Improving Feature Selection Using Genetic Algorithm for Regression and Classification Tasks

Vedant Marwadi - 23208466
Nature-Inspired Computing
Master of Computer and Information Sciences
Auckland University of Technology

***Abstract*—In a world where data is abundant yet often overwhelming, the challenge lies not in building models but in identifying which features truly matter. Effective feature selection is essential to ensure that models are not only accurate but also efficient and interpretable. This study investigates the use of Genetic Algorithm (GA) as a method for feature selection, with the aim of optimizing the performance of K-Nearest Neighbors and Decision Tree models for both regression and classification tasks. GA is chosen for its ability to efficiently explore large search spaces and identify optimal feature subsets, minimizing the risk of getting trapped in local minima compared to traditional methods. Applied to the *California Housing Prices*, *Student Performance*, and *Student Dropout and Academic Success* datasets obtained from the UC Irvine Machine Learning Repository, GA proved successful in minimizing the number of features while enhancing model performance. The results indicated that GA reduced the Mean Squared Error by up to 60.84% and improved classification accuracy by over 11.52%, with feature sets reduced by more than 50%. However, the study was limited to only two machine learning algorithms, which restricts the generalizability of the findings, as the effectiveness of GA for feature selection could vary with different algorithms. Nonetheless, the findings demonstrate GA's potential to create more streamlined and comprehensible models through effective feature selection for diverse tasks. Future research could extend this approach to a broader range of models and datasets.**

***Index Terms*—Feature Selection, Genetic Algorithm, Model Optimization, Regression and Classification, Model Performance**

I. Introduction

As artificial intelligence (AI) continues to advance, the line between human intelligence and machine capability is becoming increasingly blurred. AI is now undertaking tasks that were once considered the exclusive domain of human expertise [1]. Central to this shift is the growth of machine learning, a branch of AI that allows algorithms to autonomously extract knowledge and insights from data. What makes this transformation truly remarkable is that machines are no longer limited to simply following predetermined commands; instead, they are beginning to demonstrate abilities akin to reasoning and decision-making [2].

However, for any machine learning model to perform tasks accurately, it must be able to focus on the most relevant information within a dataset [3]. This is where feature selection becomes essential. It enables the model to identify and prioritize the key variables in a dataset, while filtering out unnecessary attributes that might otherwise reduce model performance [4]. Yet, despite its importance, traditional methods of feature selection often fall short when faced with complex datasets.

Methods such as filter, wrapper, and embedded approaches rely on relatively straightforward techniques like correlation or information gain to rank feature importance, and thus tend to struggle when datasets involve complex relationships between features [5]. The intricacies of such datasets require more sophisticated methods, and Genetic Algorithm (GA) effectively addresses this need. Inspired by the process of natural selection, GA is uniquely suited to solving optimization problems, particularly those with large search spaces, like feature selection.

GA evolves a population of potential feature subsets over multiple generations, with each subset being evaluated for its fitness to improve model performance [6]. This evolutionary approach allows GA to explore a vast range of potential solutions, gradually zeroing in on the most effective feature subsets for the given task [7].

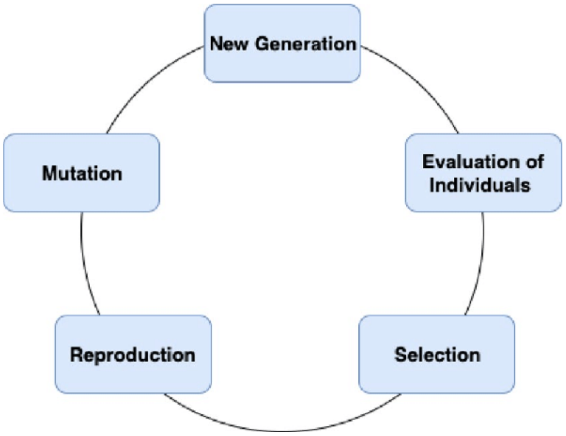

Fig. 1. The process of genetic algorithm. Adapted from [8].

The aim of this experiment is to contribute to the growing body of evidence on whether GA-based feature selection can enhance the accuracy of prediction and classification models, specifically for K-Nearest Neighbors (KNN) and Decision Trees (DT), across various real-world datasets sourced from

UCI Machine Learning Repository. KNN is selected for its simplicity and effectiveness in handling both regression and classification tasks, as well as its reliance on distance metrics, which can be significantly influenced by feature selection. DT, on the other hand, is chosen for their interpretability and ability to model complex nonlinear relationships without requiring extensive preprocessing. Together, these algorithms allow for a comprehensive evaluation of GA's feature selection efficacy across different modeling paradigms.

For the regression tasks, the *California Housing Prices* dataset is used to predict house prices. The performance of the KNN and DT is evaluated using Mean Squared Error (MSE), comparing results both with and without GA-driven feature selection. In a second test, the *Student Performance* dataset, containing a mix of numerical and categorical attributes, is used to predict student grades. As with the housing prices task, the KNN and DT models are assessed using MSE, allowing for a direct comparison of their performance before and after the application of GA-based feature selection. Finally, to explore the effects of GA in classification tasks, the *Student Dropout and Academic Success* dataset is used. In this case, the models are evaluated based on their classification accuracy, again comparing results with and without GA-influenced feature selection.

By examining these two models in such diverse contexts, the investigation seeks to establish the broader utility of GA in improving machine learning outcomes, particularly in scenarios where traditional feature selection methods may not be sufficient. Through this, the research hopes to highlight the versatility and effectiveness of GA as a valuable tool for feature selection in modern machine learning pipelines.

II. Literature Review

When it comes to balancing simplicity and effectiveness in machine learning, algorithms like KNN and DT offer straightforward yet powerful solutions [9]. The former relies on the proximity of data points to make predictions, while the latter guides decisions down clear, logical paths [10]. Their versatility allows them to handle a wide array of tasks, from classification to regression, making them attractive options for many data scientists. In spite of that, as datasets grow more complex, both approaches start to reveal significant limitations [11].

Mewada and Patil highlighted the challenges faced by KNN when dealing with extensive feature sets. They found that as the number of variables in a dataset increases, the performance of the algorithm deteriorates. In high-dimensional datasets, this method struggles to separate relevant patterns from irrelevant noise, which results in poor outcomes. [12]. DT suffer in these conditions too. Hasan found that they are particularly sensitive to irrelevant columns in the dataset, which not only reduces accuracy but also increases the risk of overfitting, making them less reliable for new data [13]. In fact, Magdalene and Sridharan demonstrated this when both approaches performed poorly on multi-dimensional thyroid datasets [14].

Building on these insights, Tajanpure and Muddana showed that reducing the dimensionality of datasets led to a notable improvement in the accuracy of both KNN and DT models [15]. Taken together, these findings highlight that while these algorithms are excellent in their own right, they struggle when faced with large amounts of data that include many variables. Hence, feature selection becomes crucial to mitigate these issues. However, finding the right feature selection method to select the most relevant variables has proven challenging. Ganesh and colleague explored an innovative approach called the Weighted Superposition Attraction Optimization Algorithm for KNN. It outperformed other techniques, but when applied to large datasets, the algorithm hit a wall with its efficiency, slowing down training times considerably [16]. Similarly, Wang and coworkers tried a randomized wrapper-based feature selection for DT but encountered major scalability issues [17].

Even attempts to incorporate hybrid and filter-based methods have seen mixed results. Singh and Selvakumar experimented with reducing dimensionality through a combination of filters and hybrid systems but found the results to be inconsistent across different datasets. Although feature reduction was achieved, overfitting and feature redundancy still persisted, indicating that the solution was far from perfect [18]. Kalaivani and Shunmuganathan tried Principal Component Analysis (PCA) to shrink the feature space in sentiment analysis models, but PCA ended up discarding valuable information, leading to a drop in accuracy for both KNN and DT [19].

These limitations have led many to explore more advanced methods, and one approach stands out: genetic algorithm (GA). GA mimics the natural selection process, optimizing feature subsets by evolving solutions over time. Unlike traditional methods, which often struggle with the scale and complexity of large datasets, GA seems to thrive in these environments [6]. Taradeh and colleagues compared GA with the Gravitational Search Algorithm and found that GA not only converged more quickly but also generalized better across a variety of datasets [20].

The strength of GA becomes even clearer when looking at specific comparisons. Rao and fellow researchers found that GA consistently outperformed Binary Chemical Reaction Optimization in feature selection tasks for KNN and DT. In their study, GA-optimized models showed a 4.58% boost in accuracy while simultaneously reducing the number of features needed [21]. Similarly, Kalaivani and Shunmuganathan demonstrated that combining GA with information gain resulted in up to an 87.50% improvement in accuracy for sentiment analysis tasks [19]. This illustrates the robustness of GA, as it seems to perform well across various tasks, including more complex language models.

Furthermore, various studies demonstrate that GA is not only powerful in high-dimensional datasets but also applicable across fields as varied as finance, social media, and beyond. For instance, Jadhav and colleagues applied GA-enhanced wrappers for credit scoring, outperforming more complex models like Naive Bayes and Support Vector Machines [22]. Meanwhile, Singh and Sood used GA-optimized KNN and Random Forest models for spambot detection on Twitter, where they found superior accuracy compared to traditional feature selection techniques like SVM-RFE [23].

Finally, one of the most fascinating applications comes from Yu and colleagues, who used GA to optimize feature selection for fuzzy KNN classifiers in hyperspectral satellite imagery. This allowed them to dramatically improve both classification accuracy and computational efficiency, which is a crucial advancement for working with real-world and high-dimensional data [24]. Hansen and his team expanded the use of GA even further by developing a method called GenForest, which applies GA-based feature selection to Random Forest models in biological data analysis, demonstrating the far-reaching potential of GA across different scientific fields [25].

In summary, both KNN and DT face significant hurdles when applied to high-dimensional datasets without feature selection. However, traditional methods often fall short of achieving optimal feature selection. In contrast, GA have proven to be an effective and powerful alternative. Building on these insights, this study aims to advance the field by employing GA for feature selection in both regression and classification tasks using KNN and Decision Tree models. By evaluating the performance of these models after GA-driven feature selection, this research seeks to provide deeper insights into the practical applications and benefits of GA in optimizing predictive models for high-dimensional datasets.

III. Methodology

The experimental setup consisted of two main phases aimed at improving model performance through feature selection. In the first phase, KNN (with k = 5) and DT were applied to both regression and classification tasks. These two models were selected for their contrasting approaches: KNN is a distance-based method, while DT is rule-based. This combination provided diverse insights into the feature selection process. Other models were not included to maintain focus on these interpretable algorithms. The goal of this phase was to establish a performance benchmark with all available features.

The second phase involved the use of GA to optimize feature selection. The goal of this phase was to identify the most relevant features for each model and improve their performance. Model evaluation varied depending on the task. For regression, Mean Squared Error was used to assess prediction accuracy. For classification, accuracy was the key metric. This approach enabled direct comparisons between models trained with the full feature set and those using optimized subsets.

Within the GA process, each potential solution was encoded as a binary vector, where each element indicated whether a specific feature was selected (1) or excluded (0). Model performance on the selected subset determined the fitness of each solution, with prediction error serving as the fitness criterion in regression tasks, and classification success rate used in classification tasks. The algorithm featured a population of 100 individuals, with tournament selection (group size of 3), a 50% crossover probability, and a 20% mutation rate. It ran until 20 generations passed without further improvement, balancing exploration and convergence.

The datasets used for evaluation were sourced from the UCI Machine Learning Repository. For the first regression task, the *California Housing Prices* dataset (20,640 instances, 8 features) was utilized because all of its variables are numerical and provides a straightforward feature space. To introduce a mix of categorical and numerical variables, the *Student Performance* data (396 instances, 43 features) was used for the second regression experiment. For the classification task, the *Student Dropout and Academic Success* dataset (4425 instances, 36 features) was leveraged.

In conclusion, the outlined methodology provides a comprehensive structure for assessing the role of GA in feature selection for improving the performance of KNN and DT models.

IV. Experiments

A. *Experiment 1*

The first experiment assessed the extent to which GA-derived feature selection affected the predictive performance of KNN and DT models for housing prices. The *California Housing Prices* dataset was used for this regression task. First of all, the dataset was loaded and converted into a Pandas DataFrame for easier manipulation. Afterwards, it was split into features and target variables. All columns were used as features, excluding the target variable, median house value. The features were then standardized using StandardScaler from sklearn to ensure that the models could effectively interpret the data. Next, the dataset was split into training (70%) and testing (30%) sets, with a random seed set to ensure the reproducibility of the results.

Subsequently, both KNN and DT models were trained using the full feature set and their performance was evaluated using MSE. Later on, GA was implemented to select the most relevant features for each model. After running the GA, the performance of the models before and after feature selection was compared to gauge the impact of GA on the predictive capabilities. In addition, the features selected by the GA were noted for each model to identify which features contributed the most to prediction performance. Furthermore, the frequency of selected features was also recorded.

B. *Experiment 2*

In the second experiment, the same GA-tailored feature selection approach was applied to predict student performance, again using KNN and DT models. For this task, the *Student Performance* dataset was utilized. The dataset was first loaded into a Pandas DataFrame. Since it contained a mixture of categorical and numerical data, appropriate preprocessing steps were applied next. Categorical features with only two unique categories, such as ”sex” (male or female), were transformed using label encoding. For categorical features with more than two unique categories, such as ”school” (representing different institutions), one-hot encoding was applied.

Once the categorical features were encoded, the dataset was split into features and target variables. The features consisted of all columns except the target variable, G3, which represented the final grade of students. The dataset was then divided into training (70%) and testing (30%) sets, with a random state to ensure that the results could be replicated.

Similar to Experiment 1, the KNN and DT models were initially trained on the complete feature set and their performance was evaluated using MSE. Ultimately, GA was used to determine the most relevant features for each model. After executing the GA, the models' performance was assessed before and after feature selection to understand the GA's impact on predictive accuracy. As was done in previous experiment, attention was also given to identifying the specific features selected by the GA and their selection frequency.

C. *Experiment 3*

The third experiment also evaluated the impact of GA-enabled feature selection on KNN and DT models, this time within a classification task. The *Student Dropout and Academic Success* dataset was used to classify students based on their likelihood of dropping out, succeeding, or graduating. The dataset was first loaded into a Pandas DataFrame, and a subsequent examination of it indicated that the target variable comprised categorical values. Thus, to prepare this column for the models, label encoding was applied, converting the categorical outcomes into numerical values.

Once the encoding was completed, the dataset was split into features and the target variable. The features consisted of all columns except the target variable, while the target was the encoded column representing student outcomes. The features were standardized using StandardScaler from sklearn to enable the models to interpret the data effectively. The dataset was then divided into training (70%) and testing (30%) sets, with a random state applied for reproducibility.

As in the previous two experiments, both KNN and DT classifiers were at first trained on the full feature set and their performance was assessed using classification accuracy. The next step involved using GA to select the most significant features for each model. After running the GA, the performance of both models was compared before and after feature selection to determine the GA's impact on classification accuracy. Just as the prior experiments, the selected features for each model were identified. Moreover, the frequency of selected feature was also monitored.

V. Results

A. *Experiment 1*

For the baseline models trained on the full feature set of 8 features in experiment 1, the KNN model produced an MSE of 0.423345, while the DT model had an MSE of 0.531282. Once the GA was applied to select the most relevant subset of features, both models demonstrated a significant reduction in MSE. The KNN model's MSE decreased to 0.282219, representing a 33.36% reduction in error, while the DT model's MSE decreased to 0.376411, reflecting a 29.15% reduction in error.

TABLE I
Experiment 1 Results Summary

Model	MSE with All Features	MSE with GA-Selected Features	Improvement (%)
KNN	0.423345	0.282219	33.36%
Decision Tree	0.531282	0.376411	29.15%

To further illustrate the performance improvement, the following bar chart compares the MSE values for both models before and after GA-engineered feature selection.

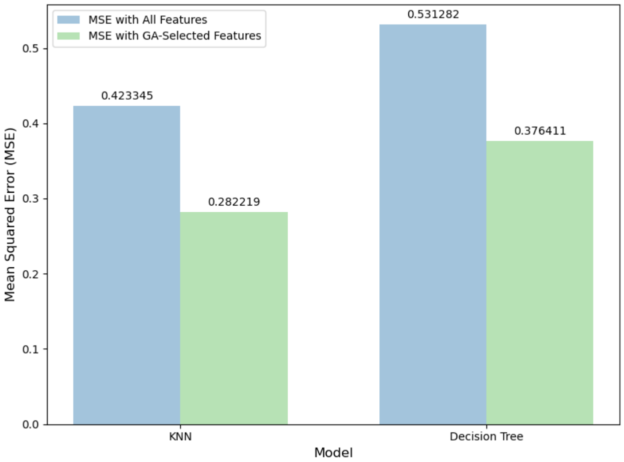

Fig. 2. Experiment 1 MSE Comparison

In this experiment, GA identified two features, Latitude and Longitude, as the most influential in predicting housing prices. Both models selected only these two columns, discarding the other six features, which included MedInc, HouseAge, AveRooms, AveBedrms, Population, and AveOccup. This reduction from eight features to two represents a 75% decrease in the number of features used for model training.

This underscores the critical insight that not all features contribute equally to predictive accuracy. By identifying and retaining only the most influential features, practitioners can avoid the pitfalls of overfitting and model complexity that can arise from utilizing a larger set of variables. The experiment illustrates how GA can serve as a powerful tool for optimizing feature selection, ultimately leading to more effective and robust models, especially in spatial datasets.

B. Experiment 2

For the baseline models trained on the full feature set of 43 features in experiment 2, the KNN model produced an MSE of 19.972773, while the Decision Tree model had a considerably higher MSE of 32.344538. Once the GA was applied to select the most relevant subset of features, both models demonstrated significant improvements. The KNN model's MSE decreased to 14.351261, representing a 28.13% reduction in error, while the Decision Tree model's MSE decreased to 12.659664, reflecting a remarkable 60.84% reduction in error.

TABLE II
EXPERIMENT 2 RESULTS SUMMARY

Model	MSE with All Features	MSE with GA-Selected Features	Improvement (%)
KNN	19.972773	14.351261	28.13%
Decision Tree	32.344538	12.659664	60.84%

To further illustrate the performance improvement, the following bar chart compares the MSE values for both models before and after GA-based feature selection.

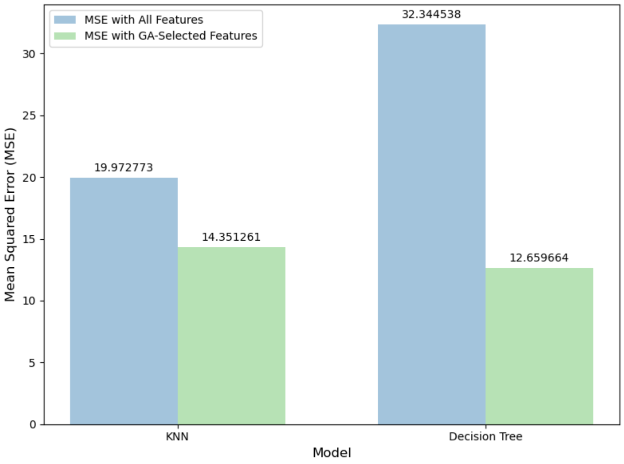

Fig. 3. Experiment 2 MSE Comparison

In this experiment, GA identified a significantly reduced subset of features from the original 43 features: the KNN model selected 23 features, while the DT model selected 24 features. This means that the GA reduced the feature space by approximately 46.51% for the KNN model and 44.19% for the DT model, cutting the feature space by almost half.

Additionally, the frequency analysis of feature selection revealed interesting patterns. Features such as ”sex” (gender), ”Pstatus” (parental cohabitation status), ”failures” (number of past class failures), ”schoolsup” (extra educational support), ”Walc” (weekend alcohol consumption), and ”absences” (number of school absences) were selected by both models, suggesting their high relevance for predicting student performance. On the other hand, features like ”Fedu” (father's level of education), ”higher” (intent to pursue higher education), ”romantic” (in a romantic relationship), and ”famrel” (quality of family relationships) were consistently not selected, indicating their lower relevance for both models.

Overall, the application of GA in this context illustrates their effectiveness in optimizing feature selection, leading to models that are not only more accurate but also more interpretable. These findings contribute to the growing body of knowledge in educational data analytics and suggest that focused feature selection can lead to actionable insights and improved outcomes in educational settings.

C. Experiment 3

For the baseline models trained on the full feature set of 36 features in experiment 3, the KNN model achieved an accuracy of 0.691265, while the DT model achieved an accuracy of 0.663404. Once the GA optimization was applied to identify the most relevant subset of features, both models exhibited notable improvements in classification accuracy. The KNN model's accuracy increased to 0.762048, representing a 10.23% improvement, while the DT model's accuracy increased to 0.736446, reflecting an 11.52% improvement.

TABLE III
EXPERIMENT 3 RESULTS SUMMARY

Model	Accuracy with All Features	Accuracy with GA-Selected Features	Improvement (%)
KNN	0.691265	0.762048	10.23%
Decision Tree	0.663404	0.736446	11.52%

To further illustrate the performance enhancement, the following bar chart compares the classification accuracy of both models before and after GA-based feature selection.

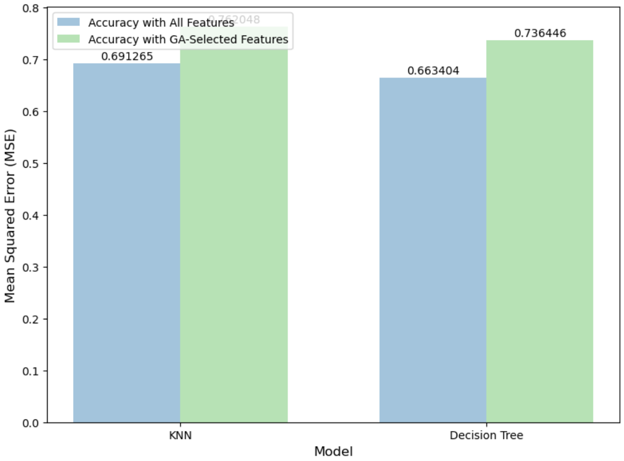

Fig. 4. Experiment 3 Accuracy Comparison

In this experiment, GA identified a significantly reduced subset of features from the original 36 features: the KNN model selected 14 features, reducing the feature set by approximately 61.11%, while the Decision Tree model selected 18 features, reducing the feature set by 50.00%.

Furthermore, the frequency analysis of feature selection revealed several noteworthy patterns. Features such as ”Course”, ”Tuition fees up to date”, ”Curricular units 1st sem (credited)”, and ”Curricular units 2nd sem (approved)” were consistently selected by both models, indicating their high relevance in predicting student outcomes. In contrast, features such as ”Marital status”, ”Application mode”, and ”Previous qualification” were consistently excluded by both models, suggesting their minimal contribution to the predictive power of the models.

Overall, these findings reinforce the notion that employing advanced feature selection techniques is vital for improving decision-making capabilities in educational settings. As educational institutions increasingly rely on data-driven insights to guide policy and practice, the results from this experiment serve as a valuable resource for educators and policymakers.

VI. Conclusion, Limitations, and Future Work

This research conducted three experiments to investigate the impact of feature selection using GA on the performance of KNN and DT models in both regression and classification scenarios. In the first experiment, which involved predicting housing prices using the *California Housing Prices* dataset, both KNN and DT models saw substantial reductions in MSE after GA-based feature selection. The models reduced their MSE by 33.36% and 29.15%, respectively, by selecting just two features, Latitude and Longitude.

In the second experiment on the *Student Performance* dataset, GA once again significantly enhanced the performance of both models by selecting 24 features from the original 43. The DT model experienced the greatest improvement, with a 60.84% reduction in MSE, while the KNN model saw a 28.13% improvement.

The third experiment, which focused on classification based on the *Student Dropout and Academic Success* dataset, further demonstrated the power of GA-influenced feature selection. The KNN and Decision Tree models improved their classification accuracy by 10.23% and 11.52%, respectively, after reducing the feature set by over 50%.

Although this research shows promising results, several limitations inherent to the study itself must be considered. Firstly, no hyperparameter tuning was conducted for any of the models. For instance, the KNN model was implemented using a fixed value of k=5, while the DT was applied with their default parameters. Even for GA, only commonly used parameter values were applied without exploring potentially more effective configurations. This lack of tuning could impact the overall performance and the observed benefits of GA-enhanced feature selection.

Furthermore, this research focused on only two specific models across regression and classification tasks. Hence, the findings may not generalize to other types of models, such as ensemble methods or deep learning models, which may respond differently to feature selection techniques.

Future studies could explore how GA-powered feature selection performs with a broader range of models to assess its applicability across diverse machine learning algorithms. Moreover, future research could also explore how GA-assisted feature selection can be integrated with automated machine learning (AutoML) systems. By including feature selection as part of the AutoML pipeline, researchers and practitioners could optimize both feature subsets and model parameters in an automated fashion.

Overall, the findings from this research underscore the importance of feature selection in machine learning. GA-aided feature selection has proved to not only improve model performance but also reduce the complexity of the feature space, making models more efficient and interpretable. By focusing on the most relevant features, models can achieve higher accuracy and make more informed predictions, particularly in tasks involving large, mixed datasets like the ones used in this study. These results highlight the potential of GA-driven optimization as a valuable tool for enhancing machine learning models across a range of predictive tasks.

References

[1] J. E. H. Korteling, G. C. van de Boer-Visschedijk, R. a. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, “Human- versus Artificial Intelligence,” *Frontiers in Artificial Intelligence*, vol. 4, Mar. 2021, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.622364/full

[2] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, p. 160, Mar. 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00592-x

[3] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, “Feature selection: A data perspective,” *ACM computing surveys (CSUR)*, vol. 50, no. 6, pp. 1–45, 2017.

[4] V. Bolón-Canedo, A. Alonso-Betanzos, L. Morán-Fernández, and B. Cancela, “Feature Selection: From the Past to the Future,” in *Advances in Selected Artificial Intelligence Areas: World Outstanding Women in Artificial Intelligence*, M. Virvou, G. A. Tsihrintzis, and L. C. Jain, Eds. Cham: Springer International Publishing, 2022, pp. 11–34. [Online]. Available: https://doi.org/10.1007/978-3-030-93052-3_2

[5] D. Theng and K. K. Bhoyar, “Feature selection techniques for machine learning: a survey of more than two decades of research,” *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575–1637, Mar. 2024. [Online]. Available: https://doi.org/10.1007/s10115-023-02010-5

[6] Z. Y. Taha, A. A. Abdullah, and T. A. Rashid, “Optimizing Feature Selection with Genetic Algorithms: A Review of Methods and Applications,” Sep. 2024, arXiv:2409.14563 [cs]. [Online]. Available: http://arxiv.org/abs/2409.14563

[7] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, “Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach,” *Information*, vol. 10, no. 12, p. 390, Dec. 2019, number: 12 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2078-2489/10/12/390

[8] S. Bourhnane, M. R. Abid, R. Lghoul, K. Zine-dine, N. EL KAMOUN, and D. Benhaddou, “Machine learning for energy consumption prediction and scheduling in smart buildings,” *SN Applied Sciences*, vol. 2, Feb. 2020.

[9] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, “Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications,” *Journal of Big Data*, vol. 11, no. 1, p. 113, Aug. 2024. [Online]. Available: https://doi.org/10.1186/s40537-024-00973-y

[10] S. D. Jadhav and H. Channe, “Comparative study of k-nn, naive bayes and decision tree classification techniques,” *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, pp. 1842–1845, 2016.

[11] N. Ukey, Z. Yang, B. Li, G. Zhang, Y. Hu, and W. Zhang, “Survey on Exact kNN Queries over High-Dimensional Data Space,” *Sensors*, vol. 23, no. 2, p. 629, Jan. 2023, number: 2 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/2/629

[12] P. Mewada and J. Patil, “Performance Analysis of k-NN on High Dimensional Datasets,” *International Journal of Computer Applications*, vol. 16, no. 2, pp. 1–5, Feb. 2011. [Online]. Available: http://www.ijcaonline.org/volume16/number2/pxc3872678.pdf

[13] A. Hasan, “Evaluation of Decision Tree Classifiers and Boosting Algorithm for Classifying High Dimensional Cancer Datasets,” *International Journal of Modeling and Optimization*, pp. 92–96, 2012. [Online]. Available: http://www.ijmo.org/show-30-77-1.html

[14] R. Magdalene and D. Sridharan, “Classification of multi-dimensional thyroid dataset using data mining techniques: comparison study,” *Advances in Natural and Applied Sciences*, vol. 9, pp. 24–29, 2015. [Online]. Available: https://www.semanticscholar.org/paper/Classification-of-multi-dimensional-thyroid-dataset-Magdalene-Sridharan/7a0ba7352c73b62dc0f6a80cd6295691797a4bbe

[15] R. Tajanpure and A. Muddana, “Circular convolution-based feature extraction algorithm for classification of high-dimensional datasets,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 1026–1039, Jan. 2021, publisher: De Gruyter. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jisys-2020-0064/html

[16] N. Ganesh, R. Shankar, R. Čep, S. Chakraborty, and K. Kalita, “Efficient Feature Selection Using Weighted Superposition Attraction Optimization Algorithm,” *Applied Sciences*, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/5/3223

[17] Z. Wang, X. Xiao, and S. Rajasekaran, “Novel and efficient randomized algorithms for feature selection,” *Big Data Min. Anal.*, vol. 3, pp. 208–224, 2020. [Online]. Available: https://consensus.app/papers/novel-randomized-feature-selection-wang/6cb8c4cdf4ac5cb589d1c88de0ca1fda/

[18] S. Singh and S. Selvakumar, “A hybrid feature subset selection by combining filters and genetic algorithm,” in *Communication & Automation International Conference on Computing*, May 2015, pp. 283–289. [Online]. Available: https://ieeexplore.ieee.org/document/7148389

[19] P. Kalaivani and K. L. Shunmuganathan, “Feature Reduction Based on Genetic Algorithm and Hybrid Model for Opinion Mining,” *Scientific Programming*, vol. 2015, no. 1, p. 961454, 2015, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2015/961454. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2015/961454

[20] M. Taradeh, M. M. Mafarja, A. A. Heidari, H. Faris, I. Aljarah, S. Mirjalili, and H. Fujita, “An evolutionary gravitational search-based feature selection,” *Inf. Sci.*, vol. 497, pp. 219–239, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0020025519304414?via%3Dihub

[21] P. Rao, A. Kumar, Q. Niyaz, P. Sidike, and V. Devabhaktuni, “Binary chemical reaction optimization based feature selection techniques for machine learning classification problems,” *Expert Syst. Appl.*, vol. 167, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417420309076?via%3Dihub

[22] S. Jadhav, H. He, and K. Jenkins, “Information gain directed genetic algorithm wrapper feature selection for credit rating,” *Appl. Soft Comput.*, vol. 69, pp. 541–553, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1568494618302242?via%3Dihub

[23] A. Singh and M. Sood, “Feature Selection Optimization Using Genetic Algorithm for Spambot Detection in an OSN,” pp. 334–345, 2019. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-4451-4_26

[24] S. Yu, S. D. Backer, and P. Scheunders, “Genetic feature selection combined with fuzzy kNN for hyperspectral satellite imagery,” *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, vol. 2, pp. 702–7042, 2000. [Online]. Available: https://ieeexplore.ieee.org/document/861676

[25] L. Hansen, E. A. Lee, K. Hestir, L. Williams, and D. Farrelly, “Controlling feature selection in random forests of decision trees using a genetic algorithm: classification of class I MHC peptides.” *Combinatorial chemistry & high throughput screening*, vol. 12 5, pp. 514–9, 2009. [Online]. Available: https://www.eurekaselect.com/article/14355