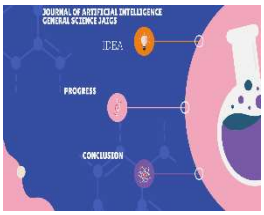
Vol.4, Issue 01, April 2024
Journal of Artificial Intelligence General Science JAIGS

https://ojs.boulibrary.com/index.php/JAIGS



An Expedited Examination of Responsible AI Frameworks: Directing Ethical AI Development

Jeff Shuford

Nationally Syndicated Business & Technology Columnist, USA

ARTICLEINFO
Article History:
Received:
01.05.2024
Accepted:
15.05.2024
Online: 22.05.2024
Keyword: artificial intelligence, software engineering, responsible AI, development frameworks, literature review

ABSTRACT

In recent years, the rapid expansion of Artificial Intelligence (AI) and its integration into various aspects of daily life have ignited significant discourse on the ethical considerations governing its application. This study addresses these concerns by swiftly reviewing multiple frameworks designed to guide the development and utilization of Responsible AI (RAI) applications. Through this exploration, we analyze each framework's alignment with the Software Development Life Cycle (SDLC) phases, revealing a predominant focus on the Requirements Elicitation phase, with limited coverage of other stages. Furthermore, we note a scarcity of supportive tools, predominantly offered by private entities. Our findings underscore the absence of a comprehensive framework capable of accommodating both technical and non-technical stakeholders across all SDLC phases, thus revealing a notable gap in the current landscape. This study sheds light on the imperative need for a unified framework encompassing all RAI principles and SDLC phases, accessible to users of varying expertise and objectives.

Introduction:

The emergence of Artificial Intelligence (AI) heralds a transformative era in science and society, as noted by Harari (2017). Alongside advancements in data processing and analysis facilitated by AI technologies (Jordan and Mitchell, 2015), the prevalence of autonomous and semi-autonomous decision systems is increasingly evident across various industries such as healthcare, automotive, banking, and manufacturing (Cornacchia et al., 2021). With AI's profound potential and widespread societal impact, discussions on the values and principles guiding its development and application have become paramount (Vayena et al., 2018; Awad et al., 2018).

Recent scholarly research and media attention have highlighted concerns surrounding AI's potential to disrupt employment, be exploited by malicious actors, evade accountability, propagate bias, and compromise fairness (n.d., 2017; Brundage et al., 2018; Zou and Schiebinger, 2018). In response, the concept of Responsible Artificial Intelligence (RAI) has been articulated, emphasizing intelligent algorithms that prioritize the needs of all stakeholders, particularly marginalized and disadvantaged users, to ensure trustworthy decision-making (Cheng et al., 2021). This entails safeguarding and informing users, mitigating adverse impacts, and maximizing long-term beneficial outcomes, with constant feedback mechanisms to uphold societal values.

In light of societal apprehensions, various public and private entities have developed resources, including ethical requirements, principles, guidelines, best practices, tools, and frameworks, to address RAI principles. This study conducts a Rapid Review (RR) of these frameworks, exploring their practical guidance and support for stakeholders involved in implementing and validating AI applications, aligning with the Software Development Life Cycle (SDLC).

Our investigation assesses the comprehensiveness and completeness of RAI frameworks in addressing principles and SDLC phases, as well as the availability of complementary tools aiding practitioners throughout the development lifecycle. The primary finding reveals a notable dearth of tools supporting the design, implementation, and auditing of RAI principles for both technical and non-technical stakeholders. Future research endeavors should concentrate on developing a comprehensive and user-friendly RAI framework, facilitating its adoption in real-world projects by AI practitioners.

The subsequent sections of this paper are structured as follows: Section 2 provides background definitions pertinent to the study; Section 3 outlines the research protocol, including research questions and methodology; Section 4 presents the results of the rapid review and addresses the research questions; Section 5 discusses the findings, emphasizing key insights; Section 6 considers potential validity threats, and finally, Section 7 offers conclusions and avenues for future research.

Background

To provide a foundation for our study, we offer preliminary definitions to elucidate the concepts guiding our work.

Responsible AI Principles

Numerous national and international organizations have established specialized expert groups on AI to address associated risks, often tasked with crafting policy documents. Notable entities include the European Commission's High-Level Expert Group on Artificial Intelligence, the UNESCO Ad Hoc Expert Group for the Recommendation on the Ethics of Artificial Intelligence, the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, the NASA Artificial Intelligence Group, and the UK AI Council, among others.

These committees have been tasked with generating reports and guidelines concerning Responsible AI (RAI). Similar initiatives are emerging within the commercial sector, particularly among AI-dependent businesses. Companies like Sony and Meta have made their AI policies and principles publicly available. Additionally, professional organizations and non-profit groups such as UNI Global Union and the Internet Society have issued statements and recommendations.

The substantial efforts of this diverse array of stakeholders in developing RAI principles and policies not only underscore the necessity for ethical guidance but also reflect their vested interest in shaping AI ethics according to their distinct priorities (Greene et al., 2019). Notably, there has been scrutiny of the private sector's involvement in AI ethics, with concerns raised that it may employ high-level soft policies either to frame a social issue as technical or to circumvent regulation altogether (Bay, 2018; Jobin et al., 2019).

However, numerous studies have highlighted how these proposals often diverge, leading to what is termed as principle proliferation (Floridi and Cowls, 2019). Consequently, various in-depth investigations have been undertaken, such as the study by Jobin et al. (2019), which identified a global convergence around five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy.

Jobin et al. (2019) noted that none of these ethical principles were present in all the documents they reviewed; however, these five principles were mentioned in more than half of the sources examined. Moreover, further thematic analysis revealed significant semantic and conceptual divergences in interpreting these principles and the specific recommendations or areas of concern derived from each of them.

Selected Definitions of AI Principles

As outlined in Section 2.1, there exists considerable uncertainty and nuance surrounding the definition of principles that primarily characterize Responsible AI, as well as regarding the definition of RAI itself. Indeed, it is sometimes referred to as Trustworthy or Ethical AI. In our study, we tackle the issue of principle proliferation by opting to focus on a specific subset of those that characterize RAI, specifically the four principles identified by Jobin et al. (2019), excluding responsibility due to its lack of clear definition.

Furthermore, to provide authoritative and precise definitions for each principle, we have chosen to utilize those offered by the High-Level Expert Group on Artificial Intelligence established by the European Commission in their Ethics guidelines for trustworthy AI (AIHLEG, 2018).

Below, we present the selected definitions for each principle. We have aligned the principles outlined in the High-Level Expert Group on AI (AIHLEG, 2018) with those identified by Jobin et al. (2019), and where the naming convention differs, we have indicated this in parenthesis. We have also mapped principles to system requirements.

Transparency

This requirement is closely linked with the concept of *explicability* [...] Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. [...]. [The] explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). [...]

Diversity and non-discrimination and fairness (Justice and fairness)

In order to achieve Trustworthy AI, [one] must enable inclusion and diversity throughout the entire AI system's life cycle. [...] this also entails ensuring equal access through inclusive design processes as well as equal treatment. [...] Bias [derives from] data sets used by AI systems (both for training and operation) [because these] may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. [...]

Technical robustness and safety (Non-maleficence)

[...] Technical robustness requires that AI systems are developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

Frameworks

In this study, our attention is directed towards frameworks that operationalize the aforementioned ethical principles of RAI.

The concept of frameworks is well-established within the realm of Software Engineering (SE). As far back as 1997, Johnson et al. (Johnson, 1997) described frameworks as "an object-oriented reuse technique" or "the skeleton of an application that can be customized by an application developer." These definitions are not contradictory; the former outlines the structure of a framework while the latter delineates its purpose.

Expanding beyond SE to a broader context, frameworks represent a form of design reuse. They can be viewed as a compendium of recommendations, guidelines, and tools to adhere to in creating a product that aligns with a predefined standard.

Study Design

Rapid Reviews (RRs) have emerged as a streamlined approach for swiftly synthesizing evidence, originally developed to assist healthcare decision-makers in promptly addressing urgent and emerging needs (Konnyu et al., 2012). By simplifying systematic review methods, rapid reviews focus on literature search efficiency while still striving to yield valid conclusions (Watt et al., 2008).

To conduct this rapid review, we adhered to the protocol proposed by Cartaxo et al. (2018), supplementing the Rapid Review process with strategies outlined by Kitchenham and Charters (2007) for systematic literature reviews. The subsequent subsections provide detailed insights into the study design and its execution.

Planning the Review

The rapid literature review presented in this study entailed the following steps:

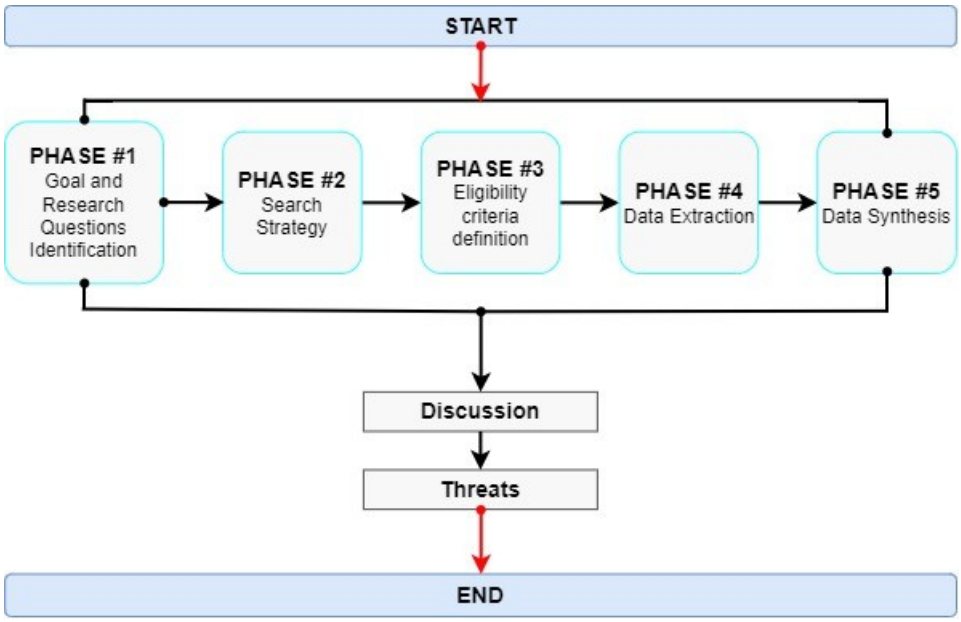
1. Goal and Research Questions: Identification of the overarching goal and associated research questions to guide the literature review.

2. Search Strategy: Formulation of a strategy to retrieve prior works published in the literature, encompassing research databases and query strings.

3. Eligibility Criteria Definition: Establishment of criteria utilized to filter collected studies.

4. Data Extraction: Definition of the process for extracting relevant data to facilitate addressing the research questions.

5. Data Synthesis: Determination of the methodology for organizing extracted relevant data to address the research questions.



Results

The Rapid Review (RR) was conducted part-time from November 15, 2022, to December 12, 2022, following the procedure outlined in Section 3.3.

Table 1: Summary of Documents Collected Grouped by Research Phase

Table 1 provides an overview of the number of relevant results obtained in each sub-phase. The search on Google Scholar yielded no useful results. In total, after consolidating resources from identified data sources, we amassed 148 unique resources, excluding duplicates.

Data Synthesis

Data were synthesized, and various statistics were computed to address the research questions (refer to the online appendix available at Barletta et al., 2023). Tables 2 and 3 offer a representative excerpt of the comprehensive dataset collected.

Research Question 1: What are the Responsible AI frameworks proposed in the literature?

All retrieved frameworks were categorized based on the type of institution proposing them, classified into three categories: Companies, Universities, and Non-Profit Organizations/Communities/Public Entities (NPG/COMM/PE). If an entity proposed multiple frameworks, it was counted only once, aiming to assess the distribution of proposals by entity type.

As depicted in Figure 2, the majority of the filtered frameworks were proposed by NPG/COMM/PE (50.7%, 70/138), followed by lucrative Companies (31.9%, 44/138), and then Universities (17.4%, 24/138).

Furthermore, the frameworks were classified into four categories based on their characteristics:

- Principle (P): Highlighting abstract ethical principles or moral values.

- Guideline (G): Offering concrete guidelines quickly translatable into design constraints or choices.

- Tool (T): Capable of verifying compliance with one or more principles and/or aiding practitioners in implementing principles or guidelines.

Data Source	**Resources retrieved**	**Resources analyzed**	**Resource selected**
Scopus	1875	1489	20
Google Scholar	91200	200	0
Algorithm Watch	167	167	80
OECD DB	356	70	38
Google Search	21,100,00	168	10

Table 2: Excerpt of the whole data classified by RAI principles taken into consideration.

Entity name	ToolName	Diversity Non-discrimination & Fairness	Privacy & Data Governance	Technical Robustness & Safety	Transparency
COMPANIES					
Meta	Facebook's five pillars of Responsible AI	Yes	Yes	Yes	Yes
UNIVERSITIES					
University of Texas at Austin	CERTIFAI	Yes	No	Yes	Yes
NO-PROFIT ORG / COMMUNITIES / GOVERNMENT ENTITIES					
NIST	AI Risk Management Framework	Yes	Yes	Yes	Yes

Table 3: Excerpt of the whole data classified by SDLC phase addressed.

Entity name	Requirements Elicitation	Design	Development	Testing	Deployment
COMPANIES					
Meta	Yes	No	No	No	No
UNIVERSITIES					
University of Texas at Austin (CERTIFAI)	No	Yes	Yes	Yes	No
NO-PROFIT ORG / COMMUNITIES / PUBLIC ENTITIES					
NIST	Yes	Yes	No	No	No

- **Other (O)**: if a resource cannot be classified into any of these categories - e.g. a list of possible attacks against an AI algorithm or a list of several questions to check in the design phase.

In Figure 3, the distribution of frameworks is illustrated by their category and grouped by proposing institution.

It is noteworthy that Companies and Non-Profit Organizations/Communities/Public Entities (NPG/COMM/PE) primarily proposed Principles (42.59% and 59.72%, respectively), whereas Universities predominantly proposed Guidelines (57.69%). In terms of the number of resources proposed, Companies and Universities accounted for 12.96% and 15.38% of Tools, respectively, while this percentage was notably lower for NPG/COMM/PE (4.17%).

Research Question 2: How much do these frameworks address various RAI principles?

In our rapid review, our focus lies mainly on analyzing frameworks that offer practical support to all stakeholders involved in the development and deployment of AI applications. Therefore, in addressing Research Question 2, we excluded frameworks categorized as Principle, which solely encompass ethical values without offering actionable advice. Additionally, we counted frameworks rather than entities; if an entity proposed multiple frameworks, each was counted individually.

Furthermore, it's essential to note that in this analysis, a principle is considered "covered" even if it's only "partially" addressed, meaning not every aspect related to that principle is discussed. For example, frameworks that address privacy solely in terms of data acquisition and storage, without addressing potential privacy attacks such as "model inversion" attack (Fredrikson et al., 2015), are still considered to cover the privacy principle.

As depicted in Figure 4, the majority of frameworks address all four RAI principles. However, there are frameworks that cover only one (15.49%) or two principles (15.49%). Particularly when three principles are addressed, the most covered are Diversity & Non-discrimination & Fairness, Privacy & Data Governance, and Transparency (9.3% for Companies, 15.8% for Universities, and 14.1% for NPG/COMM/PE). For frameworks covering two principles, the most common are Diversity & Non-discrimination & Fairness and Transparency (9.3% for Companies and 16.9% for NPG/COMM/PE). For comprehensive results, please refer to the appendix.

Conclusions

This study undertook a rapid review to offer an insight into frameworks proposed in both white and grey literature aimed at facilitating and expediting the adoption of Responsible Artificial Intelligence (RAI) practices. We formulated four research questions to obtain specific insights aligning with our research goal, thereby enhancing the informative value of this survey. To delve deeper into our investigation, we categorized the entities providing each framework into three distinct groups: Companies, Universities, and Non-Profit Organizations/Communities/Public Entities (NPG/COMM/PE), enabling us to analyze results based on the proposing entity. Additionally, this review encompassed not only scientific articles but also grey literature resources.

The findings of this study can be summarized as follows:

- Diversity of Perspectives vs. Lack of Standardization (F1): RAI frameworks are offered by a myriad of heterogeneous entities, both public and private. While this diversity enriches perspectives and fosters AI democratization, it also underscores the lack of consensus and standardization regarding best practices for RAI compliance with ethical values.

- Theoretical vs. Practical Support for AI Practitioners (F2): We classified each framework based on the Software Development Life Cycle (SDLC) phases covered. Our analysis revealed that only a few frameworks span all SDLC phases and provide practical guidance to practitioners involved in developing, testing, and deploying RAI applications. There is a notable dearth of practical validation techniques for theoretical principles and implementation guidelines. Moreover, there is a concerning deficit of tools supporting all stakeholders during the implementation and auditing phase.

- No Comprehensive Framework (F3): Despite many frameworks covering all four selected principles, we observed instances where principles were only partially addressed, mirroring the lack of standardization highlighted in F1. This rapid review underscores the current literature and industry's deficiency in complete, uniform, organized, and user-friendly RAI frameworks capable of supporting stakeholders throughout the entire Software Development Life Cycle (SDLC). It is evident that no comprehensive framework currently exists whose knowledge can be easily navigated and utilized by various stakeholders (both technical and non-technical), simplifying and accelerating the adoption of RAI practices.

References List:

[1]. Prakash, S., Malaiyappan, J. N. A., Thirunavukkarasu, K., & Devan, M. (2024). Achieving Regulatory Compliance in Cloud Computing through ML. AIJMR-Advanced International Journal of Multidisciplinary Research, 2(2).

[2]. Malaiyappan, J. N. A., Prakash, S., Bayani, S. V., & Devan, M. (2024). Enhancing Cloud Compliance: A Machine Learning Approach. AIJMR-Advanced International Journal of Multidisciplinary Research, 2(2).

[3]. Devan, M., Prakash, S., & Jangoan, S. (2023). Predictive Maintenance in Banking: Leveraging AI for Real-Time Data Analytics. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(2), 483-490.

[4]. Eswaran, P. K., Prakash, S., Ferguson, D. D., & Naasz, K. (2003). Leveraging Ip For Business Success. International Journal of Information Technology & Decision Making, 2(04), 641-650.

[5]. Prakash, S., Malaiyappan, J. N. A., Thirunavukkarasu, K., & Devan, M. (2024). Achieving Regulatory Compliance in Cloud Computing through ML. AIJMR-Advanced International Journal of Multidisciplinary Research, 2(2).

[6]. Malaiyappan, J. N. A., Prakash, S., Bayani, S. V., & Devan, M. (2024). Enhancing Cloud Compliance: A Machine Learning Approach. AIJMR-Advanced International Journal of Multidisciplinary Research, 2(2).

[7]. Biswas, A. (2019). Media Insights Engine for Advanced Media Analysis: A Case Study of a Computer Vision Innovation for Pet Health Diagnosis. International Journal of Applied Health Care Analytics, 4(8), 1-10.

[8] Chopra, B., & Raja, V. (2024). Toward Enhanced Privacy in Digital Marketing: An Integrated Approach to User Modeling Utilizing Deep Learning on a Data Monetization Platform. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 1(1), 91-105.

[9]. Raja, V. (2024). Fostering Privacy in Collaborative Data Sharing via Auto-encoder Latent Space Embedding. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 4(1), 152-162.

[10]. Raja, V. ., & chopra, B. . (2024). Exploring Challenges and Solutions in Cloud Computing: A Review of Data Security and Privacy Concerns. Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023, 4(1), 121–144. https://doi.org/10.60087/jaigs.v4i1.86

[11]. SARIOGUZ, O., & MISER, E. (2024). Data-Driven Decision-Making: Revolutionizing Management in the Information Era. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 4(1), 179-194.

[12]. Raja, V. (2024). Exploring Challenges and Solutions in Cloud Computing: A Review of Data Security and Privacy Concerns. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 4(1), 121-144.

[13]. Biswas, A. (2019). Media Insights Engine for Advanced Media Analysis: A Case Study of a Computer Vision Innovation for Pet Health Diagnosis. International Journal of Applied Health Care Analytics, 4(8), 1-10.

[14]. Talati, D. (2023). Telemedicine and AI in Remote Patient Monitoring. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 254-255.

[15]. Talati, D. (2023). Artificial Intelligence (Ai) In Mental Health Diagnosis and Treatment. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 251-253.

[16]. Talati, D. (2023). AI in healthcare domain. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2(3), 256-262.

[17]. Talati, D. (2024). AI (Artificial Intelligence) in Daily Life. Authorea Preprints.

[18]. Bhati, D., & Gupta, V. (2015). Survey—A comparative analysis of face recognition technique. Int. J. Eng. Res. General Sci, 3(2), 597-609.

[19]. Francese, R., Guercio, A., Rossano, V., & Bhati, D. (2022, June). A Multimodal Conversational Interface to Support the creation of customized Social Stories for People with ASD. In Proceedings of the 2022 International Conference on Advanced Visual Interfaces (pp. 1-5).

[20]. Jamonnak, S., Bhati, D., Amiruzzaman, M., Zhao, Y., Ye, X., & Curtis, A. (2022). VisualCommunity: a platform for archiving and studying communities. Journal of Computational Social Science, 5(2), 1257-1279.

[21]. Bhati, D., Amiruzzaman, M., Jamonnak, S., & Zhao, Y. (2021, December). Interactive visualization and capture of geo-coded multimedia data on mobile devices. In International Conference on Intelligent Human Computer Interaction (pp. 260-271). Cham: Springer International Publishing.

[22]. Joshi, R., Trivedi, M. C., Goyal, V., & Bhati, D. (2022). DNA Sequence in Cryptography: A Study. In Advances in Data and Information Sciences: Proceedings of ICDIS 2022 (pp. 557-563). Singapore: Springer Nature Singapore.

[23]. Bhati, D. (2017). Face Recognition Stationed on DT-CWT and Improved 2DPCA employing SVM Classifier. International Journal of Computer Applications, 975, 8887.

[24]. Wu, T. H., Amiruzzaman, M., Zhao, Y., Bhati, D., & Yang, J. (2024). Visualizing Routes With AI-Discovered Street-View Patterns. IEEE Transactions on Computational Social Systems.

[25]. Srivastava, A., Chandra, M., Saha, A., Saluja, S., & Bhati, D. (2023, June). Current Advances in Locality-Based and Feature-Based Transformers: A Review. In International Conference on Data & Information Sciences (pp. 321-335). Singapore: Springer Nature Singapore.

[26]. Bhati, D., Guercio, A., Rossano, V., & Francese, R. (2023, July). BookMate: Leveraging Deep Learning to Empower Caregivers of People with ASD in Generation of Social Stories. In 2023 27th International Conference Information Visualisation (IV) (pp. 403-408). IEEE.

[27]. Joshi, R., Trivedi, M. C., Goyal, V., & Bhati, D. (2022). Recent Trends for Practicing Steganography Using Audio as Carrier: A Study. In Advances in Data and Information Sciences: Proceedings of ICDIS 2022 (pp. 549-555). Singapore: Springer Nature Singapore.

[28]. Pal, M., Bhati, D., Kaushik, B., & Banka, H. Solving Classification Problem using Reduced Dimension and Eigen Structure in RSVM. International Journal of Computer Applications, 975, 8887.