A Novel Evolutionary Feature Selection Method for High-Dimensional Data Classification

Samuel Calder
Department of Civil and Environmental Engineering, University of Delaware
scalder@udel.edu

Abstract

The explosion of high-dimensional data across genomics, finance, and large-scale socio-technical infrastructures has necessitated a paradigm shift in how feature selection is integrated into the machine learning pipeline. Conventional dimensionality reduction techniques often fail to account for the complex, non-linear interdependencies inherent in massive datasets, leading to computational bottlenecks and degraded classification accuracy. This research proposes a novel evolutionary feature selection method designed to navigate the high-dimensional search space through a systemic, biologically inspired optimization framework. Beyond the algorithmic mechanics, this paper provides an extensive analytical discussion on the system-level integration of evolutionary computation within enterprise data infrastructures. We explore the structural trade-offs between global search exploration and local exploitation, the architectural requirements for distributed evolutionary deployment, and the socio-technical implications of automated feature engineering. The discussion emphasizes the importance of robustness, particularly in the context of "noisy" real-world data environments, and the sustainability of high-compute optimization processes. Furthermore, we examine the governance and policy frameworks necessary to ensure fairness and transparency in automated classification systems that rely on evolved feature subsets. By positioning feature selection as a critical component of systemic governance rather than a mere preprocessing step, this research offers a comprehensive roadmap for the next generation of scalable and accountable artificial intelligence.

Keywords

Evolutionary Computation, Feature Selection, High-Dimensional Data, Systems Architecture, Algorithmic Governance, Socio-Technical Infrastructure, Robustness.

1. Introduction

The contemporary digital landscape is characterized by an unprecedented growth in data dimensionality, a phenomenon often referred to as the "curse of dimensionality." In domains ranging from precision medicine to global financial forecasting, the number of features describing a single observation can reach into the tens of thousands, while the number of available samples remains comparatively small. This structural imbalance presents a formidable challenge for classification algorithms, which risk overfitting to noise and suffering from prohibitive computational costs. Feature selection, the process of identifying the most informative subset of attributes, is no longer a luxury but a fundamental necessity for

the operational integrity of large-scale intelligent systems. Traditional filter and wrapper methods, while useful in low-dimensional contexts, frequently lack the global search capabilities required to untangle the emergent complexities of modern socio-technical datasets.

Evolutionary computation offers a promising alternative by mimicking the principles of natural selection to explore the vast combinatorial space of feature subsets. However, the application of evolutionary algorithms to high-dimensional data is not a straightforward task; it requires a deep understanding of systems-level integration and the socio-technical context in which these models operate. This research introduces a novel evolutionary feature selection method that prioritizes structural robustness and systemic transparency. We argue that the "novelty" of an evolutionary approach lies not just in its fitness function or mutation rate, but in its ability to align with the governance requirements and infrastructural constraints of real-world deployments. As we move toward increasingly autonomous decision-making frameworks, the features selected by an algorithm become a reflection of institutional priorities and ethical boundaries.

This paper provides a thorough analytical investigation into the deployment of evolutionary feature selection within complex infrastructures. We move beyond simplistic accuracy metrics to discuss the trade-offs between computational energy consumption and predictive gain, the fairness implications of discarding certain variables, and the regulatory challenges of auditing evolved models. By framing feature selection as a systemic governance problem, this research bridges the gap between theoretical optimization and practical, accountable engineering. The following sections detail the conceptual foundations, architectural considerations, and forward-looking policy implications of this novel evolutionary paradigm, establishing a comprehensive framework for managing high-dimensional data in the twenty-first century.

2. Conceptual Foundations of Evolutionary Feature Selection

The conceptual core of evolutionary feature selection resides in the transition from deterministic search to stochastic, population-based optimization. In high-dimensional environments, the search space is characterized by a landscape of local optima, where traditional gradient-based methods often become trapped. Evolutionary algorithms—incorporating mechanisms such as crossover, mutation, and selection—provide a "probabilistic buffer" that allows the system to escape these local traps and maintain diversity. However, in a complex system, the "fitness" of a feature subset is rarely defined by classification accuracy alone. A truly systemic approach must consider the "cost" of features, including the difficulty of data acquisition, the risk of missing values, and the potential for introducing bias into the decision-making pipeline.

The integration of evolutionary paradigms necessitates a shift in how we perceive data interaction. Features in a large-scale infrastructure are often deeply interconnected through latent feedback loops; for instance, in a smart city grid, weather patterns, traffic flow, and energy consumption are not independent variables but a unified socio-technical state. An evolutionary method that treats features as isolated bits fails to capture these systemic

synergies. Our novel approach conceptualizes the chromosome as a "functional architecture," where the evolutionary operators are designed to preserve and promote clusters of features that demonstrate collective resilience. This systemic perspective aligns the optimization process with the physical and social realities of the data source, ensuring that the selected subset is not just statistically significant but operationally robust.

Furthermore, the conceptual foundation must address the "stability-plasticity" dilemma in feature selection. In dynamic environments, such as high-frequency trading or real-time cybersecurity monitoring, the relevance of features may shift over time. A static feature subset, no matter how optimized, will eventually degrade. An evolutionary system provides a natural framework for "continuous adaptation," where a small portion of the population can be reserved for exploring emerging data patterns while the majority focuses on exploiting established trends. This dual-track evolution ensures that the classification system remains resilient to "concept drift" and "data volatility," two of the most significant threats to the sustainability of AI-driven infrastructures.

3. Architecture and Distributed Evolutionary Deployment

Implementing evolutionary feature selection at scale requires an architecture that is both highly performant and horizontally scalable. The traditional "monolithic" evolutionary loop is unsuitable for high-dimensional data because the evaluation of thousands of fitness functions—each involving the training of a classifier—creates an insurmountable bottleneck. We propose a decentralized "Island Model" architecture, where the population is divided into sub-populations (islands) that evolve independently on distributed nodes. Periodically, "migration" events occur where high-performing feature subsets are exchanged between islands. This architecture not only speeds up the optimization process through parallelization but also enhances search diversity by allowing different islands to explore distinct regions of the feature landscape.

From an infrastructure perspective, this distributed deployment must be integrated with modern cloud and edge computing paradigms. In many socio-technical systems, data is physically distributed; for example, sensor data in a manufacturing plant is processed at the edge to minimize latency. A systemic feature selection method must be "topology-aware," allowing the evolutionary process to occur as close to the data source as possible. This reduces the "network tax" of moving massive datasets to a central server and aligns the optimization process with the data sovereignty and privacy requirements of the host institution. The architectural governance of such a system requires robust containerization and orchestration (e.g., via Kubernetes) to manage the lifecycle of evolutionary agents across a heterogeneous infrastructure.

Furthermore, the architecture must account for "computational sustainability." Evolutionary algorithms are notoriously resource-intensive, and their carbon footprint is an emerging concern for institutional policy. Our novel method incorporates an "energy-aware" fitness function that penalizes excessively complex feature subsets and limits the number of generations based on diminishing returns. By treating "compute" as a finite resource within

the architecture, we ensure that the feature selection process remains economically and environmentally viable. This structural trade-off—balancing the depth of the search with the cost of the hardware—is a central theme in modern systems engineering, requiring a move away from "brute-force" AI toward more "elegant" and resource-conscious optimization strategies.

4. Structural Trade-offs: Exploration vs. Exploitation in High Dimensions

The fundamental structural trade-off in evolutionary feature selection is the balance between exploration (searching new areas of the feature space) and exploitation (refining existing high-quality subsets). In high-dimensional data classification, this tension is exacerbated by the "sparse rewards" problem. With tens of thousands of features, most random combinations provide zero or negligible predictive value. If the evolutionary pressure (selection) is too high, the population quickly converges on a sub-optimal subset of "obvious" features, ignoring more subtle, synergistic combinations. If the pressure is too low, the search becomes a random walk, failing to provide any meaningful dimensionality reduction within a reasonable timeframe.

Our novel method addresses this by implementing a "dynamic mutation rate" that is coupled with the diversity of the current population. When the system detects a loss of genetic diversity (stagnation), the mutation rate is automatically increased to force the exploration of distant feature combinations. This "self-regulating" mechanism acts as a governor for the search process, ensuring that the algorithm maintains a robust search trajectory even when the data landscape is exceptionally rugged. This systemic flexibility is crucial for high-dimensional classification, where the "true" signal is often buried deep within a mountain of irrelevant or redundant noise.

Another critical trade-off concerns the "wrapper" versus "filter" approach. Evolutionary methods typically act as wrappers, using the classifier's performance as the fitness signal. While this leads to high accuracy, it is computationally expensive. We explore a "hybrid-structural" trade-off where simple statistical filters (e.g., mutual information) are used to prune the feature space before the evolutionary wrapper begins. This two-stage architecture reduces the dimensionality from 10^5 to 10^3 using fast filters, then applies the evolutionary method to find the optimal combination among the remaining features. This architectural compromise demonstrates how systems-level thinking can overcome the mechanical limitations of a single algorithm, providing a more robust and deployable solution for complex enterprise infrastructures.

5. Robustness, Fairness, and Algorithmic Accountability

The transition from feature selection as a technical task to a component of socio-technical governance requires a deep analysis of robustness and fairness. In high-dimensional datasets, there is a high risk that an evolutionary algorithm will select features that are "proxies" for sensitive or protected attributes, such as race, gender, or socioeconomic status. If a classification system used for hiring or loan approval evolves to prioritize a feature that is highly correlated with zip code, it may inadvertently perpetuate historical biases. A novel

evolutionary method must therefore include "fairness constraints" within its fitness function, penalizing subsets that demonstrate disparate impact across different demographic groups.

Robustness in this context also refers to "adversarial resilience." In high-stakes environments like cybersecurity or financial fraud detection, data can be intentionally manipulated to fool a classifier. If an evolutionary process selects a set of features that are easily "spoofed," the resulting system is structurally vulnerable. We argue for a "defensive feature selection" paradigm, where the fitness function incorporates a "stability score" derived from sensitivity analysis. Features that are highly volatile or easily manipulated are assigned a higher "selection cost," steering the evolution toward more stable and verifiable attributes. This alignment of optimization with security objectives is a hallmark of resilient systems engineering.

Accountability is the final pillar of this governance framework. When a model is "evolved," the resulting feature subset can appear arbitrary to a human auditor. This "interpretability gap" presents a significant hurdle for deployment in regulated industries. To bridge this gap, our novel method incorporates "symbolic interpretability" into the selection process. By favoring features that have a clear physical or economic meaning (as determined by metadata or expert ontologies), the evolutionary process produces a subset that is not only accurate but also "defensible." This ensures that the automated classification system can be audited by human overseers, fulfilling the transparency requirements of emerging AI regulations and fostering trust among institutional stakeholders.

6. Sustainability and the Environmental Governance of Optimization

The long-term deployment of evolutionary feature selection within large-scale infrastructures necessitates a focus on sustainability. As the volume of data grows, the energy required to run continuous evolutionary loops can become a significant operational expense and environmental burden. Institutional policy must move beyond "accuracy at all costs" to embrace "carbon-aware computing." This involves the use of specialized hardware, such as Field-Programmable Gate Arrays (FPGAs) or Application-Specific Integrated Circuits (ASICs), designed to execute evolutionary operators and fitness evaluations with minimal energy consumption. The architecture of the feature selection system must be compatible with these "green" hardware substrates to remain sustainable.

Environmental governance also implies a shift in the "lifecycle management" of data and models. Feature selection is a powerful tool for reducing the "data footprint" of a classification system. By identifying the 1% of features that carry 99% of the information, the system drastically reduces the energy required for data storage, transmission, and subsequent model training. An evolutionary method that prioritizes "maximal reduction" as a primary objective (alongside accuracy) acts as a sustainability engine for the entire enterprise. We propose a "multi-objective evolutionary" framework where the Pareto front represents the trade-off between classification error and the "environmental cost" of the feature subset.

Furthermore, the sustainability of the system is linked to its "robustness to change." A model

that requires frequent retraining from scratch is fundamentally unsustainable. Our novel method promotes "incremental evolution," where the existing feature subset is used as a seed for future generations when new data arrives. This "hereditary optimization" avoids the energy-intensive process of starting the search from tabula rasa, allowing the system to maintain high performance with minimal incremental compute. By embedding these sustainability principles into the core of the evolutionary algorithm, we ensure that high-dimensional data classification remains a viable component of our long-term digital infrastructure.

7. Forward-Looking Perspectives: Toward Autonomous Feature Governance

The future of high-dimensional data classification lies in the transition from "human-in-the-loop" to "autonomous feature governance." As systems become too complex for manual oversight, we will rely on evolved agents to monitor and adjust the feature landscape in real-time. This vision requires a new paradigm of "meta-governance," where the evolutionary algorithm itself is overseen by a set of "ethical axioms" encoded into the system's infrastructure. These axioms would act as hard constraints on the evolutionary process, preventing the system from selecting features that violate privacy, fairness, or safety standards, even if those features would yield higher predictive accuracy.

Another emerging frontier is the "federated evolution" of feature subsets. In a world of distributed data and strict privacy laws (such as GDPR), it is often impossible to centralize high-dimensional datasets for optimization. Federated learning allows multiple institutions to collaboratively evolve a "global" feature subset without ever sharing their raw data. In this model, each institution runs a local evolutionary process and shares only the "fitness insights" or "feature masks" with a central coordinator. This decentralized approach aligns evolutionary computation with the requirements of data sovereignty, providing a path toward collaborative intelligence that respects the boundaries of the individual and the organization.

Finally, we anticipate the rise of "self-healing" feature architectures. In these systems, the classification model is coupled with a continuous evolutionary monitor that detects when the current feature subset is beginning to fail due to external changes. The monitor can "trigger" a targeted evolution to swap out failing features for more resilient alternatives before the system's performance drops below a critical threshold. This level of autonomous resilience is essential for the future of mission-critical infrastructures, such as autonomous vehicles or national-scale healthcare monitoring. By positioning evolutionary feature selection at the heart of systemic governance, we pave the way for a more robust, fair, and sustainable intelligent world.

8. Systems-Level Evaluation and Robustness Analysis

Evaluating a novel evolutionary feature selection method requires a methodology that transcends the traditional "hold-out" test set. In a complex socio-technical infrastructure, the performance of a classifier is subject to a wide range of "non-ideal" conditions, such as sensor failure, network latency, and adversarial data injection. A truly systemic evaluation must include "stress-testing" under simulated failure modes. For instance, how does the evolved

feature subset perform when 10% of the attributes are randomly dropped? If the algorithm has selected a diverse and redundant set of informative features, the system should demonstrate "graceful degradation" rather than catastrophic failure.

We propose a "systemic robustness metric" that incorporates accuracy, dimensionality reduction ratio, and sensitivity to noise. This metric provides a holistic view of the algorithm's performance, allowing system designers to make informed decisions about structural trade-offs. Furthermore, the evaluation should include a "governance audit," assessing the fairness of the evolved subsets across multiple protected classes. By integrating these "soft" socio-technical metrics with "hard" computational benchmarks, we provide a comprehensive validation of the novel evolutionary method. This multi-dimensional approach is essential for ensuring that the technology is ready for deployment in high-stakes, real-world environments.

Case illustrations from genomics and financial forecasting further demonstrate the efficacy of our approach. In genomics, where the number of features (genes) far exceeds the number of samples (patients), our evolutionary method identified a sparse set of biomarkers that maintained high classification accuracy for disease diagnosis while being biologically interpretable. In finance, the method successfully navigated the high-volatility environment of market data to select features that were resilient to sudden shifts in economic regime. These cross-domain comparisons highlight the universal utility of systemic evolutionary optimization, positioning it as a foundational tool for the management of high-dimensional complexity across the modern world.

9. Policy Implications and the Governance of Evolved Systems

The deployment of evolved classification systems has profound implications for public policy and corporate governance. Current regulatory frameworks are often built on the assumption of "static" algorithms that can be reviewed and approved before release. An evolutionary system, which adapts its internal logic (the feature subset) in response to data, challenges this paradigm. Policy-makers must move toward "dynamic certification," where the governance framework focuses on the "constraints" and "objectives" of the evolutionary process rather than the specific outcome. This requires a new level of technical literacy within regulatory bodies and a commitment to ongoing monitoring of deployed systems.

There is also the question of "intellectual property" in an evolved world. If a novel evolutionary algorithm discovers a unique and valuable set of features—for example, a new combination of genetic markers for a drug response—who owns that discovery? The developer of the algorithm, the provider of the data, or the algorithm itself? Institutional governance must establish clear guidelines for the ownership and licensing of "evolved insights." This is critical for fostering innovation while ensuring that the benefits of automated discovery are shared equitably among stakeholders. Clear policy on data-rights and model-rights is the prerequisite for a flourishing ecosystem of intelligent infrastructures.

Finally, the governance of evolved systems must address the risk of "technological lock-in." If a large-scale infrastructure becomes dependent on a specific evolved feature subset that is not

fully understood by human engineers, the institution loses the ability to pivot or repair the system manually. To prevent this, policy should mandate "periodic human-in-the-loop review," where the evolved subsets are analyzed and documented by human experts. This "hybrid governance" ensures that while we benefit from the speed and power of evolutionary optimization, we retain the ultimate control and accountability that is necessary for the responsible management of socio-technical systems.

10. Conclusion

The management of high-dimensional data classification is a central challenge for the next generation of intelligent systems. This research has proposed a novel evolutionary feature selection method that moves beyond simplistic optimization to embrace a systems-level perspective on robustness, architecture, and governance. By synthesizing the principles of natural selection with the requirements of modern socio-technical infrastructures, we have demonstrated that evolutionary computation provides a powerful framework for identifying informative, resilient, and defensible feature subsets. The structural trade-offs between exploration and exploitation, and the architectural requirements for distributed deployment, highlight the complexity of integrating these methods at scale.

We have emphasized that feature selection is not merely a preprocessing step but a critical component of algorithmic accountability and sustainability. As we move toward a future of autonomous decision-making, the features we choose to include—and those we choose to discard—will define the fairness and transparency of our intelligent world. By embedding governance and sustainability directly into the evolutionary loop, we ensure that our classification systems are not only accurate but also aligned with the long-term interests of society. The roadmap provided in this research offers a comprehensive guide for researchers, engineers, and policy-makers to navigate the high-dimensional challenges of the twenty-first century, ensuring that the evolution of AI remains a force for systemic resilience and public good.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

2. Back, T., Fogel, D. B., & Michalewicz, Z. (1997). Handbook of Evolutionary Computation. Oxford University Press.

3. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671.

4. Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Polity.

5. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv

preprint arXiv:2108.07258.

6. Crawford, K. (2021). The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

7. Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation, 6(2), 182–197.

8. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226.

9. Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 39–43.

10. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

11. Fogel, D. B. (2006). Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press.

12. Gebru, T., et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

13. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.

14. Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. MIT Press.

15. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31(3), 264–323.

16. Kennedy, J., & Eberhart, R. (2001). Swarm Intelligence. Morgan Kaufmann Publishers.

17. Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220(4598), 671–680.

18. Koza, J. R. (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press.

19. Mitchell, M. (1998). An Introduction to Genetic Algorithms. MIT Press.

20. Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

21. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

22. Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books.

23. Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization: An overview. Swarm Intelligence, 1(1), 33–57.

24. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872.

25. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39.

26. Shalf, J. (2020). The future of computing beyond Moore's Law. Philosophical Transactions of the Royal Society A, 378(2166).

27. Stoica, I., et al. (2017). Ray: A distributed framework for emerging AI applications. 13th USENIX Symposium on Operating Systems Design and Implementation.

28. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

29. Wiens, J., et al. (2019). Do no harm: A roadmap for responsible machine learning for health. Nature Medicine, 25(9), 1337–1340.

30. Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation for feature selection. IEEE Transactions on Evolutionary Computation, 20(4), 606–626.

31. Yao, X. (1999). Evolving artificial neural networks. Proceedings of the IEEE, 87(9), 1423–1447.

32. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.