [190].
Defense	Detecting war attacks and enhancing defense capabilities are explained in the paper [67].	XAI techniques used to control Lethal Autonomous Weapon Systems (LAWS) are explained in the paper [384].

Table 6.5: Intersection of RS and XAI

at my stores on any given day of the year and in various weather conditions. Building an explainable model, on the other hand, allows you to determine what the major sales drivers are and use that information to increase revenue.

- **Control:** To get from proof of concept to full implementation, you must be certain that your system meets certain planned requirements, does not violate intended requirements and does not exhibit any unwanted behaviors. If the system makes a mistake, the organization must be able to recognize that something is wrong and take corrective action, if necessary, or even shut down the AI systems. By monitoring performance, highlighting faults, and offering a means to switch the system off, XAI may help your organization maintain control over AI. In terms of data privacy, XAI can assist in ensuring that only approved data is utilized for an agreed-upon purpose, and that data can be deleted if necessary.

- **Safety:** Due to several constraints on the safety and security of AI systems, they are more powerful and widespread. Sometimes these systems are also traced by hackers, due to unethical design, engineering oversights, and the effect on the environment. XAI can assist in the detection of such flaws. It's also linked to cyber detection and protection teams, which protect against hacking and deliberate manipulation of learning and reward systems.

- **Trust:** XAI also contributes to the development of trust by enhancing the predictability, repeatability, and stability of interpretable models. When stakeholders witness a consistent set of results, their trust grows. Once that trust has been created, end-users will be more willing to trust other applications they haven't seen before. This is especially crucial in the development of AI because models are likely to be used in situations where their users' actions may modify the environment, thus invalidating future directions.

- **Ethics:** Engineers and developers focus on functional requirements during the SDLC (Software development life cycle), while the business team focuses on speeding up AI implementation and improving performance. XAI evaluates Ethical values in the design phase of the AI development cycle, with clear governance and controls to guard against risks emerging from ethical oversight. Ethics is a system of principles that guide's to justify the good or bad, right or wrong decisions in trustworthiness, reliability, and honesty are explained in the paper [3].

- **Accountability:** It's critical to know who's responsible for an AI system's decisions, which necessitates a thorough XAI-enabled understanding of how the system works,

how it makes judgments or provides suggestions, how it learns and evolves over time, and how to ensure it works as intended.

- **Regulation:** Regulatory bodies and standard-setting organizations are concentrating on a variety of AI-related issues, with the establishment of governance, accuracy, transparency, and explainability requirements at the top of the list.

6.5.8 Metrics for Explainability

There are three key factors to consider when determining where interpretability is required and to what level. Validation, verification, and risk management. The main use case criticality evaluation criteria across six domains are Revenue, Rate, Rigour, Regulation, Reputation, and Risk. In practice, the criticality of use case explainability is driven predominately by three economic factors:

6.5.8.1 Economic factors

i. Potential influence of a single prediction. ii. Understanding why a forecast was made has economic value in terms of the actions that could be performed as a result of the prediction. iii. The economic value of the knowledge gained by analyzing trends and patterns in numerous forecasts (management information). Organizations must, however, place greater emphasis on variables other than economic and technological drives, such as executive risk, reputation, and rigor.

6.5.8.2 Evaluation of the use case criticality components

The following passage describes the use-case criticality components according to the business point of view.

- **Revenue:** The sum of the economic impact of a single prediction, the economic benefit of understanding why a single prediction was made, and the intelligence gained from a global understanding of the process being modeled.
- **Rate:** The number of decisions that an AI application must make every day vs three per month, for example.
- **Rigour:** The application's robustness: its accuracy and ability to generalize well to unknown inputs.
- **Regulation:** The regulation that establishes the acceptable use and level of functional validation for a particular AI application.

- **Reputation:** How the AI application interacts with business, stakeholders, and society, as well as the extent to which a specific use case may have a negative impact on the company's reputation.

- **Risk:** The potential harm caused by a negative outcome stemming from the employment of the algorithm extends beyond the immediate implications and covers the executive, operational, technological, social (including customers), ethical and workforce environments.

6.5.9 Explainability helps for decision taking

Decision-makers may find Explainability to be a useful tool for understanding the reasoning behind machine learning predictions and enabling meaningful agency by suggesting relevant information, see [92], [375], [343]. Explainability could aid in establishing their comprehension of a system's operation, which is required for them to trust a newly presented ML/AI system are explained in the paper [59]. Making model behavior understandable to the decision-makers may reduce the cognitive load required to complete the task and help users in overcoming algorithmic aversion by giving them a secure sense of comprehension are explained in these papers [123], [406]. There are five reasons for explainability helps in taking decisions.

- In order to meet regulatory criteria for BDAI (Big Data Artificial Intelligence)-based solutions.

- Promote transparency and trust within the organization.

- To make sure the systems are always optimized.

- In order to improve customer communication.

- Improve the hit rate, which will save expenses and boost productivity.

6.5.10 Problems with current explainability methods

But despite significant efforts in explainability research, many of the suggested methods are not usable when put into use are explained in the paper [2]. The stakeholders who are supposed to benefit from explainability efforts may ignore them or misunderstand them due to a lack of usability. For instance, are explained in the paper [52] found that explainability was primarily thought of as a tool for ML experts; users and decision-makers did not see it as a useful tool for them and instead thought of it as a tool designed for ML experts. They conducted interviews with data scientists and other stakeholders from 30 different organizations. Usability issues

may arise from a failure to recognize the explainability requirements of various stakeholders. Until recently, most research efforts were devoted to helping ML specialists and data scientists, frequently ignoring the needs of a wide range of other stakeholders trying to understand the operation of opaque systems, as explained in the paper [367]. To ensure that explainability can be used when used in practice, scholars nowadays appear to agree that greater attention should be paid to the needs of different stakeholders, as explained in the paper [52]. To avoid algorithmic aversion, explainability in decision-making contexts should be used with caution, and a clear purpose of strengthening confidence rather than simply increasing users' desire to use the system as explained in the paper [204]. Otherwise, using explainability to increase confidence in the system and its predictions could lead to automation bias are explained in the paper [406] and an unjustifiable sense of assurance are explained in the paper [180]. Some argue that users stop assessing each decision or explanation because explainability allows the formation of specific heuristics regarding the system are explained in the paper [38].

6.5.11 Important key points of XAI

There are five major key points of XAI in the Business sector as described below.

- **AI must be driven by the business:** Developers are primarily concerned with well-stated functional needs, while business managers are concerned with business metrics and regulatory compliance. Concerns about algorithmic influence usually gain traction only after algorithms fail or have a detrimental impact on the bottom line. Because AI software is more adaptable than traditional decision-making algorithms, problems can emerge faster and with greater impact. Explainable AI can help non-technical executives and developers connect, allowing top-level strategy to be effectively communicated to junior data scientists. This technology's lack of control and quality assurance is inherently unethical, and it must be addressed at all levels of the organization. Governance is extremely difficult without XAI.

- **Executive accountability:** With the rise of AI systems and their greater impact on businesses, individuals, and society, it's important to understand who is responsible for the decisions made by AI systems. If leaders are obliged to bear responsibility for AI, they must be aware of the risk it poses to their company. Executives would bring unforeseen risks to their profile if they didn't understand the system's rationales. Executives must have faith in a system's ability to work within set bounds to take accountability.

- **Doing the right thing, right:** AI systems are designed by humans, and ethics are included before the first line of code is written. Before developing and deploying an AI

system, it is critical to have specified ethics and core principles, as well as a governance mechanism that assures compliance. These foundations guide your company's interactions with potential customers, preventing unethical behavior. It's critical that business managers understand the risks, be held accountable for them, and, when necessary, have a formal system in place to match your technology with your company's ethical principles and risk appetite.

- **Future-proofing your AI:** The demand for interpretability is growing. The usage of powerful AI in fields like financial services is so well established that dangers should be at the top of the risk list. Other industries, including healthcare and transportation, are quickly catching up. And, as AI spreads throughout the economy, all sectors will need to assess the criticality and influence of their AI on the one hand, as well as their confidence in the outcomes on the other. There will also be a new generation of AI-specific regulations. In this regard, AI explainability is fast catching up to cyber as a threat, but it may also be a useful differentiator if handled correctly.

Explainability is a business issue that requires a response from the CEO and the board of directors, not just a technical one. Now is the time to act-get this right, and you'll be able to move forward with more confidence.

6.5.12 Limitations of XAI

Here some of the XAI limitations are described below.

- **Confidentiality:** An algorithm may be classified as confidential, a trade secret, or a security risk if it is revealed. How can we be sure that the AI system hasn't picked up a biased view of the world (or perhaps an unbiased view of a biased world) due to flaws in the training data, model, or goal function? Consider a scenario in which the software's creators are biased, either consciously or unconsciously.

- **Complexity:** Algorithms are sometimes simple to understand but quite difficult to implement. As a result, a layperson's comprehension is ludicrous, and this is an area where XAI techniques may be effective. Because XAI can build alternative algorithms that are easier to understand.

- **Unreasonableness:** Algorithms that produce conclusions based on genuine facts that are not rational, prejudiced, or out of line. How can you be sure that the decisions made by an AI system are fair? Reasonability is logical, but it also has the perspective that it is dependent on the precise knowledge input supplied to AI algorithms.

- **Injustice:** We may be able to grasp how an algorithm works, but we need to know how the system complies with a legal or moral code.
- **Control:** To progress from a proof-of-concept to a full-fledged implementation, you must be certain that your system meets all of the specified requirements and does not exhibit any undesirable behaviors. If the system makes a mistake, businesses must be able to recognize that anything is wrong to take corrective action or, in the worst-case scenario, shut down the AI system. By monitoring performance, highlighting faults, and offering a means to switch the system off, XAI may help your organization maintain control over AI. In terms of data privacy, XAI may assist in ensuring that only approved data is utilized, for an agreed-upon purpose, and that data can be deleted if necessary.
- **Safety:** There have been various faults raised about AI systems' safety and security, particularly as they become more powerful and prevalent. This can be attributed to a variety of issues, including unethical design, engineering oversights, hacking, and the impact of AI's operating environment. XAI can assist in the detection of such flaws. To protect against hacking and deliberate manipulation of learning and reward systems, it's also critical to collaborate closely with cyber detection and protection teams.
- **Trust:** XAI was utilized to increase confidence by improving the predictability, repeatability, and stability of interpretable models. When stakeholders witness a consistent set of results, their confidence grows over time. Once such trust has been created, end-users will be more willing to trust applications they haven't seen before. This is especially crucial in the development of AI because models are likely to be used in situations where their use could change the environment, thus invalidating future predictions.
- **Ethics:** Engineers and developers focus on functional requirements during the SDLC (software development life cycle), whereas business teams focus on speeding up AI implementation and improving performance. The need to fulfill these compelling aims might readily hide the ethical implications and other unforeseen consequences. It's critical that a moral compass is included in AI training from the beginning, and that AI behavior is regularly monitored through XAI evaluation after that. A formal method that matches a company's technological design and development with its ethical ideals and principles, as well as its risk appetite, may be required where suitable. PwC is aiming to incorporate ethical considerations into the design phase of the AI development cycle, with explicit governance and control in place to protect against ethical oversight risks.
- **Accountability:** It's critical to understand who is responsible for an AI system's judgments. This, in turn, necessitates a thorough XAI-enabled understanding of how the sys-

tem works, how it makes judgments or provides suggestions, how it learns and changes over time, and how to ensure it works as intended.

- **Regulation:** While AI is currently unregulated, this is expected to change as its impact on daily life grows more widespread. Regulatory bodies and standard-setting organizations are concentrating on a number of AI-related issues, with governance, accuracy, openness, and explainability at the top of the list. Protecting potentially vulnerable consumers is another regulatory priority.

6.5.13 Key Research Development in XAI

XAI has grown with numerous studies and methodologies, from interpretable linear machine learning models to deriving explanations to systems with complicated deep neural networks. The Defense Advanced Research Projects Agency in the United States is one of the primary institutions responsible for making XAI so popular (DARPA). In addition to enabling transparency in AI systems, hybrid cloud and AI platform providers such as IBM and Google have also made significant investments in this area. In addition, certain well-known open source tools, such as Dalex, SHAP, Lime, Shapash, and others, come in handy when working with XAI.

- **SHAPASH:** SHAPASH is a Python package that is used to make machine learning accessible to everyone. It offers a variety of visualizations with clear descriptions that anyone can understand. Data scientists can readily understand and share their models. A description of the most influential criteria helps end-users understand the decision proposed by a model. SHAPASH also aids data science audits by providing important information about any model and data in a single report.

- **LIME:** A Loadable Kernel Module (LKM) for Linux and Linux-based devices such as Android that enables volatile memory acquisition. This distinguishes LIME as the first tool to capture complete memory on Android devices. It also reduces user-kernel interaction during acquisition, resulting in more forensically sound memory grabs than existing Linux memory acquisition programs.

- **DARPA:** Deep Explanation, Interpretable Models, and Model Induction were three of DARPA's explainability techniques. Based on promising research at the time, their goal was to adopt new modified ML algorithms and develop explanation models based on psychological theories of understanding while maintaining a high degree of learning performance are explained in the paper [151].

- **DALEX:** DALEX is an open-source explainability package that may be used with Keras, TensorFlow, and h2o are explained in the paper [37]. Breakdown are explained in the paper [354], Lime (Local Interpretable Model-agnostic Explanations) are explained in the paper [310], and Shapley Value are explained in the paper [216] are examples of standard methods for interpretability and explanation. The Dalex package provides uniform abstraction across predictive models, model-level explanation, and predictive-level explanation.

- **SHAP:** SHAP is a model-independent explainability framework. It is currently one of the most popular ways to explain machine learning. SHAP (Shapley Addictive Explanations) is a unified game theory-based strategy for interpreting any model's output are explained in the paper [216].

- **IBM - AI Explainability 360:** In addition to advancing XAI research and development, IBM released AIX360, an open-source platform that provides data scientists and developers with a variety of explainability methods, evaluation metrics, and taxonomy of explainability approaches. This software toolkit was developed by IBM Research's Vijay Arya and colleagues to provide a platform for identifying gaps in AI systems and explaining AI models or complex datasets to people in various ways. It's claimed as the most comprehensive tool for explaining AI, covering topics like data interpretation, local and global post-hoc methodologies, person-specific explanations, and evaluation measures are explained in the paper [30].

6.5.13.1 Critical Domains

- **Health-care:** Health care workers utilize AI to speed up and improve a variety of functions, including decision-making, forecasting, risk management, and even diagnosis, by scanning medical pictures for anomalies and patterns that are undetected to the naked eye. Many healthcare practitioners now use the XAI as a critical tool, but it is often difficult to understand, causing frustration among clinicians and patients, especially when making high-stakes decisions. Here XAI is used in the following scenarios. When fairness is critical, and end-users or consumers require information to make an informed decision, When the consequence of a wrong AI choice is severe (such as a reference to unnecessary surgery), When the cost of a mistake is substantial, such as a misdiagnosis of a malignant tumor that results in excessive financial charges, increased health risks, and personal trauma, and When the AI system generates a new hypothesis, it must be confirmed by domain or subject matter experts are explained in the paper [9].

- **Criminal justice:** ML/DL algorithms are increasingly being confronted by courts in criminal, administrative, and civil matters. Judges, according to this essay, should seek explanations for algorithmic conclusions. Designing systems that explain how algorithms arrive at their conclusions or predictions is one option to overcome the "black box" problem. Judges will play a pivotal role in influencing the type and form of "explainable AI" if and when they demand these explanations (XAI). Courts can define what XAI should mean in various legal settings using common law methods. There are several benefits to having courts play this role: A pragmatic way to reach nuanced judgments is to use judicial reasoning that builds from the bottom up, employing case-by-case assessment of the evidence are explained in the paper [99].

- **Real-Estate:** In real estate AI apps can handle planned data flows, learn user behavior, streamline and expedite operations, and provide more accurate assessments and market forecasts using XAI methods.

- **Autonomous vehicles:** Vehicle automation entails using mechatronics, artificial intelligence, and multi-agent systems to help a vehicle's operator (car, aircraft, watercraft, or otherwise) are explained in these papers [164], [163]. These features, as well as the vehicles that use them, may be referred to as intelligent or smart. Semi-autonomous vehicles use automation for challenging tasks, such as navigation, to help but not replace human input, whereas robotic or autonomous vehicles rely solely on automation.

- **Algorithmic trading:** Algorithmic trading is a way of executing orders using preprogrammed automatic trading instructions that take into consideration variables including time, price, and volume are explained in the paper [206]. In comparison to human traders, this sort of trading tries to take advantage of the speed and computational capacity of computers. Algorithmic trading has gained popularity among retail and institutional traders in the twenty-first century. It is extensively utilized by investment banks, pension funds, mutual funds, and hedge funds that need to spread out the execution of a larger order or execute deals that are too quick for human traders to respond.

6.5.13.2 Mission Critical Systems

A mission-critical system is a computer, electrical, or electromechanical system that is essential to the success of an operation. When such a system fails or is disrupted, the consequences are typically severe and immediate.

Mission-critical systems include mission-essential equipment and mission-critical applications. Examples of mission-critical systems include: Operating and controlling systems for

railways and airplanes, Electric power grid management systems, Communication systems for first responders. Explainable Artificial Intelligence (XAI) offers advantages over traditional Decision Support Systems (DSS) and Recommender Systems (RS) in mission-critical domains by providing transparent and trustworthy AI-driven decision-making.

6.5.14 Which sectors need XAI but other approaches are not suitable in RS, ES, and DSS

Common suitable sectors for XAI: Health care, marketing, insurance, and financial services. In these sectors, Machine Learning (ML), Artificial Intelligence (AI) techniques, and XAI methods are used to develop more explainable models while maintaining a high level of learning prediction (accuracy) and performance.

6.5.15 Explainability of different types of algorithms and learning techniques on a subjective scale

Explainability is the ability to comprehend a given result when seen as post hoc interpretations. We have provided subjective values on a scale of 1 to 5 (with 1 being the most difficult and 5 being the easiest) to rate how easy or difficult it is for an end-user to decipher why a model made a certain decision because most of these models do not provide direct explanations as to why or how the results are achieved. Each of these learning processes has its structure, which is influenced by how new information is learned.

Table 6.6: Explainability of AI Algorithms and Learning Techniques

S.No	AI Class	Technique	Explainability (1–5)	Reasoning
1	Graphical Models	Bayesian Belief Networks (BBNs)	3.5	Represent conditional dependencies; attribute-level reasoning is traceable.
2	Supervised/Unsupervised	Decision Trees	4	Tree structure enables transparent decision rules and criteria.
3	Supervised	Logistic Regression	3	Mathematical relationships are visible, but coefficients may be complex to interpret.
4	Supervised	Support Vector Machines (SVMs)	2	Decision planes are known, but feature-level explanations are limited.
5	Unsupervised	K-means Clustering	3	Clusters are interpretable via centroids, though meaning may be abstract.
6	Deep Learning	Neural Networks (NNs)	1	Hidden layers obscure direct reasoning; low explainability.
7	Ensemble Models	Random Forest/Boosting	3	Based on multiple trees; some transparency with reduced clarity.
8	Reinforcement Learning	Q-Learning	2	Learns through reward signals; underlying logic is difficult to trace.
9	NLP	Hidden Markov Models (HMMs)	3	Sequence modeling with probabilistic transitions provides moderate interpretability.

Here Table 6.6 describes the Explainability of algorithms and learning techniques.

6.5.15.1 Architectural Similarities between DSS, ES, RS, and XAI

Table 6.7 describes the architectural similarities between DSS, ES, RS, and XAI.

Table 6.7: Architectural Similarities between DSS, ES, RS, and XAI

DSS	**ES**	**RS**	**XAI**
Inference Engine	Inference Engine	Recommendation Engine	AI Entity
Knowledge Base	Knowledge Base	Knowledge Base	XAI Database
User Interface	User Interface	User Interface	User Interface
Database	Explanation Module	Recommender Systems	Explanation Module
Translation Process	–	Translation Process	Behavior Translation Process

6.6 Case Study of Defect Prediction

In the modern era of advancing digitalization, software defects have become prevalent and incredibly costly, yet they pose significant challenges in terms of identification, anticipation, and prevention. Consequently, if software defects persist in safety-critical systems, they could lead to severe harm to individuals, jeopardize lives, and cause catastrophic incidents. In this section, we consider a case study of how Software Defect Prediction provides an insightful lens to explore the evolution of AI-driven decision-making in the realm of DSS, ES, RS, and XAI. Over the years, advancements in these areas have played a crucial role in improving software quality, mitigating risks, and enhancing overall system performance.

6.6.1 DSS

Initially, DSS for software defect prediction relied on basic statistical techniques and data analysis to identify potential problem areas. These systems would analyze historical defect data, such as bug reports and code changes, and generate statistical reports on defect rates and trends. While these early DSS provided some insights, they lacked the ability to detect specific defects or provide actionable recommendations.

Advancements in machine learning and data mining techniques paved the way for more sophisticated DSS in software defect prediction. These systems began leveraging historical defect data, code metrics, and other relevant software attributes to build predictive models. Machine learning algorithms, such as decision trees, random forests, and support vector ma-

chines, were employed to classify software components as defective or non-defective based on their features.

As DSS evolved, they began to integrate additional data sources, such as code complexity metrics, code churn, and developer activity, to enhance the accuracy of defect prediction models. By considering multiple factors, these DSS became more capable of identifying potential defect-prone areas in software code. Developers could prioritize their efforts, focus on critical components, and allocate resources effectively.

Moreover, DSS for software defect prediction started incorporating feedback loops. As new defect data became available, the models were retrained, allowing them to adapt and improve over time. This iterative process helped refine the defect prediction models, increasing their accuracy and reducing false positives or false negatives.

6.6.2 ES

In the case of software defect prediction, expert systems have been instrumental in capturing the expertise of experienced software engineers and translating it into a rule-based system. These systems analyze various factors such as code patterns, software structure, and historical defect data to identify potential defects and provide recommendations for preventive measures.

One of the key advantages of expert systems is their ability to leverage a vast amount of domain-specific knowledge. Software engineers with extensive experience in detecting and fixing defects can contribute their expertise to the system. This knowledge is captured in the form of rules that represent the decision-making processes followed by human experts. These rules are typically based on if-then statements, where certain conditions in the software code trigger specific actions or recommendations.

To develop an expert system for software defect prediction, the first step involves acquiring and organizing the relevant knowledge from domain experts. This knowledge includes coding standards, best practices, common pitfalls, and defect patterns. The acquired knowledge is then translated into a rule-based structure that can be processed by the expert system.

Once the expert system is developed, it can be deployed to analyze software code and predict potential defects. By applying the established rules to the code, the system identifies patterns and detects areas that are likely to contain defects. The expert system can also provide recommendations for preventive measures, such as suggesting alternative coding approaches or highlighting potential vulnerabilities.

One notable aspect of expert systems is their ability to learn and adapt. As the system processes more data and encounters real-world scenarios, it can update and refine its rules. This adaptive capability ensures that the expert system evolves with changing software development practices and emerging defect patterns.

Expert systems have significantly improved the accuracy and efficiency of software defect prediction. By leveraging the collective knowledge and expertise of experienced software engineers, these systems can identify potential defects early in the development process, enabling timely interventions and preventing the occurrence of critical issues. Moreover, expert systems provide consistent and reliable predictions, reducing the dependence on individual expertise and minimizing human errors.

However, it is important to note that expert systems have certain limitations. They heavily rely on the knowledge and rules provided by human experts, which means that any biases or limitations in the expert's knowledge can impact the system's performance. Additionally, expert systems may struggle with handling complex or ambiguous scenarios that go beyond the scope of predefined rules. In such cases, the system may require constant updates and refinement to ensure accurate predictions.

6.6.3 RS

In the context of software defect prediction, recommender systems leverage machine learning algorithms and collaborative filtering techniques to analyze vast amounts of historical defect data. They aim to identify patterns, similarities, and relationships between various software projects and their associated defects. This analysis enables recommender systems to generate personalized recommendations and suggestions for software developers to proactively address potential defects and improve overall code quality.

One of the primary benefits of using recommender systems for software defect prediction is their ability to leverage collective intelligence. By analyzing a wide range of historical defect data from different software projects and teams, recommender systems can identify common patterns and correlations that might be missed by individual developers or traditional rule-based systems. This collective intelligence provides valuable insights into defect-prone areas, coding practices, and potential risks, enabling developers to make informed decisions during the development process.

Recommender systems also play a crucial role in knowledge transfer and sharing within software development teams. By analyzing historical defect data, these systems can identify expertise and experience patterns of individual developers or teams. They can then recommend relevant knowledge resources, best practices, and coding guidelines to other team members. This knowledge transfer facilitates a collaborative and learning-oriented environment, where developers can benefit from the collective experience and expertise of their peers, ultimately leading to improved software quality and defect prevention.

Furthermore, recommender systems in software defect prediction can assist in resource allocation and project planning. By analyzing historical defect data and project characteristics,

these systems can recommend appropriate allocation of resources, such as assigning experienced developers to critical areas or allocating additional testing efforts in defect-prone modules. This data-driven decision-making helps optimize resource utilization, mitigate risks, and enhance overall project management.

It is important to note that the effectiveness of recommender systems in software defect prediction heavily relies on the availability and quality of historical defect data. Therefore, organizations must prioritize the collection, documentation, and maintenance of defect data throughout the software development lifecycle. The more comprehensive and accurate the historical data, the more reliable and precise the recommendations generated by the recommender systems will be.

6.6.4 XAI

Initially, traditional machine learning models, such as neural networks, were effective in predicting software defects. However, their decision-making processes were often considered black boxes, making it challenging for developers to understand and trust the predictions. This lack of interpretability raised concerns, especially in safety-critical systems where the consequences of software defects could be severe. Recognizing the need for transparent decision-making, researchers and practitioners delved into developing XAI techniques are explained in these papers [361], [239] specifically tailored for software defect prediction. These techniques aim to provide understandable explanations for the predictions made by AI models, enabling developers to comprehend the reasoning behind the defect predictions.

One approach within XAI is rule-based explanations. This technique involves extracting decision rules from the AI model, allowing developers to understand how the model reaches a particular prediction. By translating complex model behavior into interpretable rules, developers can identify the specific code patterns, metrics, or features that contribute to defect predictions. These explanations serve as valuable insights, enabling developers to focus their efforts on the relevant areas of the codebase and take appropriate preventive actions. Another XAI technique applied to software defect prediction is feature importance analysis. This method aims to identify the most influential features or factors contributing to the model's predictions. By quantifying the importance of different code attributes or metrics, such as cyclomatic complexity, code churn, or code coupling, developers gain a better understanding of the key drivers behind defect predictions. This knowledge helps prioritize efforts for defect prevention, code refactoring, or targeted testing in the areas deemed most critical by the model.

Counterfactual explanations are another XAI technique used in software defect prediction. These explanations provide alternative scenarios where defects are prevented. By analyzing counterfactuals, developers can understand the specific code changes or modifications required

to avoid predicted defects. This information guides them in making informed decisions to rectify potential weaknesses in the software codebase and prevent defects from occurring. XAI techniques also include visualization methods that present defect predictions in a visually interpretable mannerare explained in the paper [82]. Graphical representations, heatmaps, or interactive dashboards can be employed to display the areas of the codebase with the highest defect probabilities. This visual feedback allows developers to quickly grasp the overall defect landscape, identify hotspots, and prioritize their actions accordingly.

By integrating XAI techniques into software defect prediction, developers gain insights into the inner workings of AI models. They can validate, interpret, and refine the predictions, ultimately improving the model's accuracy and reliability. Additionally, XAI fosters collaboration between AI systems and human experts, empowering developers to leverage their domain knowledge and make informed decisions based on the explanations provided. The evolution of XAI in the case of software defect prediction has significantly enhanced trust, transparency, and accountability in AI-driven decision-making systems. It has bridged the gap between complex machine learning models and human understanding, empowering developers to make effective use of AI while maintaining control and ensuring software quality.

Table 6.8: Evolution of Software Defect Prediction (SDP) based on Decision Techniques

Decision Technique	**Evolution of SDP**
Decision Support Systems (DSS)	Initially, DSS provided information and reports for decision-making processes. DSS evolved to incorporate defect prediction capabilities, enabling developers to identify potential defects.
Expert Systems (ES)	ES emerged to emulate the decision-making abilities of human experts. ES captured the expertise of software engineers and translated it into a rule-based system for defect prediction.
Recommender Systems (RS)	RS utilized machine learning and collaborative filtering techniques to analyze historical defect data. RS provided personalized recommendations for preventive measures based on similarities to previous projects.
Explainable AI (XAI)	XAI techniques introduced transparency and interpretability in AI-driven defect prediction. XAI enabled understanding of the reasoning behind defect predictions, allowing for validation and refinement. XAI provided rule-based explanations, feature importance analysis, counterfactual explanations, and visualization.

Table 6.8 provides a concise overview of the evolution of Software Defect Prediction (SDP) based on four decision techniques: Decision Support Systems (DSS), Expert Systems (ES), Recommender Systems (RS), and eXplainable AI (XAI). It highlights how each technique has contributed to the advancement of SDP over time. From the initial provision of information and reports in DSS to the rule-based systems of ES, the personalized recommendations of RS, and the introduction of transparency and interpretability through XAI, these decision techniques have played significant roles in enhancing defect prediction capabilities. The table serves as a reference point to understand the progression of SDP and the impact of these decision techniques in the software development domain.

6.6.5 Using XAI for SDP

To demonstrate the effectiveness of XAI, we use JMI dataset which is widely used in SDP literature and use LIME model to explain the outcome. Recently, [54] article explained that an exhaustive combination of different feature engineering techniques was applied to find the best combination for SDP tasks. They used scaling, feature extraction, feature scaling and oversampling to get the best data transformation and use a SVM model to get the best results. In this paper, we follow the same feature engineering pipeline and train the SVM model on the transformed data. Since SVM is not intrinsically explainable, we use LIME to explain the model after training (Post-Hoc). The results are reproducible for future experimentation and use and can be found at: https://gist.github.com/nitin-bommi/2f08c2bf2acd4daa938b80761f93d10a.

The data set is divided into train and test sets in the ratio 8:2. A feature transformation pipeline is then created that transforms the input data into useful and relevant data that is free from noise and improves computational time. To demonstrate the results, we used 8 components from the PCA as these 8 components retain around 96.6% of the data, and the ratio of information retained by each principal component is shown in figure 6.9. The figure shows that most of the information is retained in the first component. Instead of using 21 features for training that increase the time taken to train the model, we use only 8 components that represent most of the data. After extracting these 8 components, we use feature selection again to select the best 4 features and drop the remaining features. This reduces the dimension of the data from 21 features to just 4 features. Table 6.9 shows the ANOVA F-test scores of different principal components computed previously. We see that the first component that extracts the maximum information has the highest score indicating that it has the greatest influence on the outcome. The training data is then augmented to match the number of majority class samples. The model is then trained on the augmented samples. The model gave an accuracy of around 72.89%.

Figure 6.10 shows the explanations. It can be seen that the model is 37% confident that it is non-defective and 63% confident that it is defective. The features that influenced the model to come to the decision are also shown. Features 4, 2, and 3 have negative coefficients while Feature 1 has a positive coefficient. It also shows that feature 4 has the most impact however, the coefficient of the model for feature 1 might be large so the overall prediction is defective. With this kind of explanation, developers can actually look at the features that cause the model to be defective. This will allow the testers and developers to improve the code in the ”non-defective” direction.

However, since explanations are given for transformed features and not for original features, it is hard to interpret what original features the developers have to change to reflect a change in the transformed features. One solution to tackle this problem is to use all the features without using any transformations. However, this will result in high dimensionality and the model will lack generalization. Another solution is to use inverse transformations, where we perform the transformations in the reverse direction. However, this results in some data loss as the data will be projected onto a new set of dimensions.

This case study shows that, though XAI is useful in explaining the predictions and helping developers and testers backtrack the outcome, it does not explain the fine details of which attributes have to be changed precisely. To overcome this problem to some extent, we can make use of counterfactual explanations that suggest the changes required to get the desired output. However, as the dimensions increase, the search space of the input also increases making it difficult to produce feasible explanations.

Table 6.9: ANOVA F-test scores for each principal component. The top four components with the highest scores are selected for downstream analysis.

PC	Score
1	499.64
2	4.95
3	238.85
4	0.04
5	15.08
6	1.71
7	0.27
8	60.21

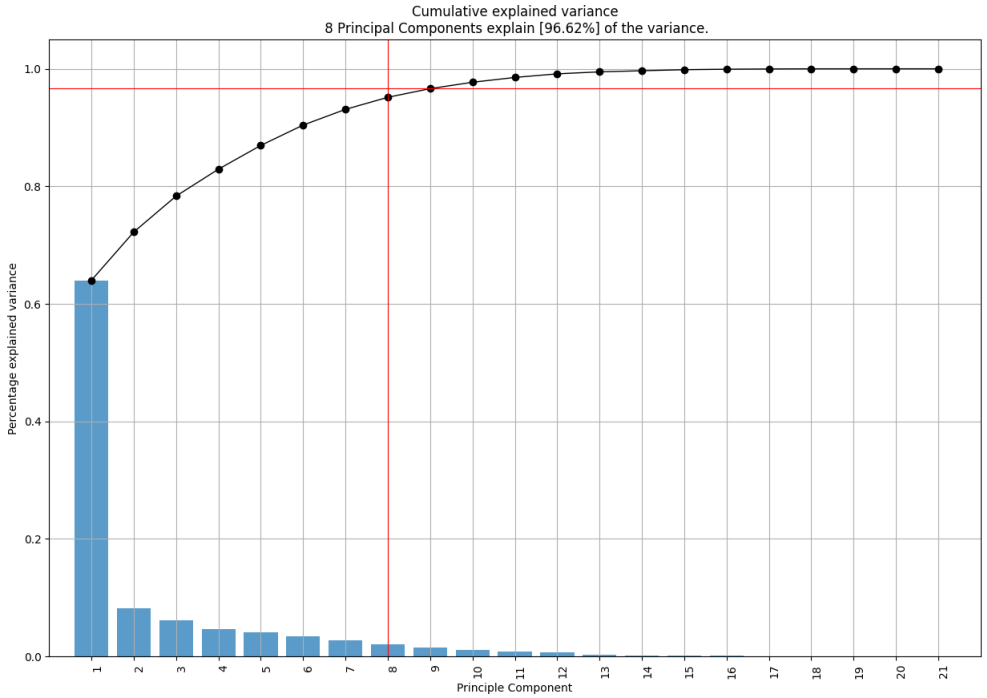


Figure 6.9: Variance explained by each principal component. The height of each bar indicates how much of the total variability in the dataset is captured by that component, which can guide the selection of the number of components for dimensionality reduction.

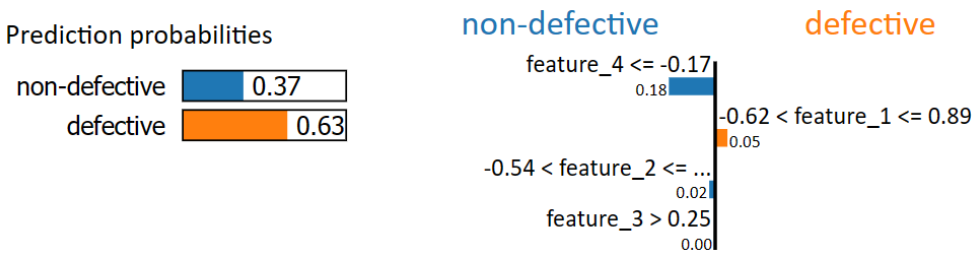


Figure 6.10: LIME explanations highlighting feature contributions for predictions on non-defective and defective software modules in the SDP dataset.

6.7 Conclusion

This chapter presents a comprehensive comparative analysis of four foundational AI-based decision-making systems: Decision Support Systems (DSS), Expert Systems (ES), Recommender Systems (RS), and Explainable AI (XAI). By critically examining their core components, strengths, limitations, and potential synergies, the chapter offers a holistic perspective on the evolving landscape of AI-driven decision making in industrial applications. The insights gained here serve as a valuable foundation for designing and implementing intelligent and interpretable decision support systems tailored to data-driven industrial contexts.

Furthermore, the chapter highlights the crucial role of Explainable AI (XAI) in enhancing transparency and trustworthiness of complex machine learning models within industrial domains. Through practical case studies, such as the application of LIME in software defect prediction, this work demonstrates how XAI facilitates responsible AI-human collaboration by making AI decisions understandable and actionable. These findings underscore the importance of integrating explainability into industrial AI systems to foster more reliable, accountable, and effective decision-making processes [2]. In the following chapters, we will delve into more specific studies and applications.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis, titled *"Data-Driven Approaches Using Explainable AI for Industrial Applications"*, presents a systematic investigation into the integration of Explainable Artificial Intelligence (XAI) across various intelligent decision-making systems relevant to industrial domains. Each chapter has addressed specific challenges in industrial applications, ranging from classical AI systems to modern interpretable machine learning solutions in recruitment and fraud detection.

7.2 Summary of Contributions

Chapter 3 presented a comparative overview of Expert Systems (ES), Recommender Systems (RS), and Explainable AI (XAI), emphasizing their architecture, applications, and relevance in industrial settings. It highlighted the limitations of ES and RS, and showcased how XAI revives interpretability in complex systems, fostering ethical and transparent decision-making in industrial contexts.

Chapter 4 proposed an interpretable resume categorization framework that integrates K-Nearest Neighbors (KNN) with Local Interpretable Model-Agnostic Explanations (LIME). This method improved transparency and fairness in automated hiring systems, addressing industrial challenges like bias, accountability, and trust in AI-driven recruitment platforms.

Chapter 5 expanded the discussion to core AI-based decision-making systems, including Decision Support Systems (DSS), ES, RS, and XAI. A holistic analysis was presented, exploring their strengths, weaknesses, and synergy. Case studies, such as LIME's application to software defect prediction, reinforced XAI's importance in enhancing transparency and trust in

mission-critical industrial applications.

Chapter 6 introduced a Hybrid Threshold Method for fraud detection by combining AUC-ROC and PR-AUC curves with LIME and SHAP for interpretability. This approach demonstrated superior performance over traditional methods, particularly in handling class imbalance. It contributed a trustworthy framework for detecting fraud in industrial systems, balancing predictive performance with explainability.

Chapter 7 proposed a novel clustering and summarization framework using Gustafson-Kessel (GK) fuzzy clustering and Sentence-BERT embeddings. The integration of LIME provided actionable insights into cluster formation, ensuring transparency in resume profiling. This approach outperformed traditional clustering methods and addressed semantic richness and high dimensionality in resume datasets, advancing talent analytics in industrial applications.

7.3 Future Directions

Based on the outcomes of this research, several promising directions for future work are outlined below:

- **Dynamic Explainability:** Extend XAI frameworks to support real-time explanations, especially for high-frequency data applications like fraud detection and recruitment analytics.

- **Interpretability of Deep Learning Models:** Explore advanced XAI techniques such as counterfactual explanations, attention mechanisms, and concept-based models to better understand deep neural networks in industrial use cases.

- **Multimodal Fusion:** Integrate diverse data modalities (e.g., transactional logs, behavioral signals, text) to develop more robust and explainable AI models suited for industrial decision-making.

- **Scalability and Optimization:** Enhance the computational efficiency and scalability of summarization and clustering algorithms to support deployment in large-scale, real-world industrial systems.

- **Policy, Ethics, and Compliance:** Align XAI frameworks with regulatory requirements, ethical standards, and industrial safety norms to ensure responsible and lawful AI deployment.

7.4 Concluding Remarks

This thesis highlights the pivotal role of Explainable AI in shaping the future of industrial applications. By embedding transparency and accountability into AI-driven decision-making systems, research contributes to creating intelligent systems that are not only effective but also interpretable and trustworthy. These contributions form a critical step toward responsible AI adoption across industrial domains, fostering seamless human-AI collaboration and laying the foundation for future innovations in ethical and explainable AI solutions tailored for industry.

References

[1] ARASH ABADPOUR. **A sequential Bayesian alternative to the classical parallel fuzzy clustering model**. *Information Sciences*, **318**:28–47, 2015. ()

[2] ASHRAF ABDUL, JO VERMEULEN, DANDING WANG, BRIAN Y LIM, AND MOHAN KANKANHALLI. **Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda**. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018. (137, 153)

[3] SAMY ABU-NASER AND MOHAMMED ALHABBASH. **MALE INFERTILITY EXPERT SYSTEM DIAGNOSES AND TREATMENT**. *The American Journal of Innovative Research and Applied Sciences.*, **2**:181–192, 04 2016. (135)

[4] FRANCISCA ADOMA ACHEAMPONG, HENRY NUNOO-MENSAH, AND WENYU CHEN. **Transformer models for text-based emotion detection: a review of BERT-based approaches**. *Artificial Intelligence Review*, **54**(8):5789–5829, 2021. ()

[5] AMINA ADADI AND MOHAMMED BERRADA. **Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)**. *IEEE Access*, **6**:52138–52160, 2018. (xi, 2, 4, 86, 125, 127, 129, 130)

[6] ACHEAMPONG FRANCISCA ADOMA, NUNOO-MENSAH HENRY, AND WENYU CHEN. **Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition**. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE, 2020. ()

[7] GEDIMINAS ADOMAVICIUS AND ALEXANDER TUZHILIN. **Context-aware recommender systems**. In *Recommender systems handbook*, pages 217–253. Springer, 2011. (124)

[8] CHARU C AGGARWAL ET AL. *Recommender systems*, **1**. Springer, 2016. (125)

[9] MUHAMMAD AHMAD, CARLY ECKERT, GREG MCKELVEY, KIYANA ZOLFAGAR, ANAM ZAHID, AND ANKUR TEREDESAI. **Death vs. data science: predicting end of life**. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 2018. (142)

[10] SHAMIM AHMED, M. SHAMIM KAISER, MOHAMMAD SHAHADAT HOSSAIN, AND KARL ANDERSSON. **A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions**. *IEEE Access*, **13**:37370–37388, 2025. (87)

[11] ULYSSE AIVODJI, FRANCESCO ALESIANI, SÉBASTIEN GAMBS, MICHAËL HUGUET, AND THOMAS MARTIN. **Fairwashing: the risk of rationalization**. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 134–140, 2019. (19, 24)

[12] AKIKO AIZAWA. **An information-theoretic perspective of tf–idf measures**. *Information Processing & Management*, **39**(1):45–65, 2003. ()

[13] BERLIAN AL KINDHI, TRI ARIEF SARDJONO, MAURIDHI HERY PURNOMO, AND GIJBERTUS JACOB VERKERKE. **Hybrid K-means, fuzzy C-means, and hierarchical clustering for DNA hepatitis C virus trend mutation analysis**. *Expert systems with applications*, **121**:373–381, 2019. (44)

[14] MADHAVI ALAMURI, BAPI RAJU SURAMPUDI, AND ATUL NEGI. **A survey of distance/similarity measures for categorical data**. In *2014 International joint conference on neural networks (IJCNN)*, pages 1907–1914. IEEE, 2014. (34)

[15] NOURAH ALANGARI, MOHAMED EL BACHIR MENAI, HASSAN MATHKOUR, AND IBRAHIM ALMOSALLAM. **Intrinsically Interpretable Gaussian Mixture Model**. *Information*, **14**(3):164, 2023. (48)

[16] MAJED ALATEEQ AND WITOLD PEDRYCZ. **Multi-Context Fuzzy Clustering: Towards Interpretable Fuzzy Clustering**. *IEEE Transactions on Fuzzy Systems*, 2024. (48)

[17] BADER ALDUGHAYFIQ, FARZEEN ASHFAQ, NZ JHANJHI, AND MAMOONA HUMAYUN. **Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP**. *Diagnostics*, **13**(11):1932, 2023. (45)

[18] LEO ALEXANDER III, Q CHELSEA SONG, LOUIS HICKMAN, AND HYUN JOO SHIN. **Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment**. *International Journal of Selection and Assessment*, **33**(1):e12499, 2025. (23)

[19] RASIM M ALGULIYEV, RAMIZ M ALIGULIYEV, AND NIJAT R ISAZADE. **An unsupervised approach to generating generic summaries of documents**. *Applied Soft Computing*, **34**:236–250, 2015. (47)

[20] IRFAN ALI, NIMRA MUGHAL, ZAHID HUSSAIN KHAND, JAVED AHMED, AND GHULAM MUJTABA. **Resume classification system using natural language processing and machine learning techniques**. *Mehran University Research Journal of Engineering & Technology*, **41**(1):65–79, 2022. (xiii, 41)

[21] SAJID ALI, TAMER ABUHMED, SHAKER EL-SAPPAGH, KHAN MUHAMMAD, JOSE M ALONSO-MORAL, ROBERTO CONFALONIERI, RICCARDO GUIDOTTI, JAVIER DEL SER, NATALIA DÍAZ-RODRÍGUEZ, AND FRANCISCO HERRERA. **Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence**. *Information fusion*, **99**:101805, 2023. (2, 14, 16)

[22] GULSUM ALICIOGLU AND BO SUN. **A survey of visual analytics for Explainable Artificial Intelligence methods**. *Computers & Graphics*, **102**:502–520, 2022. (127)

[23] STEVEN ALTER. **A taxonomy of decision support systems**. *Sloan Management Review (Pre-1986)*, **19**(1):39, 1977. (127)

[24] MIGUEL ALVAREZ-GARCIA, RAQUEL IBAR-ALONSO, AND MAR ARENAS-PARRA. **A comprehensive framework for explainable cluster analysis**. *Information Sciences*, **663**:120282, 2024. ()

[25] DAVID ALVAREZ-MELIS AND TOMMI S. JAAKKOLA. **On the Robustness of Interpretability Methods**. *CoRR*, **abs/1806.08049**, 2018. (88)

[26] SHIDEH SHAMS AMIRI, SAM MOTTAHEDI, EARL RUSTY LEE, AND SIMI HOQUE. **Peeking inside the black-box: Explainable machine learning applied to household transportation energy consumption**. *Computers, Environment and Urban Systems*, **88**:101647, 2021. ()

[27] ABDUL QUAIYUM ANSARI AND MOHAMMAD AYOUB KHAN. **Fundamentals of industrial informatics and communication technologies**. In *Handbook of Research on Industrial Informatics and Manufacturing Intelligence: Innovations and Solutions*, pages 1–19. IGI Global Scientific Publishing, 2012. (1, 2)

[28] K ARCHANA AND KG SARANYA. **Crop Yield Prediction, Forecasting and Fertilizer Recommendation using Voting Based Ensemble Classifier**. *SSRG Int. J. Comput. Sci. Eng*, **7**, 2020. ()

[29] ALEJANDRO BARREDO ARRIETA, NATALIA DÍAZ-RODRÍGUEZ, JAVIER DEL SER, ADRIEN BENNETOT, SIHAM TABIK, ALBERTO BARBADO, SALVADOR GARCÍA, SERGIO GIL-LÓPEZ, DANIEL MOLINA, RICHARD BENJAMINS, ET AL. **Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI**. *Information fusion*, **58**:82–115, 2020. (xi, 2, 4, 86, 88, 128)

[30] VIJAY ARYA, RACHEL KE BELLAMY, PIN-YU CHEN, AMIT DHURANDHAR, MICHAEL HIND, SAMUEL C HOFFMAN, STEPHANIE HOUDE, Q VERA LIAO, RONNY LUSS, ALEKSANDRA MOJSILOVIĆ, ET AL. **One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques**. *arXiv preprint arXiv:1909.03012*, 2019. (2, 4, 88, 142)

[31] YOSEF ASHIBANI AND QUSAY H MAHMOUD. **Cyber physical systems security: Analysis, challenges and solutions**. *Computers & Security*, **68**:81–97, 2017. ()

[32] JOHN O. AWOYEMI, ADEBAYO O. ADETUNMBI, AND SAMUEL A. OLUWADARE. **Credit card fraud detection using machine learning techniques: A comparative analysis**. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9, 2017. (88)

[33] CLAUDINE BADUE, RÂNIK GUIDOLINI, RAPHAEL VIVACQUA CARNEIRO, PEDRO AZEVEDO, VINICIUS B CARDOSO, AVELINO FORECHI, LUAN JESUS, RODRIGO BERRIEL, THIAGO M PAIXAO, FILIPE MUTZ, ET AL. **Self-driving cars: A survey**. *Expert Systems with Applications*, **165**:113816, 2021. (116)

[34] PRAFULLA BAFNA, DHANYA PRAMOD, AND ANAGHA VAIDYA. **Document clustering: TF-IDF approach**. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE, 2016. ()

[35] MERT BAL, M FATIH AMASYALI, HAYRI SEVER, GUVEN KOSE, AND AYSE DEMIRHAN. **Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system**. *The Scientific World Journal*, **2014**, 2014. ()

[36] ARINDAM BANERJEE, CHASE KRUMPELMAN, JOYDEEP GHOSH, SUGATO BASU, AND RAYMOND J MOONEY. **Model-based overlapping clustering**. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 532–537, 2005. (47)

[37] HUBERT BANIECKI, WOJCIECH KRETOWICZ, PIOTR PIATYSZEK, JAKUB WISNIEWSKI, AND PRZEMYSLAW BIECEK. **dalex: Responsible Machine Learning with**

Interactive Explainability and Fairness in Python. *arXiv preprint arXiv:2012.14406*, 2020. (142)

[38] GAGAN BANSAL, TONGSHUANG WU, JOYCE ZHOU, RAYMOND FOK, BESMIRA NUSHI, ECE KAMAR, MARCO TULIO RIBEIRO, AND DANIEL WELD. **Does the whole exceed its parts? the effect of ai explanations on complementary team performance**. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021. (138)

[39] PIETER BARNARD, NICOLA MARCHETTI, AND LUIZ A DASILVA. **Robust network intrusion detection through explainable artificial intelligence (XAI)**. *IEEE Networking Letters*, **4**(3):167–171, 2022. (23)

[40] NICHOLAS J BELKIN, ROBERT N ODDY, AND HELEN M BROOKS. **ASK for information retrieval: Part I. Background and theory**. *Journal of documentation*, 1982. (121, 127)

[41] VICTORIA BELLOTTI AND KEITH EDWARDS. **Intelligibility and accountability: human considerations in context-aware systems**. *Human–Computer Interaction*, **16**(2-4):193–212, 2001. ()

[42] ABLA CHOUNI BENABDELLAH, ASMAA BENGHABRIT, AND IMANE BOUHADDOU. **A survey of clustering algorithms for an industrial context**. *Procedia computer science*, **148**:291–302, 2019. (xiii, 47, 48)

[43] ALESSIO BENAVOLI, GIORGIO CORANI, AND FRANCESCA MANGILI. **Should we really use post-hoc tests based on mean-ranks?** *The Journal of Machine Learning Research*, **17**(1):152–161, 2016. (74)

[44] CHAYMAE BENFARES, YOUNÈS EL BOUZEKRI EL IDRISSI, AND KARIM HAMID. **Personalized healthcare system based on ontologies**. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pages 185–196. Springer, 2018. ()

[45] JAMES BENNETT, STAN LANNING, ET AL. **The NetFlix prize**. In *Proceedings of KDD cup and workshop*, **2007**, page 35. New York, NY, USA., 2007. ()

[46] H RUSSELL BERNARD AND GERY RYAN. **Text analysis**. *Handbook of methods in cultural anthropology*, **613**, 1998. ()

[47] MASSIMO BERTOLINI, DAVIDE MEZZOGORI, MATTIA NERONI, AND FRANCESCO ZAMMORI. **Machine Learning for industrial applications: A comprehensive literature review**. *Expert Systems with Applications*, **175**:114820, 2021. (2)

[48] RONALD BEST AND HANG ZHANG. **Alternative information sources and the information content of bank loans**. *The Journal of Finance*, **48**(4):1507–1522, 1993. ()

[49] JAMES C BEZDEK. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013. (44)

[50] JAMES C BEZDEK, ROBERT EHRLICH, AND WILLIAM FULL. **FCM: The fuzzy c-means clustering algorithm**. *Computers & Geosciences*, **10**(2-3):191–203, 1984. (xiii, 47, 48)

[51] ISABEL BEZZAOUI, CAROLIN STEIN, CHRISTOF WEINHARDT, AND JONAS FEGERT. **Explainable AI for online disinformation detection: Insights from a design science research project**. *Electronic Markets*, **35**(1):1–28, 2025. (25)

[52] UMANG BHATT, ALICE XIANG, SHUBHAM SHARMA, ADRIAN WELLER, ANKUR TALY, YUNHAN JIA, JOYDEEP GHOSH, RUCHIR PURI, JOSÉ MF MOURA, AND PETER ECKERSLEY. **Explainable machine learning in deployment**. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 648–657, 2020. (137, 138)

[53] SIDDHARTHA BHATTACHARYYA ET AL. **Data mining for credit card fraud: A comparative study**. *Decision Support Systems*, **50**(3):602–613, 2011. (25)

[54] NITIN SAI BOMMI AND ATUL NEGI. **A Standard Baseline for Software Defect Prediction: Using Machine Learning and Explainable AI**. In *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1798–1803. IEEE, 2023. (150)

[55] ALESSANDRO BONDIELLI AND FRANCESCO MARCELLONI. **On the use of summarization and transformer architectures for profiling résumés**. *Expert Systems with Applications*, **184**:115521, 2021. (45)

[56] ANDREA BONDIELLI AND FRANCESCO MARCELLONI. **On the use of summarization and transformer architectures for profiling resumes**. *Information Processing & Management*, **58**(5):102546, 2021. (xiii, 47, 48)

[57] ANDREW P BRADLEY. **The use of the area under the ROC curve in the evaluation of machine learning algorithms**. *Pattern recognition*, **30**(7):1145–1159, 1997. (88)

[58] LEO BREIMAN. **Random forests**. *Machine learning*, **45**(1):5–32, 2001. (64)

[59] ANDREA BRENNEN. **What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.** In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2020. (137)

[60] BRUCE G. BUCHANAN AND EDWARD H. SHORTLIFFE. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1984. (13)

[61] ROBIN BURKE. **Knowledge-based recommender systems**. *Encyclopedia of library and information systems*, **69**(Supplement 32):175–186, 2000. (124)

[62] NIKLAS BUSSMANN, PAOLO GIUDICI, DIMITRI MARINELLI, AND JOCHEN PAPENBROCK. **Explainable machine learning in credit risk management**. *Computational Economics*, **57**(1):203–216, 2021. (22)

[63] BERFU BÜYÜKÖZ, ALI HÜRRIYETOĞLU, AND ARZUCAN ÖZGÜR. **Analyzing ELMo and DistilBERT on socio-political news classification**. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, 2020. ()

[64] LUIS M CAMARINHA-MATOS. **Grand challenges in industrial informatics**, 2023. (1)

[65] RICARDO JGB CAMPELLO, DAVOUD MOULAVI, AND JÖRG SANDER. **Density-based clustering based on hierarchical density estimates**. *Lecture Notes in Computer Science (LNCS)*, **7819**:160–172, 2013. (46, 47)

[66] HÉCTOR CAÑAS, JOSEFA MULA, MANUEL DÍAZ-MADROÑERO, AND FRANCISCO CAMPUZANO-BOLARÍN. **Implementing industry 4.0 principles**. *Computers & industrial engineering*, **158**:107379, 2021. ()

[67] YUANJIANG CAO, XIAOCONG CHEN, LINA YAO, XIANZHI WANG, AND WEI EMMA ZHANG. **Adversarial attacks and detection on reinforcement learning-based interactive recommender systems**. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1669–1672, 2020. (134)

[68] J. CARON ET AL. **Explainable clustering: A survey**. *Artificial Intelligence Review*, **55**:101–147, 2022. (49)

[69] JANE V CARTER, JIANMIN PAN, SHESH N RAI, AND SUSAN GALANDIUK. **ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves**. *Surgery*, **159**(6):1638–1645, 2016. (88)

[70] HORNG-JINH CHANG, LUN-PING HUNG, AND CHIA-LING HO. **An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis**. *Expert systems with applications*, **32**(3):753–764, 2007. ()

[71] SHAO-TUNG CHANG, KANG-PING LU, AND MIIN-SHEN YANG. **Stepwise possibilistic c-regressions**. *Information Sciences*, **334**:307–322, 2016. ()

[72] KINJAL CHAUDHARI AND ANKIT THAKKAR. **A comprehensive survey on travel recommender systems**. *Archives of Computational Methods in Engineering*, pages 1–27, 2019. ()

[73] NITESH V CHAWLA, KEVIN W BOWYER, LAWRENCE O HALL, AND W PHILIP KEGELMEYER. **SMOTE: Synthetic Minority Over-sampling Technique**. In *Journal of Artificial Intelligence Research*, **16**, pages 321–357, 2002. (93)

[74] TIANQI CHEN, TONG HE, MICHAEL BENESTY, VADIM KHOTILOVICH, YUAN TANG, HYUNSU CHO, KAILONG CHEN, RORY MITCHELL, IGNACIO CANO, TIANYI ZHOU, ET AL. **Xgboost: extreme gradient boosting**. *R package version 0.4-2*, **1**(4):1–4, 2015. (93)

[75] WEISI CHEN, ZORAN MILOSEVIC, FETHI A RABHI, AND ANDREW BERRY. **Real-time analytics: Concepts, architectures, and ML/AI considerations**. *IEEE Access*, **11**:71634–71657, 2023. ()

[76] DAVIDE CHICCO AND GIUSEPPE JURMAN. **The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation**. *BMC genomics*, **21**:1–13, 2020. ()

[77] DAVIDE CHICCO AND GIUSEPPE JURMAN. **The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification**. *BioData Mining*, **16**(1):4, 2023. ()

[78] SRIKRISHNA CHINTALAPATI AND SHIVENDRA KUMAR PANDEY. **Artificial intelligence in marketing: A systematic literature review**. *International Journal of Market Research*, **64**(1):38–68, 2022. (31)

[79] ALEXANDER CHISLER, YULIA VOLKOVA, AND EVGENY PYSHKIN. **Handle with it addiction: A browser extension for overcoming excessive tv-series streaming**. In *Proceedings of the 12th Central and Eastern European Software Engineering Conference in Russia*, pages 1–5, 2016. ()

[80] MAURO PETER HENRIQUE CHISSANO AND JOHN MINNERY. **Roads, rates and development: Urban roads and growth in Xai-Xai, Mozambique**. *Habitat International*, **42**:48–57, 2014. ()

[81] HYUNJI CHUNG AND SANGJIN LEE. **Intelligent virtual assistant knows your life**. *arXiv preprint arXiv:1803.00466*, 2018. (112)

[82] JÜRGEN CITO, ISIL DILLIG, VIJAYARAGHAVAN MURALI, AND SATISH CHANDRA. **Counterfactual explanations for models of code**. In *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice*, pages 125–134, 2022. (149)

[83] WILLIAM J CLANCEY AND ROBERT R HOFFMAN. **Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems**. *Applied AI letters*, **2**(4):e53, 2021. (23)

[84] PAMELA K COATS. **Why expert systems fail**. *Financial Management*, pages 77–86, 1988. (121)

[85] JOSEPH COHEN, XUN HUAN, AND JUN NI. **Shapley-based explainable ai for clustering applications in fault diagnosis and prognosis**. *Journal of Intelligent Manufacturing*, pages 1–16, 2024. (45)

[86] DENNIS COLLARIS AND JARKE J VAN WIJK. **Machine learning interpretability through contribution-value plots**. In *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*, pages 1–5, 2020. (16)

[87] CRISTINA CONATI, OSWALD BARRAL, VANESSA PUTNAM, AND LEA RIEGER. **Toward personalized XAI: A case study in intelligent tutoring systems**. *Artificial intelligence*, **298**:103503, 2021. (23)

[88] MARK G CORE, H CHAD LANE, MICHAEL VAN LENT, DAVE GOMBOC, STEVE SOLOMON, MILTON ROSENBERG, ET AL. **Building explainable artificial intelligence systems**. In *AAAI*, pages 1766–1773, 2006. (xi, 132)

[89] RICHARD M CORMACK. **A review of classification**. *Journal of the Royal Statistical Society: Series A (General)*, **134**(3):321–353, 1971. ()

[90] THOMAS COVER AND PETER HART. **Nearest neighbor pattern classification**. *IEEE transactions on information theory*, **13**(1):21–27, 1967. (30)

[91] MARK WILLIAM CRAVEN. *Extracting comprehensible models from trained neural networks*. The University of Wisconsin-Madison, 1996. ()

[92] CHRISTINE M CUTILLO, KARLIE R SHARMA, LUCA FOSCHINI, SHINJINI KUNDU, MAXINE MACKINTOSH, AND KENNETH D MANDL. **Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency**. *NPJ Digital Medicine*, **3**(1):1–5, 2020. (137)

[93] SEYYED MOHAMMAD HOSSEIN DADGAR, MOHAMMAD SHIRZAD ARAGHI, AND MORTEZA MASTERY FARAHANI. **A novel text mining approach based on TF-IDF and Support Vector Machine for news classification**. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, pages 112–116. IEEE, 2016. (30)

[94] DEBASHIS DAS, LAXMAN SAHOO, AND SUJOY DATTA. **A survey on recommendation system**. *International Journal of Computer Applications*, **160**(7), 2017. (124)

[95] LUIS DE-MARCOS, ADRIÁN DOMÍNGUEZ, JOSEBA SAENZ-DE NAVARRETE, AND CARMEN PAGÉS. **An empirical study comparing gamification and social networking on e-learning**. *Computers & education*, **75**:82–91, 2014. ()

[96] DASWIN DE SILVA, SEPPO SIERLA, DAMMINDA ALAHAKOON, EVGENY OSIPOV, XINGHUO YU, AND VALERIY VYATKIN. **Toward intelligent industrial informatics: A review of current developments and future directions of artificial intelligence in industrial applications**. *IEEE Industrial Electronics Magazine*, **14**(2):57–72, 2020. ()

[97] ETHEL-MICHELE DE VILLIERS, CLAUDE FAUQUET, THOMAS R BROKER, HANS-ULRICH BERNARD, AND HARALD ZUR HAUSEN. **Classification of papillomaviruses**. *Virology*, **324**(1):17–27, 2004. ()

[98] GUILHERME DEAN PELEGRINA AND SAJID SIRAJ. **Shapley Value-Based Approaches to Explain the Quality of Predictions by Classifiers** . *IEEE Transactions on Artificial Intelligence*, **5**(08):4217–4231, August 2024. (88)

[99] ASHLEY DEEKS. **The judicial demand for explainable artificial intelligence**. *Columbia Law Review*, **119**(7):1829–1850, 2019. (143)

[100] MARÍA DEL MAR ROLDÁN-GARCÍA, JOSÉ GARCÍA-NIETO, AND JOSÉ F ALDANA-MONTES. **Enhancing semantic consistency in anti-fraud rule-based expert systems**. *Expert Systems with Applications*, **90**:332–343, 2017. ()

[101] SOLEDAD DELGADO, CLARA HIGUERA, JORGE CALLE-ESPINOSA, FEDERICO MORÁN, AND FRANCISCO MONTERO. **A SOM prototype-based cluster analysis methodology**. *Expert Systems with Applications*, **88**:14–28, 2017. (46)

[102] SEBASTIAN DETERDING, DAN DIXON, RILLA KHALED, AND LENNART NACKE. **From game design elements to gamefulness: defining" gamification"**. In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15, 2011. ()

[103] LUC DEVROYE. **On the Inequality of Cover and Hart in Nearest Neighbor Discrimination**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-3**(1):75–78, 1981. (30)

[104] VARUN DOGRA, AMAN SINGH, SAHIL VERMA, KAVITA, NZ JHANJHI, AND MN TALIB. **Analyzing DistilBERT for sentiment classification of banking financial news**. In *Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021*, pages 501–510. Springer, 2021. ()

[105] WATERMAN A DONALD ET AL. **A guide to expert systems**. *Addison-Wesley: Reading, MA). Fay R, Treloar G and Iyer-Raniga U (2000) Life-cycle energy analysis of buildings: a case study. Building Research and Information*, **28**(1):31–41, 1986. ()

[106] JIQIAN DONG, SIKAI CHEN, MOHAMMAD MIRALINAGHI, TIANTIAN CHEN, PEI LI, AND SAMUEL LABI. **Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems**. *Transportation research part C: emerging technologies*, **156**:104358, 2023. (23)

[107] FINALE DOSHI-VELEZ AND BEEN KIM. **Towards a rigorous science of interpretable machine learning**. *arXiv preprint arXiv:1702.08608*, 2017. (7, 31)

[108] FINALE DOSHI-VELEZ AND BEEN KIM. **Towards a rigorous science of interpretable machine learning**. `https://arxiv.org/abs/1702.08608`, 2017. arXiv preprint arXiv:1702.08608, Accessed: 2025-06-30. (18, 24)

[109] FILIP KARLO DOŠILOVIĆ, MARIO BRČIĆ, AND NIKICA HLUPIĆ. **Explainable artificial intelligence: A survey**. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018. ()

[110] MAURO DRAGONI, IVAN DONADELLO, AND CLAUDIO ECCHER. **Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice**. *Artificial Intelligence in Medicine*, **105**:101840, 2020. (127)

[111] DANIEL J DUBOIS AND GIORDANO TAMBURRELLI. **Understanding gamification mechanisms for software development**. In *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, pages 659–662, 2013. ()

[112] PIERPAOLO D'URSO. **Informational Paradigm, management of uncertainty and theoretical formalisms in the clustering framework: A review**. *Information Sciences*, **400**:30–62, 2017. ()

[113] RUDRESH DWIVEDI, DEVAM DAVE, HET NAIK, SMITI SINGHAL, RANA OMER, PANKESH PATEL, BIN QIAN, ZHENYU WEN, TEJAL SHAH, GRAHAM MORGAN, ET AL. **Explainable AI (XAI): Core ideas, techniques, and solutions**. *ACM Computing Surveys*, **55**(9):1–33, 2023. ()

[114] YOGESH K DWIVEDI, LAURIE HUGHES, ELVIRA ISMAGILOVA, GERT AARTS, CRISPIN COOMBS, TOM CRICK, YANQING DUAN, ROHITA DWIVEDI, JOHN EDWARDS, ALED EIRUG, ET AL. **Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy**. *International journal of information management*, **57**:101994, 2021. (86, 130)

[115] EROL EGRIOGLU, CH ALADAG, UFUK YOLCU, VEDIDE R USLU, AND N ALP ERILLI. **Fuzzy time series forecasting method based on Gustafson–Kessel fuzzy clustering**. *Expert Systems with Applications*, **38**(8):10355–10357, 2011. ()

[116] WAFAA S EL-KASSAS, CHERIF R SALAMA, AHMED A RAFEA, AND HODA K MOHAMED. **Automatic text summarization: A comprehensive survey**. *Expert systems with applications*, **165**:113679, 2021. ()

[117] CHARLES A ELLIS, MOHAMMAD SE SENDI, ELOY GEENJAAR, SERGEY M PLIS, ROBYN L MILLER, AND VINCE D CALHOUN. **Algorithm-agnostic explainability for unsupervised clustering**. *arXiv preprint arXiv:2105.08053*, 2021. (45)

[118] OSSAMA EMBARAK. **Explainable Artificial Intelligence for Services Exchange in Smart Cities**. In *Explainable Artificial Intelligence for Smart Cities*, pages 13–30. CRC Press, 2021. ()

[119] SEAN EOM AND E KIM. **A survey of decision support system applications (1995–2001)**. *Journal of the Operational Research Society*, **57**:1264–1278, 2006. (113)

[120] MC ER. **Decision support systems: a summary, problems, and future trends**. *Decision support systems*, **4**(3):355–363, 1988. (116)

[121] MARTIN ESTER, HANS-PETER KRIEGEL, JÖRG SANDER, AND XIAOWEI XU. **A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise**. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231, 1996. (46, 47)

[122] GLENN J FALA, KATHRYN T CLAYTON, AND DIANE M MASCIANTONIO. **Applying expert systems to health care management**. In *Proceedings of the 1995 ACM symposium on Applied computing*, pages 237–241, 1995. ()

[123] XIAODONG FAN, SOOYOUNG OH, MICHAEL MCNEESE, JOHN YEN, HAYDEE CUEVAS, LAURA STRATER, AND MICA R ENDSLEY. **The influence of agent reliability on trust in human-agent collaboration**. In *Proceedings of the 15th European conference on Cognitive ergonomics: the ergonomics of cool interaction*, pages 1–8, 2008. (137)

[124] SEBASTIAN FARQUHAR, JANNIK KOSSEN, LORENZ KUHN, AND YARIN GAL. **Detecting hallucinations in large language models using semantic entropy**. *Nature*, **630**(8017):625–630, 2024. (43)

[125] SHAGUFTA FATEMA AND MOHAMMAD MERAJ AHMAD KHAN. **Artificial Intelligence in Talent Acquisition: Assessing the Impact on Recruitment Processes**. *Journal of Research and Innovation in Technology and Management (JRITM)*, **1**(3):1–5, 2024. (83)

[126] EDWARD FEIGENBAUM, PAMELA MCCORDUCK, AND H PENNY NII. *The rise of the expert company*. Times Books, 1988. ()

[127] EDWARD A FEIGENBAUM. **Expert systems in the 1980s**. *State of the art report on machine intelligence. Maidenhead: Pergamon-Infotech*, 1981. (117)

[128] EDWARD A FEIGENBAUM AND PAMELA MCCORDUCK. *The fifth generation*. Addison-Wesley Pub., 1983. ()

[129] RONEN FELDMAN AND JAMES SANGER. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007. (30)

[130] ALBERTO FERNANDEZ, FRANCISCO HERRERA, OSCAR CORDON, MARIA JOSE DEL JESUS, AND FRANCESCO MARCELLONI. **Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?** *IEEE Computational intelligence magazine*, **14**(1):69–81, 2019. ()

[131] RICHARD FORSYTH. **The architecture of expert systems**. *Expert systems: principles and case studies*, pages 9–17, 1984. ()

[132] DALLAS FRASER, ANDREW KANE, AND FRANK WM TOMPA. **Choosing math features for BM25 ranking with Tangent-L**. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–10, 2018. (45, 52)

[133] SORELLE A FRIEDLER, CARLOS SCHEIDEGGER, SURESH VENKATASUBRAMANIAN, SONAM CHOUDHARY, EVAN P HAMILTON, AND DEREK ROTH. **A comparative study of fairness-enhancing interventions in machine learning**. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019. (30)

[134] DANIEL FRYER, INGA STRÜMKE, AND HIEN NGUYEN. **Shapley values for feature selection: The good, the bad, and the axioms**. *Ieee Access*, **9**:144352–144360, 2021. (86, 88)

[135] BUDDHIMA GAMLATH, XINRUI JIA, ADAM POLAK, AND OLA SVENSSON. **Nearly-tight and oblivious algorithms for explainable clustering**. *Advances in Neural Information Processing Systems*, **34**:28929–28939, 2021. ()

[136] PRAKHAR GANESH, YAO CHEN, XIN LOU, MOHAMMAD ALI KHAN, YIN YANG, HASSAN SAJJAD, PRESLAV NAKOV, DEMING CHEN, AND MARIANNE WINSLETT. **Compressing large-scale transformer-based models: A case study on bert**. *Transactions of the Association for Computational Linguistics*, **9**:1061–1080, 2021. ()

[137] DAMIEN GARREAU AND ULRIKE LUXBURG. **Explaining the explainer: A first theoretical analysis of LIME**. In *International conference on artificial intelligence and statistics*, pages 1287–1296. PMLR, 2020. ()

[138] ISAK GATH AND AMIR B. GEVA. **Unsupervised optimal fuzzy clustering**. *IEEE Transactions on pattern analysis and machine intelligence*, **11**(7):773–780, 1989. (47)

[139] MOHAMED YACINE GHERAIBIA AND CHARLES GOUIN-VALLERAND. **Intelligent mobile-based recommender system framework for smart freight transport**. In *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, pages 219–222, 2019. ()

[140] LEILANI H GILPIN, DAVID BAU, BEN Z YUAN, AYESHA BAJWA, MICHAEL SPECTER, AND LALANA KAGAL. **Explaining explanations: An overview of interpretability of machine learning**. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. ()

[141] MICHAEL GLEICHER. **A framework for considering comprehensibility in modeling**. *Big data*, **4**(2):75–88, 2016. ()

[142] STEVEN GLOBERMAN, THOMAS W ROEHL, AND STEPHEN STANDIFIRD. **Globalization and electronic commerce: Inferences from retail brokering**. *Journal of International Business Studies*, **32**(4):749–768, 2001. ()

[143] DAVID GOLDBERG, DAVID NICHOLS, BRIAN M OKI, AND DOUGLAS TERRY. **Using collaborative filtering to weave an information tapestry**. *Communications of the ACM*, **35**(12):61–70, 1992. (121)

[144] S MALEK MOHAMADI GOLSEFID, MH FAZEL ZARANDI, AND IB TURKSEN. **Multi-central general type-2 fuzzy clustering approach for pattern recognitions**. *Information Sciences*, **328**:172–188, 2016. ()

[145] BRYCE GOODMAN AND SETH FLAXMAN. **European Union regulations on algorithmic decision-making and a "right to explanation"**. *AI magazine*, **38**(3):50–57, 2017. ()

[146] ALLAN DAVID GORDON. *Classification*. CRC Press, 1999. ()

[147] ANJANA GOSAIN AND SONIKA DAHIYA. **Performance analysis of various fuzzy clustering algorithms: a review**. *Procedia Computer Science*, **79**:100–111, 2016. (48)

[148] R. GUIDOTTI ET AL. **A survey of methods for explaining black box models**. *ACM Computing Surveys*, **51**(5):1–42, 2018. (49)

[149] ASELA GUNAWARDANA AND GUY SHANI. **A survey of accuracy evaluation metrics of recommendation tasks.** *Journal of Machine Learning Research*, **10**(12), 2009. ()

[150] DAVID GUNNING. **Explainable Artificial Intelligence (XAI)**. `https://www.darpa.mil/program/explainable-artificial-intelligence`, 2017. Accessed: 2025-06-30. (14)

[151] DAVID GUNNING AND DAVID AHA. **DARPA's explainable artificial intelligence (XAI) program**. *AI Magazine*, **40**(2):44–58, 2019. (127, 130, 141)

[152] DAVID GUNNING, MARK STEFIK, JAESIK CHOI, TIMOTHY MILLER, SIMONE STUMPF, AND GUANG-ZHONG YANG. **XAI—Explainable artificial intelligence**. *Science Robotics*, **4**(37), 2019. ()

[153] CHARU GUPTA, AMITA JAIN, AND NISHEETH JOSHI. **Fuzzy logic in natural language processing–a closer view**. *Procedia computer science*, **132**:1375–1384, 2018. ()

[154] DONALD E GUSTAFSON AND WILLIAM C KESSEL. **Fuzzy clustering with a fuzzy covariance matrix**. In *1978 IEEE conference on decision and control including the 17th symposium on adaptive processes*, pages 761–766. IEEE, 1979. (xiii, 26, 45, 47, 48, 61, 64)

[155] ZHONGYANG HAN, JUN ZHAO, QUANLI LIU, AND WEI WANG. **Granular-computing based hybrid collaborative fuzzy clustering for long-term prediction of multiple gas holders levels**. *Information Sciences*, **330**:175–185, 2016. ()

[156] NITIN HARDENIYA, JACOB PERKINS, DEEPTI CHOPRA, NISHEETH JOSHI, AND ITI MATHUR. *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016. (34, 52)

[157] J. A. HARDING, M. SHAHBAZ, SRINIVAS, AND A. KUSIAK. **Data Mining in Manufacturing: A Review**. *Journal of Manufacturing Science and Engineering*, **128**(4):969–976, 12 2006. (2)

[158] RICHARD J HATHAWAY, JAMES C BEZDEK, AND YINGKANG HU. **Generalized fuzzy c-means clustering strategies using L/sub p/norm distances**. *IEEE transactions on Fuzzy Systems*, **8**(5):576–582, 2000. ()

[159] FRANCISCO HERRERA, FRANCISCO CHARTE, ANTONIO J RIVERA, MARÍA J DEL JESUS, FRANCISCO HERRERA, FRANCISCO CHARTE, ANTONIO J RIVERA, AND MARÍA J DEL JESUS. *Multilabel classification*. Springer, 2016. ()

[160] LENNART HOFEDITZ, SÜNJE CLAUSEN, ALEXANDER RIESS, MILAD MIRBABAIE, AND STEFAN STIEGLITZ. **Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring**. *Electronic Markets*, **32**(4):2207–2233, 2022. (23)

[161] ANDREAS HOLZINGER, ANNA SARANTI, CHRISTOPH MOLNAR, PRZEMYSLAW BIECEK, AND WOJCIECH SAMEK. **Explainable AI methods-a brief overview**. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 13–38. Springer, 2020. ()

[162] DANILO HORTA AND RICARDO CAMPELLO. **Comparing hard and overlapping clusterings**. *Journal of Machine Learning Research*, **16**:2949–2997, 2015. (47)

[163] JUNYAN HU, PARIJAT BHOWMICK, INMO JANG, FARSHAD ARVIN, AND ALEXANDER LANZON. **A decentralized cluster formation containment framework for multirobot systems**. *IEEE Transactions on Robotics*, **37**(6):1936–1955, 2021. (143)

[164] JUNYAN HU, PARIJAT BHOWMICK, AND ALEXANDER LANZON. **Group Coordinated Control of Networked Mobile Robots With Applications to Object Transportation**. *IEEE Transactions on Vehicular Technology*, **70**(8):8269–8274, 2021. (143)

[165] HUAJIE HUANG, BO LIU, XIAOYU XUE, JIUXIN CAO, AND XINYI CHEN. **Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique**. *Applied Soft Computing*, **154**:111368, 2024. (5)

[166] JOONAS HÄMÄLÄINEN, SUSANNE JAUHIAINEN, AND TOMMI KÄRKKÄINEN. **Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering**. *Algorithms*, **10**(3), 2017. ()

[167] HESAM IZAKIAN, WITOLD PEDRYCZ, AND IQBAL JAMAL. **Clustering spatiotemporal data: An augmented fuzzy c-means**. *IEEE transactions on fuzzy systems*, **21**(5):855–868, 2012. ()

[168] ANIL K JAIN AND RICHARD C DUBES. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988. ()

[169] SARIKA JAIN, ANJALI GROVER, PRAVEEN SINGH THAKUR, AND SOURABH KUMAR CHOUDHARY. **Trends, problems and solutions of recommender system**. In *International Conference on Computing, Communication & Automation*, pages 955–958. IEEE, 2015. (124)

[170] DIETMAR JANNACH, MARKUS ZANKER, ALEXANDER FELFERNIG, AND GERHARD FRIEDRICH. *Recommender systems: an introduction*. Cambridge University Press, 2010. (123, 125)

[171] SÉRGIO JESUS, CATARINA BELÉM, VLADIMIR BALAYAN, JOÃO BENTO, PEDRO SALEIRO, PEDRO BIZARRO, AND JOÃO GAMA. **How can I choose an explainer? An application-grounded evaluation of post-hoc explanations**. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 805–815, 2021. (15)

[172] KAREL JEŽEK AND JOSEF STEINBERGER. **Automatic text summarization (the state of the art 2007 and new challenges)**. In *Proceedings of Znalosti*, pages 1–12, 2008. ()

[173] TAEHO JO. **Using K Nearest Neighbors for text segmentation with feature similarity**. In *2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pages 1–5. IEEE, 2017. (31)

[174] SUN JUAN, CHEN QUANGONG, LONG RUIJUN, AND JIANG WENLAN. **An application of the analytic hierarchy process and fuzzy logic inference in a decision support system for forage selection**. *New Zealand Journal of Agricultural Research*, **47**(3):327–331, 2004. ()

[175] BERNAD JUMADI DEHOTMAN SITOMPUL, OPIM SALIM SITOMPUL, AND POLTAK SIHOMBING. **Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm**. In *Journal of Physics: Conference Series*, **1235**, page 012015. IOP Publishing, 2019. (63)

[176] M HUMAYN KABIR, KHONDOKAR FIDA HASAN, MOHAMMAD KAMRUL HASAN, AND KEYVAN ANSARI. **Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform**. *arXiv preprint arXiv:2111.00601*, 2021. ()

[177] SINAN KAPLAN, HANNU UUSITALO, AND LASSE LENSU. **A unified and practical user-centric framework for explainable artificial intelligence**. *Knowledge-Based Systems*, **283**:111107, 2024. ()

[178] NICOLAOS B KARAYIANNIS. **MECA: Maximum entropy clustering algorithm**. In *Proceedings of 1994 IEEE 3rd international fuzzy systems conference*, pages 630–635. IEEE, 1994. ()

[179] LEONARD KAUFMAN AND PETER J ROUSSEEUW. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009. ()

[180] HARMANPREET KAUR, HARSHA NORI, SAMUEL JENKINS, RICH CARUANA, HANNA WALLACH, AND JENNIFER WORTMAN VAUGHAN. **Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning**. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020. (138)

[181] RAFE KETTLER. **A Guide to Python's Magic Methods**. *URL: https://rszalski. github. io/magicmethods/# intro*, 2015. ()

[182] FARHINA SARDAR KHAN, SYED SHAHID MAZHAR, KASHIF MAZHAR, DHOHA A. ALSALEH, AND AMIR MAZHAR. **Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions**. *Artificial Intelligence Review*, **58**(8):232, 2025. ()

[183] LAITH T KHRAIS. **Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce**. *Future Internet*, **12**(12):226, 2020. ()

[184] SHAH KHUSRO, ZAFAR ALI, AND IRFAN ULLAH. **Recommender systems: issues, challenges, and research opportunities**. In *Information Science and Applications (ICISA) 2016*, pages 1179–1189. Springer, 2016. (124, 125)

[185] BUOMSOO KIM, JINSOO PARK, AND JIHAE SUH. **Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information**. *Decision Support Systems*, **134**:113302, 2020. (23)

[186] YOUNG-IL KIM, DAE-WON KIM, DOHEON LEE, AND KWANG H LEE. **A cluster validation index for GK cluster analysis based on relative degree of sharing**. *Information Sciences*, **168**(1-4):225–242, 2004. ()

[187] SHAKTI KINGER AND VRUSHALI KULKARNI. **Explainable AI for Deep Learning Based Disease Detection**. In *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*, pages 209–216, 2021. ()

[188] LAWRENCE A KLEIN, PING YI, AND HUALIANG TENG. **Decision support system for advanced traffic management through data fusion**. *Transportation Research Record*, **1804**(1):173–178, 2002. ()

[189] SUNIL KUMAR KOPPARAPU. **Automatic extraction of usable information from unstructured resumes to aid search**. In *2010 IEEE International Conference on Progress in Informatics and Computing*, **1**, pages 99–103. IEEE, 2010. (30)

[190] BASTIAN KRÄMER, MORITZ STANG, CATHRINE NAGL, AND WOLFGANG SCHÄFERS. **Explainable AI in a Real Estate Context-Exploring the Determinants of Residential Real Estate Values**. *a Real Estate Context-Exploring the Determinants of Residential Real Estate Values (December 20, 2021)*, 2021. (134)

[191] RAGHU KRISHNAPURAM AND JONGWOO KIM. **A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms**. *IEEE Transactions on Fuzzy systems*, **7**(4):453–461, 1999. ()

[192] RAGHURAM KRISHNAPURAM AND JAMES M KELLER. **A possibilistic approach to clustering**. *IEEE transactions on fuzzy systems*, **1**(2):98–110, 1993. ()

[193] KOSTIANTYN KUCHER, ELMIRA ZOHREVANDI, AND CARL AL WESTIN. **Towards Visual Analytics for Explainable AI in Industrial Applications**. *Analytics*, **4**(1):7, 2025. (2)

[194] UDO KUCKARTZ. *Qualitative text analysis: A guide to methods, practice and using software*. Sage, 2014. ()

[195] OUREN KUIPER, MARTIN VAN DEN BERG, JOOST VAN DEN BURGT, AND STEFAN LEIJNEN. **Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities**. *arXiv preprint arXiv:2111.02244*, 2021. ()

[196] ANDREW KUSIAK. **Smart manufacturing must embrace big data**. *Nature*, **544**(7648):23–25, 2017. (2)

[197] ISAAC LAGE, EMILY CHEN, JEFFREY HE, MENAKA NARAYANAN, BEEN KIM, SAM GERSHMAN, AND FINALE DOSHI-VELEZ. **An evaluation of the human-interpretability of explanation**. *arXiv preprint arXiv:1902.00006*, 2019. (131)

[198] SOANPET SREE LAKSHMI AND T ADI LAKSHMI. **Recommendation systems: Issues and challenges**. *International Journal of Computer Science and Information Technologies*, **5**(4):5771–5772, 2014. (125)

[199] YANN LECUN, YOSHUA BENGIO, AND GEOFFREY HINTON. **Deep learning**. *nature*, **521**(7553):436–444, 2015. (15)

[200] HEA IN LEE, IL YOUNG CHOI, HYUN SIL MOON, AND JAE KYEONG KIM. **A multi-period product recommender system in the online food market based on recurrent neural networks**. *Sustainability*, **12**(3):969, 2020. (134)

[201] JAY LEE, BEHRAD BAGHERI, AND HUNG-AN KAO. **Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics**. In *International proceeding of int conference on industrial informatics (INDIN)*, pages 1–6. Citeseer, 2014. ()

[202] GEYU LIANG, SENNE MICHIELSSEN, AND SALAR FATTAHI. **Enhancing Performance of Explainable AI Models with Constrained Concept Refinement**. *arXiv preprint arXiv:2502.06775*, 2025. ()

[203] TING-PENG LIANG AND CHIH-PING WEI. **Introduction to the special issue: Mobile commerce applications**. *International journal of electronic commerce*, **8**(3):7–17, 2004. ()

[204] Q VERA LIAO, DANIEL GRUEN, AND SARAH MILLER. **Questioning the AI: informing design practices for explainable AI user experiences**. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020. (138)

[205] ANDY LIAW AND MATTHEW WIENER. **Classification and regression by randomForest**. *R news*, **2**(3):18–22, 2002. (64)

[206] TOM CW LIN. **The new investor**. *UCLA L. Rev.*, **60**:678, 2012. (143)

[207] ZACHARY C LIPTON. **The Mythos of Model Interpretability. Queue 16, 3, Article 30 (June 2018), 27 pages**, 2018. (64, 131)

[208] DAVIDE LIU, GEORGE PHILIPPE FARAJALLA, AND ALEXANDRE BOULENGER. **Transformer-based Banking Products Recommender System**. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2641–2642, 2021. ()

[209] FEI TONY LIU, KAI MING TING, AND ZHI-HUA ZHOU. **Isolation forest**. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. (89)

[210] MING LIU, HONGJUN ZHANG, ZESHUI XU, AND KUN DING. **The fusion of fuzzy theories and natural language processing: A state-of-the-art survey**. *Applied Soft Computing*, **162**:111818, 2024. ()

[211] XIAODONG LIU, XIANCHANG WANG, AND WITOLD PEDRYCZ. **Fuzzy clustering with semantic interpretation**. *Applied Soft Computing*, **26**:21–30, 2015. ()

[212] YANG LIU AND MIRELLA LAPATA. **Text summarization with pretrained encoders**. *arXiv preprint arXiv:1908.08345*, 2019. ()

[213] OCTAVIO LOYOLA-GONZALEZ, ANDRES EDUARDO GUTIERREZ-RODRÍGUEZ, MIGUEL ANGEL MEDINA-PÉREZ, RAUL MONROY, JOSÉ FRANCISCO MARTÍNEZ-TRINIDAD, JESÚS ARIEL CARRASCO-OCHOA, AND MILTON GARCIA-BORROTO. **An explainable artificial intelligence model for clustering numerical databases**. *IEEE Access*, **8**:52370–52384, 2020. (46)

[214] SIMONE A LUDWIG. **MapReduce-based fuzzy c-means clustering algorithm: implementation and scalability**. *International journal of machine learning and cybernetics*, **6**:923–934, 2015. ()

[215] SZYMON ŁUKASIK, PIOTR A KOWALSKI, MAŁGORZATA CHARYTANOWICZ, AND PIOTR KULCZYCKI. **Clustering using flower pollination algorithm and Calinski-Harabasz index**. In *2016 IEEE congress on evolutionary computation (CEC)*, pages 2724–2728. IEEE, 2016. (47, 63)

[216] SCOTT M LUNDBERG AND SU-IN LEE. **A unified approach to interpreting model predictions**. *Advances in neural information processing systems*, **30**, 2017. (2, 4, 7, 13, 15, 16, 43, 66, 86, 87, 88, 90, 95, 96, 142)

[217] J. MACQUEEN. **Some methods for classification and analysis of multivariate observations**. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, pages 281–297. University of California Press, 1967. (47)

[218] HADEER MAHMOUD, ABDELFATAH HEGAZY, AND MOHAMED H KHAFAGY. **An approach for big data security based on Hadoop distributed file system**. In *2018 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 109–114. IEEE, 2018. (116, 125)

[219] MOSTAFA MAHMOUD, NAJLA ALGADI, AND AHMED ALI. **Expert system for banking credit decision**. In *2008 International conference on computer science and information technology*, pages 813–819. IEEE, 2008. ()

[220] SARA RAGAB MAHMOUD AND CHIN KIM GAN. **Classification of solar variability using K-means method for the evaluation of solar photovoltaic systems performance**. *International Journal of Renewable Energy Research*, **12**:692–702, 2022. ()

[221] MOUTUSY MAITY AND MAYUKH DASS. **Consumer decision-making across modern and traditional channels: E-commerce, m-commerce, in-store**. *Decision Support Systems*, **61**:34–46, 2014. ()

[222] SAURAV MANCHANDA AND GEORGE KARYPIS. **Text segmentation on multilabel documents: A distant-supervised approach**. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1170–1175. IEEE, 2018. (31)

[223] XIANGKE MAO, HUI YANG, SHAOBIN HUANG, YE LIU, AND RONGSHENG LI. **Extractive summarization using supervised and unsupervised learning**. *Expert systems with applications*, **133**:173–181, 2019. (47)

[224] GEORGE M MARAKAS. *Decision support systems in the 21st century*, **134**. Prentice Hall Upper Saddle River, 2003. (113)

[225] UJJWAL MAULIK AND SANGHAMITRA BANDYOPADHYAY. **Performance evaluation of some clustering algorithms and validity indices**. *IEEE Transactions on pattern analysis and machine intelligence*, **24**(12):1650–1654, 2002. (63)

[226] GEOFFREY J MCLACHLAN AND DAVID PEEL. *Finite Mixture Models*. Wiley, 2000. (46, 47)

[227] MATTHIAS R MEHL. **Quantitative text analysis.** *Handbook of multimethod measurement in psychology*, page 141–156, 2006. ()

[228] NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, AND ARAM GALSTYAN. **A survey on bias and fairness in machine learning**. *ACM computing surveys (CSUR)*, **54**(6):1–35, 2021. (30)

[229] PREM MELVILLE AND VIKAS SINDHWANI. **Recommender systems.** *Encyclopedia of machine learning*, **1**:829–838, 2010. ()

[230] MELKAMU ABAY MERSHA, JUGAL KALITA, ET AL. **Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms**. *Procedia Computer Science*, **244**:121–132, 2024. (45)

[231] CHRISTIAN MESKE AND ENRICO BUNDE. **Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support**. In *International Conference on Human-Computer Interaction*, pages 54–69. Springer, 2020. (134)

[232] JIAJU MIAO AND WEI ZHU. **Precision–recall curve (PRC) classification trees**. *Evolutionary intelligence*, **15**(3):1545–1569, 2022. (88)

[233] RADA MIHALCEA AND PAUL TARAU. **Textrank: Bringing order into text**. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004. (52)

[234] DEREK MILLER. **Leveraging BERT for extractive text summarization on lectures**. *arXiv preprint arXiv:1906.04165*, 2019. (53)

[235] TIM MILLER. **Explanation in artificial intelligence: Insights from the social sciences**. *Artificial Intelligence*, **267**:1–38, 2019. (19, 87)

[236] SUELI A MINGOTI AND JOAB O LIMA. **Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms**. *European journal of operational research*, **174**(3):1742–1759, 2006. (44)

[237] MARVIN MINSKY, RAY KURZWEIL, AND STEVE MANN. **The society of intelligent veillance**. In *2013 IEEE International Symposium on Technology and Society (ISTAS): Social Implications of Wearable Computing and Augmediated Reality in Everyday Life*, pages 13–17. IEEE, 2013. ()

[238] MARWA HUSSIEN MOHAMED, MOHAMED HELMY KHAFAGY, AND MOHAMED HASAN IBRAHIM. **Recommender systems challenges and solutions survey**. In *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, pages 149–155. IEEE, 2019. ()

[239] AHMAD HAJI MOHAMMADKHANI, NITIN SAI BOMMI, MARIEM DABOUSSI, ONKAR SABNIS, CHAKKRIT TANTITHAMTHAVORN, AND HADI HEMMATI. **A systematic literature review of explainable AI for software engineering**. *arXiv preprint arXiv:2302.06065*, 2023. (148)

[240] ATHAR HUSSEIN MOHAMMED AND ALI H ALI. **Survey of bert (bidirectional encoder representation transformer) types**. In *Journal of Physics: Conference Series*, **1963**, page 012173. IOP Publishing, 2021. ()

[241] DIANA-ADERINA MOISUC AND MIHAI-CONSTANTIN AVORNICULUI. **Architectural model of expert systems**. In *5th International Symposium Engineering Management and Competitiveness*, 2015. (116)

[242] CHRISTOPH MOLNAR. *Interpretable Machine Learning*. Lulu.com, 2020. (18, 64)

[243] GRÉGOIRE MONTAVON, WOJCIECH SAMEK, AND KLAUS-ROBERT MÜLLER. **Methods for interpreting and understanding deep neural networks**. *Digital Signal Processing*, **73**:1–15, 2018. ()

[244] SAJAD MOOSAVI, MARYAM FARAJZADEH-ZANJANI, ROOZBEH RAZAVI-FAR, VASILE PALADE, AND MEHRDAD SAIF. **Explainable AI in Manufacturing and Industrial Cyber–Physical Systems: A Survey**. *Electronics*, **13**(17), 2024. (1)

[245] JULIA MOOSBAUER, JULIA HERBINGER, GIUSEPPE CASALICCHIO, MARIUS LINDAUER, AND BERND BISCHL. **Explaining hyperparameter optimization via partial dependence plots**. *Advances in neural information processing systems*, **34**:2280–2291, 2021. (17)

[246] ANDREA MORICHETTA, PEDRO CASAS, AND MARCO MELLIA. **EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis**. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, pages 22–28, 2019. (45)

[247] MICHAL MOSHKOVITZ, SANJOY DASGUPTA, CYRUS RASHTCHIAN, AND NAVE FROST. **Explainable k-means and k-medians clustering**. In *International conference on machine learning*, pages 7055–7065. PMLR, 2020. (46)

[248] MAXIM MOZGOVOY AND EVGENY PYSHKIN. **Unity application testing automation with appium and image recognition**. In *International Conference on Tools and Methods for Program Analysis*, pages 139–150. Springer, 2017. ()

[249] ZHANG MU, CHEN YI, ZHANG XIAOHONG, AND LAI JUNYONG. **Study on the recommendation technology for tourism information service**. In *2009 Second International Symposium on Computational Intelligence and Design*, **1**, pages 410–415. IEEE, 2009. ()

[250] YAZAN MUALLA, AMRO NAJJAR, TIMOTHEUS KAMPIK, IGOR TCHAPPI, STÉPHANE GALLAND, AND CHRISTOPHE NICOLLE. **Towards explainability for a civilian uav fleet management using an agent-based approach**. *arXiv preprint arXiv:1909.10090*, 2019. ()

[251] DOST MUHAMMAD AND MALIKA BENDECHACHE. **Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis**. *Computational and structural biotechnology journal*, **24**:542–560, 2024. (22)

[252] CPC MUNAISECHE, DR KAPARANG, AND PARABELEM TINNO DOLF ROMPAS. **An Expert system for diagnosing eye diseases using forward chaining method**. In *IOP Conference Series: Materials Science and Engineering*, **306**, page 012023. IOP Publishing, 2018. ()

[253] MARK A MUSEN, BLACKFORD MIDDLETON, AND ROBERT A GREENES. **Clinical decision-support systems**. In *Biomedical informatics*, pages 795–840. Springer, 2021. (116)

[254] ANAM MUSTAQEEM, SYED MUHAMMAD ANWAR, AND MUHAMMAD MAJID. **A modular cluster based collaborative recommender system for cardiac patients**. *Artificial intelligence in medicine*, **102**:101761, 2020. ()

[255] CATALDO MUSTO, GIOVANNI SEMERARO, PASQUALE LOPS, MARCO DE GEMMIS, AND GEORGIOS LEKKAS. **Personalized finance advisory through case-based recommender systems and diversification strategies**. *Decision Support Systems*, **77**:100–111, 2015. ()

[256] VIRAL NAGORI AND BHUSHAN TRIVEDI. **Types of expert system: Comparative study**. *Asian Journal of Computer and Information Systems*, **2**(2), 2014. (118)

[257] D.P.H. NAPOLITANO, LORENZO VAIANI, AND LUCA CAGLIERO. **Learning Confidence Intervals for Feature Importance: A Fast Shapley-based Approach**. In *EDBT/ICDT Workshops*, pages 2259–2284, 2023. (88)

[258] SARANG NARKHEDE. **Understanding auc-roc curve**. *Towards data science*, **26**(1):220–227, 2018. (86, 88)

[259] LJUBICA NEDOVIĆ AND VLADAN DEVEDŽIĆ. **Expert systems in finance—a cross-section of the field**. *Expert Systems with Applications*, **23**(1):49–66, 2002. (116)

[260] NITIN NEWALIYA, VIKAS SIWACH, HARKESH SEHRAWAT, AND YUDHVIR SINGH. **Exploring Maritime Movement Information: An Explainable AI Approach using Hi-DBSCAN and SHAP**. *International Journal of Systematic Innovation*, **8**(4):131–145, 2024. (48)

[261] HUNG TRUONG THANH NGUYEN, HUNG QUOC CAO, KHANG VO THANH NGUYEN, AND NGUYEN DINH KHOI PHAM. **Evaluation of explainable artificial intelligence: Shap, lime, and cam**. In *Proceedings of the FPT AI Conference*, pages 1–6, 2021. ()

[262] JEAN JACQUES OHANA, STEVE OHANA, ERIC BENHAMOU, DAVID SALTIEL, AND BEATRICE GUEZ. **Explainable AI (XAI) models applied to the multi-agent environment of financial markets**. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 189–207. Springer, 2021. (134)

[263] GENJI OHARA, KEIGO KIMURA, AND MINEICHI KUDO. **R-LIME: Rectangular Constraints and Optimization for Local Interpretable Model-agnostic Explanation Methods**. In *International Conference on Pattern Recognition*, pages 80–95. Springer, 2025. (88)

[264] B. OJEDA-MAGANA, R. RUELAS, M.A. CORONA-NAKAMURA, AND D. ANDINA. **An Improvement to the Possibilistic Fuzzy c-Means Clustering Algorithm**. In *2006 World Automation Congress*, pages 1–8, 2006. (81)

[265] ABDULLAH CAGLAR OKSUZ, ANISA HALIMI, AND ERMAN AYDAY. **Autolycus: Exploiting explainable artificial intelligence (xai) for model extraction attacks against interpretable models**. *Proceedings on Privacy Enhancing Technologies*, 2024. (25)

[266] MERYEM OUAHILAL, MOHAMMED EL MOHAJIR, MOHAMED CHAHHOU, AND BADR EDDINE EL MOHAJIR. **A comparative study of predictive algorithms for business analytics and decision support systems: Finance as a case study**. In *2016 International Conference on Information Technology for Organizations Development (IT4OD)*, pages 1–6. IEEE, 2016. (113)

[267] LIRON PANTANOWITZ. **Introduction to informatics**. In *Practical Informatics for Cytopathology*, pages 1–4. Springer, 2013. ()

[268] REZA MEIMANDI PARIZI. **On the gamification of human-centric traceability tasks in software testing and coding**. In *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 193–200. IEEE, 2016. ()

[269] MOON-HEE PARK, JIN-HYUK HONG, AND SUNG-BAE CHO. **Location-based recommendation system using bayesian user's preference model in mobile devices**. In *International conference on ubiquitous intelligence and computing*, pages 1130–1139. Springer, 2007. ()

[270] MOONYOUNG PARK, MARINA PURGINA, AND MAXIM MOZGOVOY. **Learning English grammar with WordBricks: classroom experience**. In *Proceedings of the 2016 IEEE International Conference on Teaching and Learning in Education*, pages 220–223, 2016. ()

[271] JONATHON K PARKER, LAWRENCE O HALL, AND JAMES C BEZDEK. **Comparison of scalable fuzzy clustering methods**. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–9. IEEE, 2012. (48)

[272] P. PATEL ET AL. **Machine learning for human resource management: A review and directions**. *International Journal of Advanced Research in Computer Science*, **11**(5):10–16, 2020. (49)

[273] DAN PATTERSON. *Introduction to artificial intelligence and expert systems*. Prentice-Hall, Inc., 1990. (xi, 117, 118)

[274] URJA PAWAR, DONNA O'SHEA, SUSAN REA, AND RUAIRI O'REILLY. **Explainable ai in healthcare**. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–2. IEEE, 2020. ()

[275] MICHAEL J PAZZANI. **A framework for collaborative, content-based and demographic filtering**. *Artificial intelligence review*, **13**(5):393–408, 1999. (124)

[276] OSCAR PEDREIRA, FÉLIX GARCÍA, NIEVES BRISABOA, AND MARIO PIATTINI. **Gamification in software engineering–A systematic mapping**. *Information and software technology*, **57**:157–168, 2015. ()

[277] MIRJANA PEJIĆ BACH, ARIAN IVEC, AND DANIJELA HRMAN. **Industrial Informatics: Emerging Trends and Applications in the Era of Big Data and AI**. *Electronics*, **12**(10):2238, 2023. ()

[278] DULCE G PEREIRA, ANABELA AFONSO, AND FÁTIMA MELO MEDEIROS. **Overview of Friedman's test and post-hoc analysis**. *Communications in Statistics-Simulation and Computation*, **44**(10):2636–2653, 2015. (73)

[279] NYMPHIA PEREIRA AND SATISHKUMAR L VARMA. **Financial planning recommendation system using content-based collaborative and demographic filtering**. In *Smart Innovations in Communication and Computational Sciences*, pages 141–151. Springer, 2019. (134)

[280] GABRIJELA PERKOVIĆ, ANTUN DROBNJAK, AND IVICA BOTIČKI. **Hallucinations in LLMs: Understanding and Addressing Challenges**. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 2084–2088, 2024. (43)

[281] SLOBODAN PETROVIC. **A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters**. In *Proceedings of the 11th Nordic workshop of secure IT systems*, **2006**, pages 53–64. Citeseer, 2006. (63)

[282] DUC T PHAM AND ASHRAF A AFIFY. **Machine-learning techniques and their applications in manufacturing**. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, **219**(5):395–412, 2005. ()

[283] IVENS PORTUGAL, PAULO ALENCAR, AND DONALD COWAN. **The use of machine learning algorithms in recommender systems: A systematic review**. *Expert Systems with Applications*, **97**:205–227, 2018. (125)

[284] DANIEL J POWER. *Decision support systems: concepts and resources for managers*. Greenwood Publishing Group, 2002. (113)

[285] DJ POWER. **What is a DSS**. *The Online executive journal for data-intensive decision support*, **1**(3):223–232, 1997. (113)

[286] CP PRAMOD AND GOPINATHA NATH PILLAI. **K-Means clustering based Extreme Learning ANFIS with improved interpretability for regression problems**. *Knowledge-Based Systems*, **215**:106750, 2021. (48)

[287] P PRIYANGA AND AR NADIRA BANU KAMAL. **Methods of Mining the Data from Big Data and Social Networks Based on Recommender System**. *International Journal of Advanced Networking & Applications (IJANA)*, **8**(5):55–60, 2017. ()

[288] ANDRIA PROCOPIOU AND THOMAS M CHEN. **Explainable AI in Machine/Deep Learning for Intrusion Detection in Intelligent Transportation Systems for Smart Cities**. In *Explainable Artificial Intelligence for Smart Cities*, pages 297–321. CRC Press, 2021. (134)

[289] EVGENY PYSHKIN. **Designing human-centric applications: Transdisciplinary connections with examples**. In *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, pages 1–6. IEEE, 2017. ()

[290] EVGENY PYSHKIN AND MIKHAIL PONOMAREV. **Mathematical equation structural syntactical similarity patterns: A tree overlapping algorithm and its evaluation**. *Informatica*, **40**(4), 2016. ()

[291] SHAHZAD QAISER AND RAMSHA ALI. **Text mining: use of TF-IDF to examine the relevance of words to documents**. *International Journal of Computer Applications*, **181**(1):25–29, 2018. ()

[292] REXHEP RADA, ERIND BEDALLI, SOKOL SHURDHI, AND BETIM ÇIÇO. **A comparative analysis on prototype-based clustering methods**. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–5. IEEE, 2023. (46)

[293] MANISH RAGHAVAN, SOLON BAROCAS, JON KLEINBERG, AND KAREN LEVY. **Mitigating bias in algorithmic hiring: Evaluating claims and practices**. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020. (83)

[294] MOHAMMED MAHMUDUR RAHMAN, ZULKIFLY BIN MOHD ZAKI, NAJWA HAYAATI BINTI MOHD ALWI, AND MD MONIRUL ISLAM. **A hybrid approach to improve recommendation system in E-tourism**. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 1*, pages 787–797. Springer, 2019. (xi, 122)

[295] ARUN RAI. **Explainable AI: From black box to glass box**. *Journal of the Academy of Marketing Science*, **48**(1):137–141, 2020. (128)

[296] K RAJALAKSHMI, S CHANDRA MOHAN, AND S DHINESH BABU. **Decision support system in healthcare industry**. *International Journal of computer applications*, **26**(9):42–44, 2011. ()

[297] K RAMESH, MN INDRAJITH, YS PRASANNA, SANDIP S DESHMUKH, CHANDU PARIMI, AND TATHAGATA RAY. **Comparison and assessment of machine learning approaches in manufacturing applications**. *Industrial Artificial Intelligence*, **3**(1):2, 2025. (1)

[298] JUAN RAMOS ET AL. **Using tf-idf to determine word relevance in document queries**. In *Proceedings of the first instructional conference on machine learning*, **242**, pages 29–48. Citeseer, 2003. (45)

[299] NATHANAËL RANDRIAMIHAMISON, NATHALIE VIALANEIX, AND PIERRE NEUVIAL. **Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints**. *Journal of Classification*, **38**(2):363–389, 2021. (48)

[300] P MERCY NESA RANI, T RAJESH, AND R SARAVANAN. **Expert systems in agriculture: A review**. *Journal of Computer Science and Applications*, **3**(1):59–71, 2011. ()

[301] M. RAVI ET AL. **Hybrid Thresholding for Enhanced Performance and Interpretability in Fraud Detection: Integrating LIME and SHAP for Trustworthy AI Based Decision Making**. *Computers, Materials & Continua*, 2025. Accepted. (4, 8, 25, 85)

[302] MUDAVATH RAVI AND ATUL NEGI. **Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations**. In CHALONG SOMBATTHEERA, PENG WENG, AND JUN PANG, editors, *Multidisciplinary Trends in Artificial Intelligence. MIWAI 2024. Lecture Notes in Computer Science*, **15431**. Springer, Singapore, 2025. (4, 8, 26, 29)

[303] MUDAVATH RAVI AND ATUL NEGI. **Evolution of AI-Driven Decision Making With Decision Support Systems, Expert Systems, Recommender Systems, and XAI**. *IETE Technical Review*, 2025. Published. (5, 8, 14)

[304] MUDAVATH RAVI AND ATUL NEGI. **Leveraging LIME Explainability and Gustafson-Kessel Fuzzy Clustering for Resume Grouping and Semantic Text Summarization**. *Knowledge-Based Systems*, 2025. Under Review. (4, 8)

[305] MUDAVATH RAVI AND ATUL NEGI. **A Multi-tiered Solution for Personalized Baggage Item Recommendations using FastText and Association Rule Mining**, 2025. ()

[306] MUDAVATH RAVI, ATUL NEGI, AND SANJAY CHITNIS. **A comparative review of expert systems, recommender systems, and explainable AI**. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–8. IEEE, 2022. (xi, 4, 5, 31, 87, 117, 118, 132)

[307] N REIMERS. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. *arXiv preprint arXiv:1908.10084*, 2019. (26, 45, 53, 64)

[308] YAN REN, XIAODONG LIU, AND WANQUAN LIU. **DBCAMM: A novel density based clustering algorithm via using the Mahalanobis metric**. *Applied soft computing*, **12**(5):1542–1554, 2012. ()

[309] PAUL RESNICK AND HAL R VARIAN. **Recommender systems**. *Communications of the ACM*, **40**(3):56–58, 1997. (121, 127)

[310] MARCO TULIO RIBEIRO, SAMEER SINGH, AND CARLOS GUESTRIN. **" Why should i trust you?" Explaining the predictions of any classifier**. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. (2, 4, 7, 13, 15, 16, 43, 66, 86, 90, 94, 95, 142)

[311] MARCO TULIO RIBEIRO, SAMEER SINGH, AND CARLOS GUESTRIN. **Model-agnostic interpretability of machine learning**. *arXiv preprint arXiv:1606.05386*, 2016. (31, 87, 88, 90)

[312] FRANCESCO RICCI, LIOR ROKACH, BRACHA SHAPIRA, AND PAUL B. KANTOR. **Introduction to Recommender Systems Handbook**. In FRANCESCO RICCI, LIOR ROKACH, BRACHA SHAPIRA, AND PAUL B. KANTOR, editors, *Recommender Systems Handbook*, pages 1–35. Springer, 2011. (12)

[313] EVE RICHARDSON, RAPHAEL TREVIZANI, JASON A GREENBAUM, HANNAH CARTER, MORTEN NIELSEN, AND BJOERN PETERS. **The receiver operating characteristic curve accurately assesses imbalanced datasets**. *Patterns*, 2024. (88)

[314] CARLOTTA RIGOTTI AND EDUARD FOSCH-VILLARONGA. **Fairness, AI & Recruitment**. *Computer Law & Security Review*, **53**:105966, 2024. (83)

[315] STEPHEN E ROBERTSON AND STEVE WALKER. **Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval**. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer, 1994. (x, 52, 56)

[316] LIOR ROKACH AND ODED MAIMON. **Clustering methods**. *Data mining and knowledge discovery handbook*, pages 321–352, 2005. ()

[317] PETER J ROUSSEEUW. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis**. *Journal of Computational and Applied Mathematics*, **20**:53–65, 1987. (xiii, 47, 48)

[318] PRADEEP KUMAR ROY, SARABJEET SINGH CHOWDHARY, AND ROCKY BHATIA. **A Machine Learning approach for automation of Resume Recommendation system**. *Procedia Computer Science*, **167**:2318–2327, 2020. (30)

[319] CYNTHIA RUDIN. **Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead**. *Nature Machine Intelligence*, **1**(5):206–215, 2019. (15, 24)

[320] ENRIQUE H RUSPINI. **A new approach to clustering**. *Information and control*, **15**(1):22–32, 1969. ()

[321] WADDAH SAEED AND CHRISTIAN OMLIN. **Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities**. *Knowledge-Based Systems*, **263**:110273, 2023. ()

[322] MICHAEL J SAILOR AND JAMIE R LINK. **"Smart dust": nanostructured devices in a grain of sand**. *Chemical Communications*, **11**:1375–1383, 2005. ()

[323] TAKAYA SAITO AND MARC REHMSMEIER. **The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets**. *PloS one*, **10**(3):e0118432, 2015. (88)

[324] TATSUYA SAKAI AND TAKAYUKI NAGAI. **Explainable autonomous robots: a survey and perspective**. *Advanced Robotics*, **36**(5-6):219–238, 2022. (23)

[325] AHMED M SALIH, ZAHRA RAISI-ESTABRAGH, ILARIA BOSCOLO GALAZZO, PETIA RADEVA, STEFFEN E PETERSEN, KARIM LEKADIR, AND GLORIA MENEGAZ. **A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME**. *Advanced Intelligent Systems*, page 2400304, 2024. ()

[326] WOJCIECH SAMEK, GRÉGOIRE MONTAVON, ANDREA VEDALDI, LARS KAI HANSEN, AND KLAUS-ROBERT MÜLLER. *Explainable AI: interpreting, explaining and visualizing deep learning*, **11700**. Springer Nature, 2019. ()

[327] AMINA SAMIH, AMINA ADADI, AND MOHAMMED BERRADA. **Towards a knowledge-based explainable recommender system**. In *Proceedings of the 4th International Conference on Big Data and Internet of Things*, pages 1–5, 2019. (125, 133)

[328] VICTOR SANH, L DEBUT, J CHAUMOND, AND T WOLF. **DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019**. *arXiv preprint arXiv:1910.01108*, 2019. (53, 58)

[329] AMIT SAXENA, MUKESH PRASAD, AKSHANSH GUPTA, NEHA BHARILL, OM PRAKASH PATEL, ARUNA TIWARI, MENG JOO ER, WEIPING DING, AND CHIN-TENG LIN. **A review of clustering techniques and developments**. *Neurocomputing*, **267**:664–681, 2017. (45)

[330] ERICH SCHUBERT, JÖRG SANDER, MARTIN ESTER, HANS-PETER KRIEGEL, AND XIAOWEI XU. **DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN**. *ACM Transactions on Database Systems (TODS)*, **42**(3):1–21, 2017. (48)

[331] JOSEPH SEERING, MICHAL LURIA, CONNIE YE, GEOFF KAUFMAN, AND JESSICA HAMMER. **It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community**. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020. ()

[332] NIMA GERAMI SERESHT, RODOLFO LOURENZUTTI, AND AMINAH ROBINSON FAYEK. **A fuzzy clustering algorithm for developing predictive models in construction applications**. *Applied Soft Computing*, **96**:106679, 2020. ()

[333] LISA SERIR, EMMANUEL RAMASSO, AND NOUREDDINE ZERHOUNI. **Evidential evolving Gustafson–Kessel algorithm for online data streams partitioning using belief function theory**. *International journal of approximate reasoning*, **53**(5):747–768, 2012. ()

[334] EMILY SETO, KEVIN J LEONARD, JOSEPH A CAFAZZO, JAN BARNSLEY, CATERINA MASINO, AND HEATHER J ROSS. **Developing healthcare rule-based expert systems: case study of a heart failure telemonitoring system**. *International journal of medical informatics*, **81**(8):556–565, 2012. (116)

[335] KETAN RAJSHEKHAR SHAHAPURE AND CHARLES NICHOLAS. **Cluster quality analysis using silhouette score**. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020. (63)

[336] LLOYD S SHAPLEY ET AL. **A value for n-person games**. *Theory of Games*, 1953. (88)

[337] LALITA SHARMA AND ANJU GERA. **A survey of recommendation system: Research challenges**. *International Journal of Engineering Trends and Technology (IJETT)*, **4**(5):1989–1992, 2013. (123)

[338] EDWARD H. SHORTLIFFE. *Computer-Based Medical Consultations: MYCIN*. Elsevier, 1976. (14)

[339] MESHAL SHUTAYWI AND NEZAMODDIN N KACHOUIE. **Silhouette analysis for performance evaluation in machine learning with applications to clustering**. *Entropy*, **23**(6):759, 2021. (63)

[340] SVETLANA SICULAR, KENNETH BRANT, AND GARTNER I GARTNER. **Hype cycle for artificial intelligence, 2018**. *Edited by Gartner I. Gartner*, pages 1–73, 2018. (130)

[341] BHAGYA NATHALI SILVA, MURAD KHAN, AND KIJUN HAN. **Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities**. *Sustainable Cities and Society*, **38**:697–713, 2018. ()

[342] LILIANE SILVA, RONILDO MOURA, ANNE MP CANUTO, REGIVAN HN SANTIAGO, AND BENJAMIN BEDREGAL. **An interval-based framework for fuzzy clustering applications**. *IEEE Transactions on Fuzzy Systems*, **23**(6):2174–2187, 2015. ()

[343] AUSTE SIMKUTE, EWA LUGER, BRONWYN JONES, MICHAEL EVANS, AND RHIANNE JONES. **Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable**. *Journal of Responsible Technology*, **7**:100017, 2021. (137)

[344] DYLAN SLACK, ANNA HILGARD, SAMEER SINGH, AND HIMABINDU LAKKARAJU. **Reliable post hoc explanations: Modeling uncertainty in explainability**. *Advances in neural information processing systems*, **34**:9391–9404, 2021. (15)

[345] DYLAN SLACK, SOPHIE HILGARD, EMILY JIA, SAMEER SINGH, AND HIMABINDU LAKKARAJU. **Fooling lime and shap: Adversarial attacks on post hoc explanation methods**. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. (15, 88)

[346] BRENT SMITH AND GREG LINDEN. **Two Decades of Recommender Systems at Amazon.com**. *IEEE Internet Computing*, **21**(3):12–18, 2017. ()

[347] ALISON SMITH-RENNER, RON FAN, MELISSA BIRCHFIELD, TONGSHUANG WU, JORDAN BOYD-GRABER, DANIEL S WELD, AND LEAH FINDLATER. **No explainability without accountability: An empirical study of explanations and feedback in interactive ml**. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13, 2020. ()

[348] LE HOANG SON. **A novel kernel fuzzy clustering algorithm for geo-demographic analysis**. *Information Sciences—Informatics and Computer Science, Intelligent Systems, Applications: An International Journal*, **317**(C):202–223, 2015. ()

[349] LE HOANG SON. **Generalized picture distance measure and applications to picture fuzzy clustering**. *Applied Soft Computing*, **46**(C):284–295, 2016. ()

[350] XINYUAN SONG, TIANYANG WANG, MING LIU, YUNZE WANG, BENJI PENG, SILIN CHEN, QIAN NIU, JUNYU LIU, KEYU CHEN, MING LI, POHSUN FENG, ZIQIAN BI, YICHAO ZHANG, CHENG FEI, CAITLYN YIN, AND KAI QI YAN. **Explainable**

AI Across Domains: Techniques, Domain-Specific Applications, and Future Directions. *OSF Preprint*, 12 2024. ()

[351] CHRISTOFOROS N SPARTALIS, THEODOROS SEMERTZIDIS, AND PETROS DARAS. **Balancing xai with privacy and security considerations**. In *European Symposium on Research in Computer Security*, pages 111–124. Springer, 2023. (25)

[352] RALPH H SPRAGUE JR AND ERIC D CARLSON. *Building effective decision support systems*. Prentice Hall Professional Technical Reference, 1982. (113)

[353] BHARGAV SRINIVASA-DESIKAN. *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd, 2018. ()

[354] MATEUSZ STANIAK AND PRZEMYSLAW BIECEK. **Explanations of model predictions with live and breakDown packages**. *arXiv preprint arXiv:1804.01955*, 2018. (142)

[355] MARK STEFIK, MICHAEL YOUNGBLOOD, PETER PIROLLI, CHRISTIAN LEBIERE, ROBERT THOMSON, ROBERT PRICE, LESTER D NELSON, ROBERT KRIVACIC, JACOB LE, KONSTANTINOS MITSOPOULOS, ET AL. **Explaining autonomous drones: An XAI journey**. *Applied AI Letters*, **2**(4):e54, 2021. (23)

[356] PHILIP J STONE. **Thematic text analysis: New agendas for analyzing text content**. *Text analysis for the social sciences*, pages 35–54, 2020. ()

[357] ERIK ŠTRUMBELJ AND IGOR KONONENKO. **Explaining prediction models and individual predictions with feature contributions**. *Knowledge and information systems*, **41**:647–665, 2014. (95)

[358] MASASHI SUGIYAMA. *Introduction to statistical machine learning*. Morgan Kaufmann, 2015. (15)

[359] XIAOQIAN SUN AND SEBASTIAN WANDELT. **Transportation mode choice behavior with recommender systems: A case study on Beijing**. *Transportation research interdisciplinary perspectives*, **11**:100408, 2021. (134)

[360] REED T SUTTON, DAVID PINCOCK, DANIEL C BAUMGART, DANIEL C SADOWSKI, RICHARD N FEDORAK, AND KAREN I KROEKER. **An overview of clinical decision support systems: benefits, risks, and strategies for success**. *NPJ digital medicine*, **3**(1):1–10, 2020. ()

[361] CHAKKRIT TANTITHAMTHAVORN, JÜRGEN CITO, HADI HEMMATI, AND SATISH CHANDRA. **Explainable ai for se: Challenges and future directions**. *IEEE Software*, **40**(3):29–33, 2023. (148)

[362] SEAN TAO. **Deep neural network ensembles**. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 1–12. Springer, 2019. ()

[363] TIMUR TASHMETOV, KAMOLIDDIN TASHMETOV, RAVSHAN ALIEV, AND MUHAMADAZIZ RASULMUHAMEDOV. **Fuzzy information and expert systems for analysis of failure of automatic and telemechanic systems on railway transport**. *Chemical Technology, Control and Management*, **2020**(5):168–172, 2020. ()

[364] LUKA TESLIC, BENJAMIN HARTMANN, OLIVER NELLES, AND IGOR SKRJANC. **Nonlinear system identification by Gustafson–Kessel fuzzy clustering and supervised local model network learning for the drug absorption spectra process**. *IEEE Transactions on Neural Networks*, **22**(12):1941–1951, 2011. ()

[365] LIANG TIANGANG, CHEN QUANGONG, REN JIZHOU, AND WANG YUANSU. **A GIS-based expert system for pastoral agricultural development in Gansu Province, PR China**. *New Zealand Journal of Agricultural Research*, **47**(3):313–325, 2004. ()

[366] ILHAM ESA TIFFANI. **Optimization of naïve bayes classifier by implemented unigram, bigram, trigram for sentiment analysis of hotel review**. *Journal of Soft Computing Exploration*, **1**(1):1–7, 2020. ()

[367] RICHARD TOMSETT, DAVE BRAINES, DAN HARBORNE, ALUN PREECE, AND SUPRIYO CHAKRABORTY. **Interpretable to whom? A role-based model for analyzing interpretable machine learning systems**. *arXiv preprint arXiv:1806.07552*, 2018. (138)

[368] THI NGOC TRANG TRAN, ALEXANDER FELFERNIG, CHRISTOPH TRATTNER, AND ANDREAS HOLZINGER. **Recommender systems in the healthcare domain: state-of-the-art and research issues**. *Journal of Intelligent Information Systems*, **57**(1):171–201, 2021. (134)

[369] EFRAIM TURBAN. *Decision support and expert systems: Managerial perspectives*. Macmillan Library Reference, 1990. ()

[370] EFRAIM TURBAN. *Decision support and expert systems*. Pearson Education India, 2005. (12)

[371] EFRAIM TURBAN AND PAUL R WATKINS. **Integrating expert systems and decision support systems**. *Mis Quarterly*, pages 121–136, 1986. (xi, 119)

[372] NIELS VAN BERKEL, BENJAMIN TAG, JORGE GONCALVES, AND SIMO HOSIO. **Human-centred artificial intelligence: a contextual morality perspective**. *Behaviour & Information Technology*, **41**(3):502–518, 2022. (30)

[373] GERRIT H VAN BRUGGEN, ALE SMIDTS, AND BEREND WIERENGA. **Improving decision making by means of a marketing decision support system**. *Management Science*, **44**(5):645–658, 1998. (113)

[374] MARTIN VAN DEN BERG AND OUREN KUIPER. **XAI in the financial sector: a conceptual framework for explainable AI (XAI)**. *https://www. hu. nl/-/media/hu/documenten/onderzoek/projecten/*, 2020. (134)

[375] BLAKE VANBERLO, MATTHEW AS ROSS, JONATHAN RIVARD, AND RYAN BOOKER. **Interpretable machine learning approaches to prediction of chronic homelessness**. *Engineering Applications of Artificial Intelligence*, **102**:104243, 2021. (137)

[376] SANDRA WACHTER, BRENT MITTELSTADT, AND LUCIANO FLORIDI. **Why a right to explanation of automated decision-making does not exist in the general data protection regulation**. *International Data Privacy Law*, **7**(2):76–99, 2017. (21)

[377] PAUL WALTON. **Information and Inference**. *Information*, **8**(2), 2017. ()

[378] HUANJING WANG, QIANXIN LIANG, JOHN T HANCOCK, AND TAGHI M KHOSHGOFTAAR. **Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods**. *Journal of Big Data*, **11**(1):44, 2024. (88)

[379] SHEN WANG, M ATIF QURESHI, LUIS MIRALLES-PECHUAN, THIEN HUYNH-THE, THIPPA REDDY GADEKALLU, AND MADHUSANKA LIYANAGE. **Applications of explainable AI for 6G: Technical aspects, use cases, and research challenges**. *arXiv preprint arXiv:2112.04698*, 2021. ()

[380] SHOUJIN WANG, LONGBING CAO, YAN WANG, QUAN Z SHENG, MEHMET A ORGUN, AND DEFU LIAN. **A survey on session-based recommender systems**. *ACM Computing Surveys (CSUR)*, **54**(7):1–38, 2021. (124)

[381] XIAOHONG WANG. **Bioartificial organ manufacturing technologies**. *Cell transplantation*, **28**(1):5–17, 2019. ()

[382] XU WANG AND YUSHENG XU. **An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index**. In *IOP Conference Series: Materials Science and Engineering*, **569**, page 052024. IOP Publishing, 2019. (63)

[383] ZIJIE J WANG, ROBERT TURKO, OMAR SHAIKH, HAEKYU PARK, NILAKSH DAS, FRED HOHMAN, MINSUK KAHNG, AND DUEN HORNG POLO CHAU. **CNN explainer: learning convolutional neural networks with interactive visualization**. *IEEE Transactions on Visualization and Computer Graphics*, **27**(2):1396–1406, 2020. (16)

[384] AIDEN WARREN AND ALEK HILLAS. **Friend or frenemy? The role of trust in human-machine teaming and lethal autonomous weapons systems**. *Small Wars & Insurgencies*, **31**(4):822–850, 2020. (134)

[385] LEANDER WEBER, SEBASTIAN LAPUSCHKIN, ALEXANDER BINDER, AND WOJCIECH SAMEK. **Beyond explaining: Opportunities and challenges of XAI-based model improvement**. *Information Fusion*, **92**:154–176, 2023. (2, 4, 5, 88)

[386] JOHN WECKERT AND ANDREW MAIN. **In Defence of Simple Expert Systems: a case study with some observations**. *LASIE: Library Automated Systems Information Exchange*, **23**(4/5):62–70, 1993. ()

[387] CHIH-HSIU WEI AND CHIN-SHYURNG FAHN. **The multisynapse neural network and its application to fuzzy clustering**. *IEEE transactions on neural networks*, **13**(3):600–618, 2002. ()

[388] KATHARINA WEITZ, DOMINIK SCHILLER, RUBEN SCHLAGOWSKI, TOBIAS HUBER, AND ELISABETH ANDRÉ. **"Let me explain!": exploring the potential of virtual agents in explainable AI interaction design**. *Journal on Multimodal User Interfaces*, **15**(2):87–98, 2021. ()

[389] N PETER WHITEHEAD, WILLIAM T SCHERER, AND MICHAEL C SMITH. **Systems thinking about systems thinking a proposal for a common language**. *IEEE Systems Journal*, **9**(4):1117–1128, 2014. ()

[390] ADHIKA PRAMITA WIDYASSARI, SUPRIADI RUSTAD, GURUH FAJAR SHIDIK, EDI NOERSASONGKO, ABDUL SYUKUR, AFFANDY AFFANDY, ET AL. **Review of automatic text summarization techniques & methods**. *Journal of King Saud University-Computer and Information Sciences*, **34**(4):1029–1046, 2022. ()

[391] B.W. WIRTZ ET AL. **Artificial intelligence and the future of work: Human–AI interaction in HR management**. *Journal of Business Research*, **100**:366–374, 2018. (49)

[392] CAROLYN YU TUNG WONG, FARES ANTAKI, PETER WOODWARD-COURT, ARIEL YUHAN ONG, AND PEARSE A KEANE. **The role of saliency maps in enhancing ophthalmologists' trust in artificial intelligence models**. *Asia-Pacific Journal of Ophthalmology*, **13**(4):100087, 2024. (22)

[393] TUNG-YU WU AND YOU-TING WANG. **Locally interpretable one-class anomaly detection for credit card fraud detection**. In *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 25–30. IEEE, 2021. (88)

[394] THORSTEN WUEST, DANIEL WEIMER, CHRISTOPHER IRGENS, AND KLAUS-DIETER THOBEN. **Machine learning in manufacturing: advantages, challenges, and applications**. *Production & Manufacturing Research*, **4**(1):23–45, 2016. ()

[395] JUNWEI XIAO, JIANFENG LU, AND XIANGYU LI. **Davies Bouldin Index based hierarchical initialization K-means**. *Intelligent Data Analysis*, **21**(6):1327–1338, 2017. (63)

[396] FEIYU XU, HANS USZKOREIT, YANGZHOU DU, WEI FAN, DONGYAN ZHAO, AND JUN ZHU. **Explainable AI: A brief survey on history, research areas, approaches and challenges**. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019. ()

[397] RUI XU AND DON WUNSCH. *Clustering*. John Wiley & Sons, 2008. ()

[398] RUI XU AND DONALD WUNSCH. **Survey of clustering algorithms**. *IEEE Transactions on neural networks*, **16**(3):645–678, 2005. (xiii, 45, 46, 47, 48)

[399] KIRAN KUMAR REDDY YANAMALA. **Dynamic bias mitigation for multimodal AI in recruitment ensuring fairness and equity in hiring practices**. *Journal of Artificial Intelligence and Machine Learning in Management*, **6**(2):51–61, 2022. (83)

[400] M-S YANG. **A survey of fuzzy clustering**. *Mathematical and Computer modelling*, **18**(11):1–16, 1993. ()

[401] MIIN-SHEN YANG AND CHIEN-YO LAI. **A robust automatic merging possibilistic clustering method**. *IEEE Transactions on Fuzzy Systems*, **19**(1):26–41, 2010. ()

[402] MIIN-SHEN YANG, KUO-LUNG WU, JUNE-NAN HSIEH, AND JIAN YU. **Alpha-cut implemented fuzzy clustering algorithms and switching regressions**. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **38**(3):588–603, 2008. ()

[403] WENLI YANG, YUCHEN WEI, HANYU WEI, YANYU CHEN, GUAN HUANG, XIANG LI, RENJIE LI, NAIMENG YAO, XINYI WANG, XIAOTONG GU, ET AL. **Survey on explainable AI: From approaches, limitations and applications aspects**. *Human-Centric Intelligent Systems*, **3**(3):161–188, 2023. (2)

[404] ZEBIN YANG, AGUS SUDJIANTO, XIAOMING LI, AND AIJUN ZHANG. **Inherently Interpretable Tree Ensemble Learning**. *arXiv preprint arXiv:2410.19098*, 2024. (15)

[405] CHIOU-CHERNG YEH AND MIIN-SHEN YANG. **Evaluation measures for cluster ensembles based on a fuzzy generalized Rand index**. *Applied Soft Computing*, **57**:225–234, 2017. (62)

[406] MICHAEL YEOMANS, ANUJ SHAH, SENDHIL MULLAINATHAN, AND JON KLEINBERG. **Making sense of recommendations**. *Journal of Behavioral Decision Making*, **32**(4):403–414, 2019. (137, 138)

[407] KA YEE YEUNG AND WALTER L RUZZO. **Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data**. *Bioinformatics*, **17**(9):763–774, 2001. (62)

[408] SANG WON YOON, JUAN DIEGO VELÁSQUEZ, BK PARTRIDGE, AND SHIMON Y NOF. **Transportation security decision support system for emergency response: A training prototype**. *Decision Support Systems*, **46**(1):139–148, 2008. (116)

[409] HAIYAN YU, LERONG JIANG, JIULUN FAN, AND RONG LAN. **Double-suppressed possibilistic fuzzy Gustafson–Kessel clustering algorithm**. *Knowledge-Based Systems*, **276**:110736, 2023. ()

[410] JIAN YU, MIIN-SHEN YANG, ET AL. **Analysis of parameter selection for Gustafson–Kessel fuzzy clustering using Jacobian matrix**. *IEEE Transactions on Fuzzy Systems*, **23**(6):2329–2342, 2015. ()

[411] JIAN YU, MIIN-SHEN YANG, ET AL. **Deterministic annealing Gustafson-Kessel fuzzy clustering algorithm**. *Information Sciences*, **417**:435–453, 2017. ()

[412] KUN YU, GANG GUAN, AND MING ZHOU. **Resume information extraction with cascaded hybrid model**. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 499–506, 2005. (30)

[413] JAN ZACHARIAS, MORITZ VON ZAHN, JOHANNES CHEN, AND OLIVER HINZ. **Designing a feature selection method based on explainable artificial intelligence**. *Electronic Markets*, **32**(4):2159–2184, 2022. (2, 4, 5, 88)

[414] LOTFI A ZADEH. **Fuzzy sets**. *Information and control*, **8**(3):338–353, 1965. ()

[415] MARZIE ZARINBAL, MH FAZEL ZARANDI, AND IB TURKSEN. **Interval type-2 relative entropy fuzzy C-means clustering**. *Information Sciences*, **272**:49–72, 2014. ()

[416] CHAOBO ZHANG, PIETER-JAN HOES, SHUWEI WANG, AND YANG ZHAO. **Intrinsically interpretable machine learning-based building energy load prediction method with high accuracy and strong interpretability**. *Energy and Built Environment*, 2024. (18)

[417] SHAOHONG ZHANG AND HAU-SAN WONG. **ARImp: A generalized adjusted rand index for cluster ensembles**. In *2010 20th International Conference on Pattern Recognition*, pages 778–781. IEEE, 2010. ()

[418] SHAOHONG ZHANG, HAU-SAN WONG, AND YING SHEN. **Generalized adjusted rand indices for cluster ensembles**. *Pattern Recognition*, **45**(6):2214–2226, 2012. (62)

[419] YUJIA ZHANG, KUANGYAN SONG, YIMING SUN, SARAH TAN, AND MADELEINE UDELL. **" Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations**. *arXiv preprint arXiv:1904.12991*, 2019. (31)

[420] YUE ZHAO, ZAIN NASRULLAH, AND ZHENG LI. **Pyod: A python toolbox for scalable outlier detection**. *Journal of machine learning research*, **20**(96):1–7, 2019. (89)

[421] SHUISHENG ZHOU, DONG LI, ZHUAN ZHANG, AND RUI PING. **A new membership scaling fuzzy C-means clustering algorithm**. *IEEE transactions on fuzzy systems*, **29**(9):2810–2818, 2020. ()

[422] YIJIA ZHOU, KYLE A GALLIVAN, AND ADRIAN BARBU. **Scalable clustering: Large scale unsupervised learning of gaussian mixture models with outliers**. *Journal of Computational and Graphical Statistics*, pages 1–12, 2024. (48)

[423] DÁVID ZIBRICZKY12. **Recommender systems meet finance: a literature review**. In *Proc. 2nd Int. Workshop Personalization Recommender Syst*, pages 1–10, 2016. (134)

[424] SAURABH BHAUSAHEB ZINJAD, AMRITA BHATTACHARJEE, AMEY BHILEGAONKAR, AND HUAN LIU. **ResumeFlow: An LLM-facilitated Pipeline for Personalized Resume Generation and Refinement**. In *Proceedings of the 17th ACM Web Conference 2024*, New York, NY, USA, 2024. Association for Computing Machinery. (43)

[425] RADIM ŘEHŮŘEK AND PETR SOJKA. **Software Framework for Topic Modelling with Large Corpora**. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, pages 46–50, Valletta, Malta, 2010. University of Malta. (52)

2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI

Mudavath Ravi
School Of Computer And Information Sciences
University of Hyderabad
Hyderabad, Telangana
India
19mcpc07@uohyd.ac.in

Atul Negi
School Of Computer And Information Sciences
University of Hyderabad
Hyderabad, Telangana
India
atul.negi@uohyd.ac.in

Sanjay Chitnis
School of Computer Science and Engineering
REVA University
Bengaluru, Karnataka
India
sanjay.chitnis@reva.edu.in

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9824265

***Abstract*—Previously Expert Systems (ES) dominated Artificial Intelligence (AI) applications and various ES were developed in multiple domains. However, due to knowledge acquisition bottlenecks, these systems fell out of use. With the rise in Machine Learning (ML) and Deep Learning (DL) approaches, another category of systems called Recommender Systems (RS) is now developed for various application domains. As ML/DL systems acted like black boxes, explainable AI (XAI) came into the picture to provide explanations for the recommendations or predictions made. In this paper, we review the architectural similarities and differences between these three approaches along with applications and future directions. It is important to study these to predict the future of RS and any possible resurgence of ES, developments in XAI and application domains.**

***Index Terms*—Expert Systems (ES), Artificial Intelligence (AI), Recommender Systems (RS), Machine Learning (ML), Deep Learning (DL), Search Engine (SE), Rule Engine (RE), and Explainable Artificial Intelligence (XAI).**

I. Introduction

An Expert System (ES) is a piece of computer software created to solve typical problems in a particular field of expertise. ES is a subset of AI. The early ES, which were amongst the first successful practical approaches to AI, were developed in the 1970s. They were applied to tackle difficult challenges of expert decision-making by extracting knowledge from experts and storing it as both facts and heuristics in a knowledge base, in a manner similar to a human expert. ES were expected to capture expert knowledge on a certain topic and are capable of solving a specific set of problems in that domain. These systems were tailored to a particular field, such as medicine or science, accountancy, agriculture, bio-engineering, Geology, Finance, etc. These are some of the disciplines where ESs have been developed [15] [20] [19]. However, Knowledge Acquisition remains a bottleneck for ES. From the early years of the new millennium, Machine Learning (ML) methods have become more important. Recently, Deep Learning (DL) has gathered momentum and several applications are now being developed. While these three approaches look different, there are certain similarities that we will like to bring out in this paper.

This paper is organized as follows: Section-II introduces Expert Systems whereas section-III describes the architecture of the Expert Systems and explains the relationship between ES and RS. Section-IV provides the overview of Recommender Systems and Section-V provides the overview of Explainable Artificial Intelligence. Table-II describes the various applications of Recommender Systems, Expert systems, and XAI.

II. Expert Systems

An ES is computer software designed to tackle complicated problems in a given domain that are typically solved by human experts. An ES was modeled to be like a human brain, simulating the thinking of a human expert. An ES was to undertake analysis, design, monitoring, and decision-making, among other things. For example, when we access a website for searching items or products using keywords with typographical errors, ES recommends the user to correct those keywords for searching. Nowadays Google, Amazon, Flipkart, banking, and other E- commerce companies build their own systems like RS or ES for the convenience of their customers.

Expert systems were said to be quite mature and thus successful [22]. ES were expected to outperform skilled humans in medical diagnosis, predicting crop disease, and extracting minerals from ore. DENDRAL is the first ES to investigate how scientists generate hypotheses and uncover new ones [50] and it was used to determine the structure of chemical compounds. Later many ES were designed in like, MYCIN,

Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations

Mudavath Ravi and Atul Negi( )

School of Computer and Information Sciences, University of Hyderabad, Hyderabad 500046, Telangana, India
atul.negi@uohyd.ac.in

Abstract. In the realm of automated resume categorization, the pressing need for both accuracy and interpretability in machine learning models remains a challenge. This research addresses the challenges of accuracy and interpretability in automated resume categorization by integrating K-Nearest Neighbors (KNN) with Local Interpretable Model-agnostic Explanations (LIME). Utilizing a diverse dataset from various industries, the study involves preprocessing resumes with TF-IDF vectorization and training the KNN algorithm with optimized parameters. LIME provides local explanations for individual predictions, enhancing transparency in the decision-making process. The framework is evaluated on standard metrics and demonstrated with real-world resumes, showcasing its effectiveness in practical recruitment scenarios. This approach advances the literature by combining KNN's simplicity with LIME's interpretability, promoting trust and fairness in automated recruitment systems.

Keywords: Resume Categorization · Automated Recruitment · Machine Learning · K-Nearest Neighbors (KNN) · Local Interpretable Model-agnostic Explanations (LIME) · Interpretability

1 Introduction

In the digital era, where resumes are processed by algorithms and classifiers, transparency in decision-making is paramount. Our approach focuses on enhancing model interpretability using the Local Interpretable Model-agnostic Explanations (LIME) framework. Resumes are not just strings of words but reflections of individual journeys, and the opacity of algorithms can obscure decision-making when making life changing decisions about people.

LIME helps reveal the thought processes behind model decisions. By employing LIME with models like K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Logistic Regression, we can understand not

C. Sombattheera et al. (Eds.): MIWAI 2024, LNAI 15431, pp. 407–418, 2025.
https://doi.org/10.1007/978-981-96-0692-4_33

Ravi Mudavath <19mcpc07@uohyd.ac.in>

CMC 65304: Accepted for Publication

5 messages

Intelligent Journal System <admin1@tspsubmission.com> Fri, May 30, 2025 at 12:14 PM
To: RAVI Mudavath <19mcpc07@uohyd.ac.in>, Atul Negi <atul.negi@uohyd.ac.in>

CMC-Computers, Materials & Continua
ISSN:1546-2226

Dear Atul Negi,

We are pleased to inform you that your submission ID: 65304 titled "Hybrid Thresholding for Enhanced Performance and Interpretability in Fraud Detection: Integrating LIME and SHAP for Trustworthy AI Based Decision Making" to CMC-Computers, Materials & Continua has been officially accepted.

The APC of your article is: 2200 USD.

Total payment is due in 10 days.

Please log in to the system promptly to stay updated on the latest status concerning your manuscript. Within our manuscript system, you will find such details as payment, copy-editing, typesetting, and proofreading as your manuscript progresses through the subsequent stages of processing.

We appreciate your dedication to advancing scholarly knowledge and your collaboration with CMC. Once again, congratulations on this significant achievement.

Sincerely,

CMC-Computers, Materials & Continua

Home Page: https://techscience.com/journal/cmc

Paper Submission: https://ijs.tspsubmission.com/homepage

Email: cmc@techscience.com

IETE TECHNICAL REVIEW
https://doi.org/10.1080/02564602.2025.2512086

REVIEW ARTICLE

Evolution of AI-Driven Decision Making with Decision Support Systems, Expert Systems, Recommender Systems, and XAI

Mudavath Ravi[1], Atul Negi[1], Nitin Sai Bommi[2] and Nusrat Rouf[1]

[1]School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India; [2]Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA

ABSTRACT

In contemporary society, decision-making processes have grown increasingly intricate across sectors such as business, healthcare, finance, and technology. To navigate this complexity, sophisticated tools like Decision Support Systems (DSS), Expert Systems (ES), Recommender Systems, and explainable Artificial Intelligence (XAI) have emerged, all aimed at improving decision-making efficiency. DSS, developed in the 1960s and 1970s, strategically integrates automation and data processing to enhance human decision-making. Unlike ES, which attempts to replicate human expertise, DSS collaboratively assists decision-makers. ES excel in specialized domains by mimicking human decision-making processes. Recommender Systems, are user-centric, transform digital interactions by analyzing preferences to provide personalized recommendations. XAI addresses the need for transparency in AI-driven decisions by clarifying outcomes from complex algorithms. In this paper, we aim to bridge the gap between traditional decision-making methods and emerging explanation-driven architectures by conducting a comparative analysis of DSS, ES, Recommender Systems, and XAI. It explores their historical origins, objectives, methodologies, applications, challenges, limitations, and user contexts. By elucidating their strengths and limitations, this analysis offers insights for decision-makers, researchers, and practitioners aiming to improve decision-making across diverse domains. At the end we explain the case study on software defect prediction.

KEYWORDS

Decision Support Systems (DSS); Expert Systems (ES); Recommender Systems (RS); and Explainable Artificial Intelligence (XAI)

1. INTRODUCTION

In today's rapidly evolving world, decision-making has become increasingly intricate, spanning diverse domains such as business, healthcare, finance, and technology. In this landscape, the demand for informed decisions is paramount, leading to the emergence of decision support technologies as indispensable tools. Among these, Explainable AI (XAI) is notable for addressing the need for transparency and interpretability in AI-driven decisions, providing clear insights into underlying algorithms and enhancing trust. XAI, alongside other technologies like Recommender Systems, Expert Systems (ES), and Decision Support Systems (DSS), contributes significantly to decision-making effectiveness. Recommender Systems personalize interactions with digital content, ES replicate human expertise, and DSS provide interactive tools for analyzing complex data. Understanding the principles and limitations of each technology is vital for leveraging them effectively. This paper aims to comprehensively explore XAI, Recommender Systems, ES, and DSS, offering valuable insights to decision-makers, researchers, and practitioners striving to enhance decision-making across diverse domains.

1.1 Background and Context

In today's complex decision-making landscape spanning business, healthcare, finance, and technology, decision-makers encounter intricate challenges and vast datasets. To navigate this complexity, decision support technologies have emerged as indispensable aids, offering valuable insights and analytical capabilities. These tools play a pivotal role in assisting decision-makers by providing data-driven insights, facilitating scenario analysis, and optimizing resource allocation. Understanding their underlying principles and applications is crucial for harnessing their full potential. This paper aims to explore the landscape of decision support technologies, highlighting their significance in contemporary decision-making contexts and addressing the challenges faced in decision-making.

1.2 Research Objectives

Our aim is to compare and analyze Decision Support Systems (DSS), Expert Systems (ES), Recommender Systems, and Explainable AI (XAI). We seek to understand their historical origins, architectural foundations, methodologies, challenges, applications,

CERTIFICATE

OF PARTICIPATION

This certificate is awarded to

Ravi Mudavath and Atul Negi

for the paper entitled

Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with Lime Explanations

presented in The 17th Multi-Disciplinary International Conference on Artificial Intelligence. Held doing November 11-15, 2024 in Pattaya, Thailand.

MIWAI
Pattaya 2024
www.miwai.org

Chattrakul Sombattheera
Mahasarakham University, Thailand
MIWAI 2024 General Co-Chair

Paul Weng
Duke Kunshan University, China
MIWAI 2024 General Co-Chair

Springer | Lecture Notes in Artificial Intelligence | Lecture Notes in Computer Science | AIAT

Data Driven Approaches using Explainable AI for Industrial Applications

by Ravi M

Submission date: 30-Aug-2025 09:58PM (UTC+0530)
Submission ID: 2738249745
File name: Ravi_Thesis_final_plagrism_report.pdf (8.81M)
Word count: 48414
Character count: 269482

Data Driven Approaches using Explainable AI for Industrial Applications

ORIGINALITY REPORT

15%	6%	12%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Mudavath Ravi, Atul Negi. "Chapter 33 Enhancing Transparency and Fairness in Automated Resume Categorization: A KNN-Based Approach with LIME Explanations", Springer Science and Business Media LLC, 2025 Publication	6%
2	www.pwc.com Internet Source	1%
3	hexaware.com Internet Source	<1%
4	Auste Simkute, Ewa Luger, Bronwyn Jones, Michael Evans, Rhianne Jones. "Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable", Journal of Responsible Technology, 2021 Publication	<1%
5	www.mdpi.com Internet Source	<1%
6	"Explainable AI in Health Informatics", Springer Science and Business Media LLC, 2024 Publication	<1%
7	github.com Internet Source	<1%

Students own publication cited. overall #1, #54 is to be excluded 15-7=8% similarity

Atul Negi
1/9/25

Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46 (India)

8	Submitted to Brunel University Student Paper	<1%
9	Alessandro Bondielli, Francesco Marcelloni. "On the use of summarization and transformer architectures for profiling résumés", Expert Systems with Applications, 2021 Publication	<1%
10	Submitted to University of Surrey Student Paper	<1%
11	3fdef50c-add3-4615-a675-a91741bcb5c0.usrfiles.com Internet Source	<1%
12	"Advances in Data and Information Sciences", Springer Science and Business Media LLC, 2024 Publication	<1%
13	arxiv.org Internet Source	<1%
14	Submitted to University of St Andrews Student Paper	<1%
15	Pethuru Raj, B. Sundaravadivazhagan, A. Saleem Raja, Mohammed M. Alani. "Edge AI for Industry 5.0 and Healthcare 5.0 Applications", CRC Press, 2025 Publication	<1%
16	Mukta Sharma, Amit Kumar Goel, Priyank Singhal. "Chapter 7 Explainable AI Driven Applications for Patient Care and Treatment", Springer Science and Business Media LLC, 2023 Publication	<1%

17	"Advances in Knowledge Discovery and Data Mining", Springer Science and Business Media LLC, 2006 Publication	<1 %
18	Submitted to Sheffield Hallam University Student Paper	<1 %
19	doctorpenguin.com Internet Source	<1 %
20	Submitted to Arab Open University Student Paper	<1 %
21	Om Prakash Jena, Mrutyunjaya Panda, Utku Kose. "Medical Data Analysis and Processing using Explainable Artificial Intelligence", CRC Press, 2023 Publication	<1 %
22	www.irejournals.com Internet Source	<1 %
23	www.isteonline.in Internet Source	<1 %
24	"Advances in Knowledge Discovery and Data Mining", Springer Science and Business Media LLC, 2018 Publication	<1 %
25	Xinyuan Song, HSIEH,WEI-CHE, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang. "A Comprehensive Guide to Explainable AI: From Classical Models to LLMs", Open Science Framework, 2024 Publication	<1 %
26	ijrpr.com Internet Source	<1 %

27	eprints.nottingham.ac.uk Internet Source	<1 %
28	opus.lib.uts.edu.au Internet Source	<1 %
29	Ilhan Uysal, Román Rodríguez Aguilar, Jafar Ahmad Abed Alzubi, Mehmet Bilen. "Digital Transformation and XAI in Healthcare", CRC Press, 2025 Publication	<1 %
30	nano-ntp.com Internet Source	<1 %
31	Submitted to King's Own Institute Student Paper	<1 %
32	www.pwc.co.uk Internet Source	<1 %
33	"Fundamentals and Methods of Machine and Deep Learning", Wiley, 2022 Publication	<1 %
34	Submitted to The Robert Gordon University Student Paper	<1 %
35	Submitted to itera Student Paper	<1 %
36	Rakan A. Alsowail. "Optimizing Network and Systems Management for Fraud Detection: A High-Performance Approach Using Random Light Gradient-Based CatBoost Ensemble with Enhanced Gold Rush Algorithm", Journal of Network and Systems Management, 2025 Publication	<1 %
37	Submitted to Sri Balaji University, Pune Student Paper	<1 %
	openresearch-repository.anu.edu.au	

38	Internet Source	<1%
39	Azadeh Mokari, Simone Eiserloh, Oleg Ryabchykov, Ute Neugebauer, Thomas Bocklitz. "A comparative study of robustness to noise and interpretability in U-Net-based denoising of Raman spectra", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2025 Publication	<1%
40	M Sivasankari, R Shakthi, L Madhu Jothi, M Kalaiyarasi. "Investigations on Swarm Optimization Algorithms for Classification of EEG Signals in Detection of Epilepsy", 2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR), 2023 Publication	<1%
41	Submitted to Massey University Student Paper	<1%
42	Mohammad Shanaa, Sherief Abdallah. "A Hybrid Anomaly Detection Framework Combining Supervised and Unsupervised Learning for Credit Card Fraud Detection", F1000Research, 2025 Publication	<1%
43	Alex Khang. "Shaping Cutting-Edge Technologies and Applications for Digital Banking and Financial Services", Routledge, 2025 Publication	<1%
44	Walayat Hussain, Luis Martínez López, Manoj Sahni, Zhen-Song Chen, Jun Liu. "Cutting-Edge Artificial Intelligence - Advances and	<1%

	Implications in Real-World Applications", CRC Press, 2025 Publication	
45	Zulfikar Ali Ansari, Md Shamsul Haque Ansari, Ahmed Khan, NZ Jhanjhi, Gopisetty Rathnamma. "Optimized and interpretable machine learning framework for early breast cancer detection", Health and Technology, 2025 Publication	<1 %
46	bth.diva-portal.org Internet Source	<1 %
47	d-nb.info Internet Source	<1 %
48	docs.google.com Internet Source	<1 %
49	tierarztliche.com Internet Source	<1 %
50	Submitted to Symbiosis International University Student Paper	<1 %
51	fastercapital.com Internet Source	<1 %
52	medium.com Internet Source	<1 %
53	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 1", CRC Press, 2025 Publication	<1 %
54	Mudavath Ravi, Atul Negi, Sanjay Chitnis. "A Comparative Review of Expert Systems, Recommender Systems, and Explainable AI",	<1 %

Professor
School of CIS
Prof. C.R. Rao Road,
Central University
Hyderabad-46 (India)

Publication

55	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	<1 %
56	Submitted to City University College of Ajman Student Paper	<1 %
57	David Gunning, David Aha. "DARPA's Explainable Artificial Intelligence (XAI) Program", AI Magazine, 2019 Publication	<1 %
58	Submitted to Hellenic Open University Student Paper	<1 %
59	Yash Panchal, Manan Mer, Abhiroop Ghosh. "Residential Property Price Prediction Using Machine Learning: MakanSETU", 2022 International Seminar on Application for Technology of Information and Communication (iSemantic), 2022 Publication	<1 %
60	www.grafiati.com Internet Source	<1 %
61	Natasa Kleanthous, Abir Hussain. "Machine Learning in Farm Animal Behavior using Python", CRC Press, 2025 Publication	<1 %
62	Pengxv Chen, Anmin Zhang, Shenwen Zhang, Taoning Dong, Xi Zeng, Shuai Chen, Peiru Shi,	<1 %

interpretation analysis method based on Transformer neural network model with multi-task classification variables", Reliability Engineering & System Safety, 2025
Publication

63	Submitted to ICTS Student Paper	<1 %
64	Nazmul Siddique, Mohammad Shamsul Arefin, Md Zahid Hasan, M Shamim Kaiser. "Applied Intelligence for Healthcare Informatics - Techniques and Applications", CRC Press, 2025 Publication	<1 %
65	Ruixin Wang, Kaijie Xu, Yixi Wang. "Augmentation of Soft Partition with a Granular Prototype Based Fuzzy C-Means", Mathematics, 2024 Publication	<1 %
66	Submitted to University of Nottingham Student Paper	<1 %
67	ebin.pub Internet Source	<1 %
68	www.frontiersin.org Internet Source	<1 %
69	yamanashi.repo.nii.ac.jp Internet Source	<1 %
70	Ashutosh Yadav, Mansaf Alam, Kiran Chaudhary. "AI-Driven Finance in the VUCA World", CRC Press, 2025 Publication	<1 %
71	Submitted to Aston University Student Paper	

72	Zhuoyu Li, Siying Hao, Shujun Shi, Lin Li, Ziwei Tao. "An explainable machine learning model for early warning of hypertensive and hypotensive anomalies in maintenance hemodialysis patients", BMC Nephrology, 2025 Publication	<1 %
73	www.tnsroindia.org.in Internet Source	<1 %
74	"Pattern Recognition. ICPR International Workshops and Challenges", Springer Science and Business Media LLC, 2021 Publication	<1 %
75	Ahmed Shany Khusheef. "Optimized multimodal anomaly detection in fused deposition modeling: real-time monitoring with clustering classifiers and data fusion", Progress in Additive Manufacturing, 2025 Publication	<1 %
76	Submitted to Dublin City University Student Paper	<1 %
77	Submitted to Galway Mayo Institute of Technology (GMIT) Student Paper	<1 %
78	Junjie Wu. "Advances in K-means Clustering", Springer Science and Business Media LLC, 2012 Publication	<1 %
79	Submitted to Liverpool John Moores University Student Paper	<1 %

	Habitat Mapping with Clustering Methods Using Remote Sensing", Information, 2023 Publication	
81	jultika.oulu.fi Internet Source	<1 %
82	mau.diva-portal.org Internet Source	<1 %
83	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
84	www.mtsu.edu Internet Source	<1 %
85	www.open-access.bcu.ac.uk Internet Source	<1 %
86	www.researchgate.net Internet Source	<1 %
87	"Hybrid Artificial Intelligent Systems", Springer Science and Business Media LLC, 2019 Publication	<1 %
88	Arvind Dagur, Dhirendra Kumar Shukla, Nazarov Fayzullo Makhmadiyarovich, Akhatov Akmal Rustamovich, Jabborov Jamol Sindorovich. "Artificial Intelligence and Information Technologies", CRC Press, 2024 Publication	<1 %
89	John Senior, Éva Gyarmathy. "The Ethics, Psychology, and Theology of AI - Exploring the Notion of Singularity", Routledge, 2025 Publication	<1 %
90	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning,	<1 %

Publication

91	Submitted to University of Witwatersrand Student Paper	<1 %
92	Usharani Bhimavarapu, Parvathaneni Naga Srinivasu. "Chapter 12 Enhancing Patient Data Clustering in Smart Healthcare: A Semi-supervised Approach for Person-Centric HealthCare Treatment and Resource Optimization", Springer Science and Business Media LLC, 2025 Publication	<1 %
93	bmlweb.vuse.vanderbilt.edu Internet Source	<1 %
94	ijsrem.com Internet Source	<1 %
95	jeroenooge.be Internet Source	<1 %
96	waseda.repo.nii.ac.jp Internet Source	<1 %
97	www.utupub.fi Internet Source	<1 %
98	"Machine Learning and Knowledge Extraction", Springer Science and Business Media LLC, 2020 Publication	<1 %
99	Submitted to York St John University Student Paper	<1 %
100	Zhenyu Xia, Suvash C. Saha. "FinGraphFL: Financial Graph-Based Federated Learning for Enhanced Credit Card Fraud Detection", Mathematics, 2025	<1 %

101	Internet Source	<1 %
102	David H. Eberly. "Game Physics", CRC Press, 2019 Publication	<1 %
103	Submitted to Higher Education Commission Pakistan Student Paper	<1 %
104	S. R. Reeja, Bore Gowda, Y. S. Rammohan, Ganesan Prabu Sankar, G. Jayalatha. "Engineering Science and Technology: Innovations for the Future", CRC Press, 2025 Publication	<1 %
105	censius.ai Internet Source	<1 %
106	dro.dur.ac.uk Internet Source	<1 %
107	www.preprints.org Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		