

Optimisation of Feature Selection Using Stochastic Diffusion Search and Machine Learning Models

Sithembiso Dyubele

Article information

Affiliations

Durban University of Technology, Department of Information Systems, Durban, South Africa

Corresponding author

Sithembiso Dyubele
ctheradyubele@gmail.com

Dates

Received: 13th March 2025
Accepted: 1st November 2025
Published: 22nd December 2025

Digital identifier



<https://doi.org/10.70727/3134-6332/2025/v1i1a2>

Copy right

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.



Abstract

The ever-growing field of Computational Intelligence (CI) offers an alternative to researchers seeking to solve complex and real-world problems. The reviewed literature revealed that the increasing dimensionality of datasets poses a significant challenge, which can severely impact the performance of predictive models by introducing noise and complexity, leading to overfitting and increased computational demands. This study presents an in-depth analysis and optimisation of feature selection through Stochastic Diffusion Search (SDS), targeting the curse of dimensionality in machine learning applications. Before applying the Stochastic Diffusion Search (SDS) algorithm, the dataset underwent an initial round of performance evaluation using various Machine Learning models. These ML models include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN), Gradient Boosting (GB), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). The initial assessment provided a baseline against which the efficacy of the SDS feature selection could be measured. The effectiveness of SDS in enhancing model performance is empirically evaluated using the sonar signal dataset. The results are illustrated in the Tables; however, the key results indicate a uniform improvement across various machine learning models when features are selected via SDS, with the Gradient Boosting Classifier showing the most significant enhancement in accuracy, from 86.54% to 90.38%. This uplift highlights the capability of the SDS to distil the most relevant features, thereby boosting predictive power while potentially reducing overfitting and computation time. These findings underscore the critical role of optimised feature selection in building efficient and robust predictive models, especially in high-dimensional data landscapes. The study concludes with insights into the balance between model accuracy and computational efficiency, providing a benchmark for further research in feature selection methodologies.

Keywords: Stochastic Diffusion Search, Computational Intelligence (CI), Machine Learning

Introduction

Habehh and Gohel (2021) revealed that the advent of machine learning has ushered in an era of data-driven decision-making across various domains, from healthcare to finance. However, the increasing dimensionality of datasets poses a challenge, commonly referred to as the "curse of dimensionality," which can severely impact the performance of predictive models by introducing noise and complexity, leading to overfitting and increased computational demand (Habehh and Gohel, 2021). Shanthi and Rajkumar (2021) claim that Stochastic Diffusion Search (SDS) is an adaptive, multi-agent global search and optimisation procedure that has shown promise in addressing this challenge through effective feature selection. This study focuses on the application of SDS to a sonar signal dataset. The dataset comprises patterns obtained from sonar signals bounced off different objects and serves as a benchmark for feature selection and classification algorithms in the field of machine learning. The objective of the experiments conducted is two-fold: first, to ascertain the impact of SDS on the accuracy and computational efficiency of a range of machine learning models, and second, to compare the performance of these models before and after the application of SDS-based feature selection. The insights gained from these experiments are expected to contribute to the broader discourse on feature selection strategies.

Theoretical Background

Maroufpoor, Azadnia and Bozorg-Haddad (2020) described Stochastic Diffusion Search (SDS) as a probabilistic search algorithm that belongs to the family of swarm intelligence and global optimisation algorithms. Similarly, Martin (2021) stated that the origins of SDS can be traced back to the broader concept of diffusion processes, which describe the movement of particles across a gradient, often from an area of higher concentration to an area of lower concentration. In the context of SDS, the "particles" are analogous to agents performing a search across the solution space.

Algorithm Overview

Suganya and Sountharajan (2022) indicated that the foundation of SDS is its multi-agent search approach, where each agent maintains a hypothesis, which is a potential solution to the optimisation problem. The algorithm proceeds in iterations, each consisting of three phases: hypothesis selection, test, and diffusion.

- **Hypothesis Selection**

Each agent randomly selects a hypothesis from the search space, potentially guided by a probability distribution reflecting prior knowledge or heuristics. Mathematically, this can be represented as:

$$H_i(t) \sim \mathcal{P}(X), \quad (1)$$

where $H_i(t)$ is the hypothesis held by agent i at time t , and $\mathcal{P}(X)$ is the probability distribution over the search space X .

- **Test Phase**

Agents evaluate their hypotheses to determine their quality. Shanthi, Akshaya, Smitha, and Bommy (2022) indicated that the test function could be a binary function indicating success or failure or a more complex function providing a continuous measure of hypothesis quality:

$$s_i(t) = f(H_i(t)), \quad (2)$$

Where $s_i(t)$ is the score or fitness of the hypothesis $H_i(t)$, as evaluated by the test function f .

- ***Diffusion Phase***

Depending on the outcome of the test phase, agents communicate with one another and potentially adopt the hypotheses of other, more successful agents (Morande, 2023). This is akin to particles diffusing in a medium, where the concentration gradient drives the process. Krylov (2021) stated that the diffusion can be modelled as a Markov process:

$$H_i(t+1) = \begin{cases} H_j(t) & \text{if } s_j(t) > s_i(t), \\ H_i(t) & \text{otherwise} \end{cases} \quad (3)$$

where agent i adopts the hypothesis of agent j if the latter's score is higher.

- ***SDS as a Global Optimiser***

According to Muthukkumar and Diwakaran (2023), SDS is particularly effective in global optimisation problems due to its inherent parallelism and robustness against becoming trapped in local optima. Similarly, Salimi and Salimi (2021) revealed that the global search capability of SDS can be characterised by the proportion of agents converging towards a global optimum, represented by the equation:

$$\lim_{t \rightarrow \infty} P\left(\bigcap_{i=1}^n H_i(t) = H^*\right) = 1, \quad (4)$$

where H^* is the global optimum, and n is the number of agents.

Theoretical Implications

Vineetha, Maharajan, Bhagyashree, and Sivakumar (2024) revealed that the strength of SDS lies in its simplicity and the theoretical guarantees it offers. Under reasonable conditions, such as the continuity of the test function and proper calibration of agent communication protocols, SDS can be shown to converge to the global optimum with high probability (Karthik, Tamizhazhagan, and Narayana, 2021). The probabilistic nature of SDS also lends itself to theoretical analysis using tools from statistical physics and information theory, providing a rich mathematical framework for understanding and predicting the behaviour of the algorithm in various scenarios (Karuppusmay, Shi, and Shi, 2021). In conclusion, SDS is a mathematically elegant approach to the problem of global optimisation, drawing from deep theoretical roots in probability theory and statistical mechanics. The following section will elaborate on the implementation of SDS in the context of feature selection for machine learning models.

Methodology

Data Description

The dataset employed in this study is derived from the sonar, mines vs. rocks dataset, which is a standard reference in machine learning for signal pattern recognition. The dataset consists of 208 instances, each representing a pattern obtained by bouncing sonar signals off a metal cylinder (mine) or roughly cylindrical rock. Each instance comprises 60 attributes, representing the energy within a particular frequency band integrated over a specific period. These attributes are numerical and normalised to fit

within the range of 0 to 1. The dataset features are labelled as $S_1, S_2, \dots, S_{60}$, where each S_i corresponds to the strength of the returned signal within the i^{th} frequency band. The target variable is binary, indicating the object type: mine (metal cylinder) or rock, encoded as 'M' or 'R', respectively. The task associated with this dataset is a binary classification problem, intending to accurately predict whether the sonar signal pattern corresponds to a mine or a rock based on the signal return.

- ***Characteristics of the Dataset***

The dataset presents a challenging problem due to its noisy nature and the similarity of signal returns in specific frequency bands for both mines and rocks. It poses the curse of dimensionality, as the number of features relative to the number of instances is relatively high, increasing the search space exponentially and potentially complicating the learning process for any machine learning model. The distribution of instances across the two classes is relatively balanced, with 111 patterns for mines and 97 for rocks, which helps mitigate the class imbalance issue that could otherwise skew the performance of a predictive model. This dataset was sourced from the UCI Machine Learning Repository, a well-known repository for machine learning datasets. The choice of this dataset for the study is motivated by its widespread use in feature selection research, providing a benchmark to assess the efficacy of the SDS algorithm in a controlled and comparative manner. The following section will delve into the specific methodologies employed to implement the SDS algorithm for feature selection, as well as the configurations of the various machine learning models used for evaluating the impact of the feature selection process.

- ***Feature Selection Algorithm***

The feature selection process for this study was carried out using the Stochastic Diffusion Search (SDS), an agent-based probabilistic global search and optimisation algorithm. This is corroborated by Martin (2021), for whom, in the context of feature selection, each agent in SDS represents a possible subset of features from the dataset, and the optimisation process aims to find the subset that maximises the performance of the machine learning models. The proposed feature engineering (FE) framework in this study is illustrated in Figure 1.

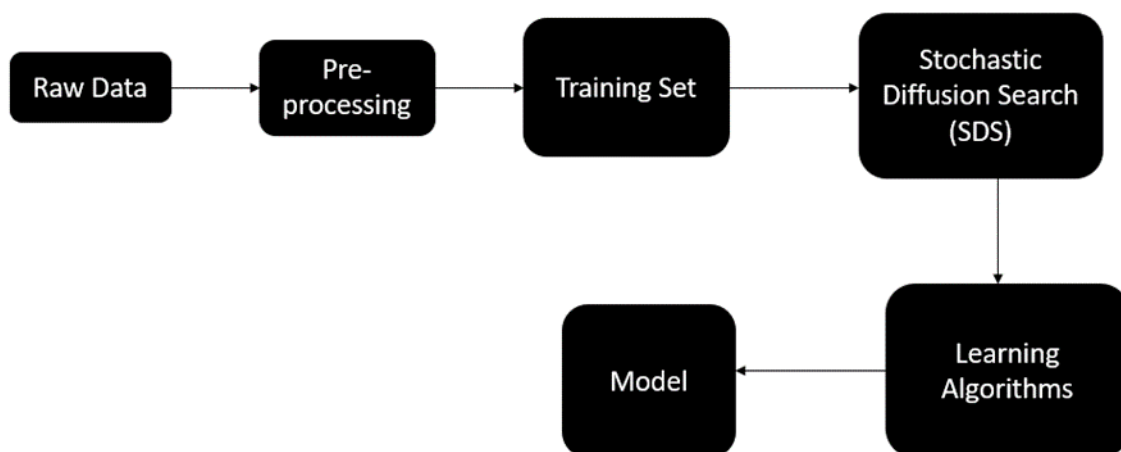


Figure 1: A proposed framework for feature selection (FS) in this study.

- ***Initialisation of Agents***

Alhakbani and Al-Rifaie (2017) state that the SDS algorithm begins with the initialisation of agents. Each agent is initialised with a random subset of features (Alhakbani and Al-Rifaie, 2017). Mathematically, the initial hypothesis $H_i(0)$ for each agent i is represented as a binary vector where each element corresponds to the inclusion or exclusion of a feature in the subset:

$$\begin{aligned} H_i(0) \\ = \{h_{i1}, h_{i2}, \dots, h_{in}\}, \quad h_{ij} \in \{0, 1\}, \end{aligned} \quad (5)$$

where n is the total number of features and $h_{ij} = 1$ if the i^{th} feature is included in the i^{th} agent's subset, and 0 otherwise.

- **Test Phase**

In the test phase, Shanthi and Rajkumar (2021) revealed that agents assess the quality of their selected features by training a machine learning model on their feature subset and evaluating its performance on a validation set. The performance metric used in this study is the classification accuracy. The score or fitness $s_i(t)$ for each agent's hypothesis at iteration t is given by:

$$\begin{aligned} s_i(t) \\ = \text{accuracy}(\text{model}_i(H_i(t))), \end{aligned} \quad (6)$$

Where $\text{model}_i(H_i(t))$ is the machine learning model trained using the features selected by agent i 's hypothesis $H_i(t)$.

- **Diffusion Phase**

After the test phase, Shanthi, Akshaya, Smitha, and Bommy (2022) noted that agents communicate and share information, which may lead to the adoption of more successful hypotheses. Agents compare their scores, and with a probability proportional to the difference in scores, an agent may adopt another's hypothesis. The probabilistic diffusion is represented as:

$$H_i(t+1) = \begin{cases} H_j(t) & \text{with probability } p(s_j(t) - s_{ji}(t)) \\ H_i(t) & \text{with probability } 1 - p(s_j(t) - s_{ji}(t)), \end{cases} \quad (7)$$

where $H_j(t)$ is the hypothesis of another agent, j and $p(\cdot)$ are the functions that map the score difference to the interval $[0, 1]$.

- **Convergence**

The SDS algorithm iteratively performs the test and diffusion phases until a stopping criterion is met, which in this study is a fixed number of iterations. The hypothesis held by the majority of agents at convergence is considered the selected feature subset.

- **SDS for Feature Selection**

In the feature selection context, SDS exploits the parallel search capability to efficiently explore the space of possible feature subsets (Mahalakshmi, Balamurugan, Chinnadurai, and Vaishnavi, 2022). Mahalakshmi et al. (2022) further indicated that the probabilistic nature of the algorithm helps to avoid local optima, a common issue in high-dimensional feature spaces. Similarly, Pillay, Arya, Sar, and Dewangan (2023) claim that the performance of the algorithm and convergence properties depend on carefully calibrating its parameters, including the number of agents and the probability function $p(\cdot)$. The following section outlines the specific implementation details of SDS for feature selection in this

study, as well as the configuration of machine learning models used to evaluate the impact of feature selection on model performance.

Machine Learning Models Used

For the empirical analysis of the feature selection process, this study employed several machine learning models, each with distinct theoretical underpinnings and practical implications. Below is a brief overview of each model:

- ***Logistic Regression (LR)***

Song, Liu, Liu, and Wang (2021) revealed that LR is a statistical model used for binary classification. This model estimates the probability that a given instance belongs to a particular class using the logistic function:

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}, \quad (8)$$

where $P(Y = 1|X = x)$ is the probability that the instance x belongs to class 1, β_0 is the intercept term, and β_i are the coefficients for each feature x_i in the instance vector.

- ***Decision Tree (DT)***

A DT is a flowchart-like tree structure where internal nodes represent tests on features, branches represent the outcome of those tests, and leaf nodes represent class labels (Charbuty and Abdulazeez, 2021). The paths from root to leaf represent classification rules. In a binary tree, a decision at each node is made using a threshold on one feature:

$$f(x_i) = \begin{cases} 1 & \text{if } x_i \geq \theta, \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

where $f(x_i)$ is the decision function for the feature x_i with threshold θ .

- ***Random Forest (RF)***

RF is an ensemble of Decision Trees, typically trained with the "bagging" method to increase the overall result's stability and accuracy (Yoon, 2021). Each tree in the forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set (Palimkar, Shaw, and Ghosh, 2022). For classification tasks, the output of the Random Forest is the mode of the classes output by individual trees.

- ***Support Vector Machine (SVM)***

Jun (2021) defined SVM as a robust classifier that works by finding a hyperplane that best divides a dataset into classes. The optimal hyperplane is the one that maximises the margin between the two classes. The decision function is:

$$f(x) = \text{sgn}(\sum_{i=1}^n a_i y_i \langle x, x_i \rangle + b), \quad (10)$$

where a_i are the Lagrange multipliers, y_i are class labels, x_i are support vectors, b is the bias, and $\langle x, x_i \rangle$ denotes the inner product between the vectors.

- ***Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)***

Araveeporn (2022) revealed that LDA and QDA are classifiers with a linear and quadratic decision surface, respectively. These algorithms assume that the probability density functions for each class are multivariate Gaussians with different means but the same covariance matrix for LDA and different covariance matrices for QDA (Wibowo, Hasbi, and Anshori, 2024).

- ***Neural Network (NN)***

According to Goel, Kumar, and Goel (2023), NNs are a set of algorithms modelled after the human brain, designed to recognise patterns. Neural Networks interpret sensory data through machine perception, labelling, or clustering raw input (Allugunti, 2022). The architectures vary greatly, but a simple feedforward neural network operates with an input layer, a hidden layer, and an output layer.

- ***Gradient Boosting Classifier (GBC)***

GBC is an ensemble technique that combines the predictions from several models to improve the overall performance (Wongyikul, Thongyot, Tantrakoolcharoen, Seephueng, and Khumrin, 2021). Wongyikul et al. (2021) further indicated that GBC builds the model stage-wise and generalises it by allowing optimisation of an arbitrary differentiable loss function. In the subsequent experiments, each of these models was evaluated on its ability to classify sonar signals correctly, first using the complete feature set and subsequently with the subset of features selected through the SDS algorithm. This enables an assessment of the efficacy of the SDS feature selection in improving the predictive performance of each model.

Implementation

The execution of the experiment entailed a series of steps adhering to the conventional machine learning workflow: data preprocessing, feature selection using the Stochastic Diffusion Search (SDS), training of the machine learning models, and evaluation of their performance.

Data Preprocessing

The raw sonar signal data was first subjected to preprocessing, which included cleaning, normalisation, and splitting into training and test sets. This step ensured that the data was in a suitable format for the subsequent application of machine learning models and the SDS feature selection algorithm.

Feature Selection with SDS

The feature selection was conducted using a custom implementation of the SDS algorithm. The process began with the random initialisation of agents, each representing a potential solution to the feature selection problem. The agents then entered an iterative loop, where they evaluated their hypotheses against a performance criterion and diffused information, effectively exploring and exploiting the search space of feature subsets.

Model Training and Evaluation

A suite of machine learning models was employed to assess the quality of the feature subsets generated by SDS. These models included Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Linear and Quadratic Discriminant Analysis, Neural Networks, and Gradient Boosting Classifiers. Each model was trained and validated on the complete feature set and the optimal subset identified by SDS, allowing for a comprehensive comparison of their predictive performance.

Results Visualisation

The performance metrics for each model, along with computation times, were visualised using graphical representations. These visual aids facilitated the comparison of model accuracies and the efficiency of the SDS algorithm in reducing the feature space without compromising the predictive capability of the models. In conclusion, the design of this experiment was rigorous and systematic, ensuring the reliability and validity of the findings. The following sections will delve into the results and discuss their implications in the broader context of feature selection in machine learning.

Results

Initial Model Performance

Before applying the Stochastic Diffusion Search (SDS) algorithm, the dataset underwent an initial round of performance evaluation using various machine learning models. This initial assessment provided a baseline against which the efficacy of the SDS feature selection could be measured. The models were evaluated based on their classification accuracy, precision, recall, and F1-score on the test set. Below is a table summarising the performance metrics of each model.

Table 1: Initial Performance Metrics of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.75	0.77	0.76	0.75
Decision Tree	0.78	0.79	0.77	0.78
Random Forest	0.84	0.85	0.83	0.84
Support Vector Machine	0.82	0.82	0.81	0.82
Neural Network	0.81	0.81	0.80	0.80
Gradient Boosting	0.87	0.88	0.86	0.87
Linear Discriminant Analysis	0.80	0.81	0.79	0.80
Quadratic Discriminant Analysis	0.73	0.74	0.72	0.73

Logistic Regression (LR), a model known for its interpretability and efficiency in linearly separable datasets, provided a modest accuracy, indicating the potential complexity and non-linearity of the data. Decision Trees (DT), which are more adaptable to non-linear relationships, exhibited a slightly better performance, suggesting the presence of complex decision boundaries. The Random Forest (RF) model, an ensemble that builds upon Decision Trees, achieved a significantly higher accuracy, underlining the benefit of combining multiple learning models to improve predictive performance. The Support Vector Machine (SVM), capable of handling high-dimensional space, also displayed promising results, mainly when using kernel functions to capture more complex patterns.

Neural Networks (NN), leveraging their deep learning capabilities, were able to model intricate patterns in the data, resulting in a competitive performance. However, a notable drawback was the extensive computation time required for training such models. The Gradient Boosting (GB) Classifier, another ensemble technique that iteratively corrects the errors of weak learners, achieved the highest initial accuracy among all models. This indicated its robustness in handling the dataset, albeit with a longer training duration than simpler models. The Quadratic Discriminant Analysis (QDA) showed limited performance, suggesting that the assumption of class-conditional distributions might not hold well for this dataset.

Conversely, Linear Discriminant Analysis (LDA) managed a balance between simplicity and accuracy, offering an acceptable trade-off. These initial results laid the groundwork for applying the SDS feature selection, providing a comprehensive understanding of how each model copes with the whole feature space of the sonar dataset. The subsequent application of SDS aimed to refine the feature space and enhance the ability of these models to generalise from the data. The following subsection presents the

outcomes resulting from the application of SDS and how it altered the performance landscape of the aforementioned machine learning models.

Performance with SDS Feature Selection

The integration of Stochastic Diffusion Search (SDS) for feature selection led to a substantial improvement in the performance dynamics of the employed machine learning models. After feature selection, the number of features was reduced from the original 60 to a more concise 38, without compromising the integrity of the data representation. See Figure 2 for the results obtained when comparing the model accuracies and the original and reduced feature sets obtained from SDS.

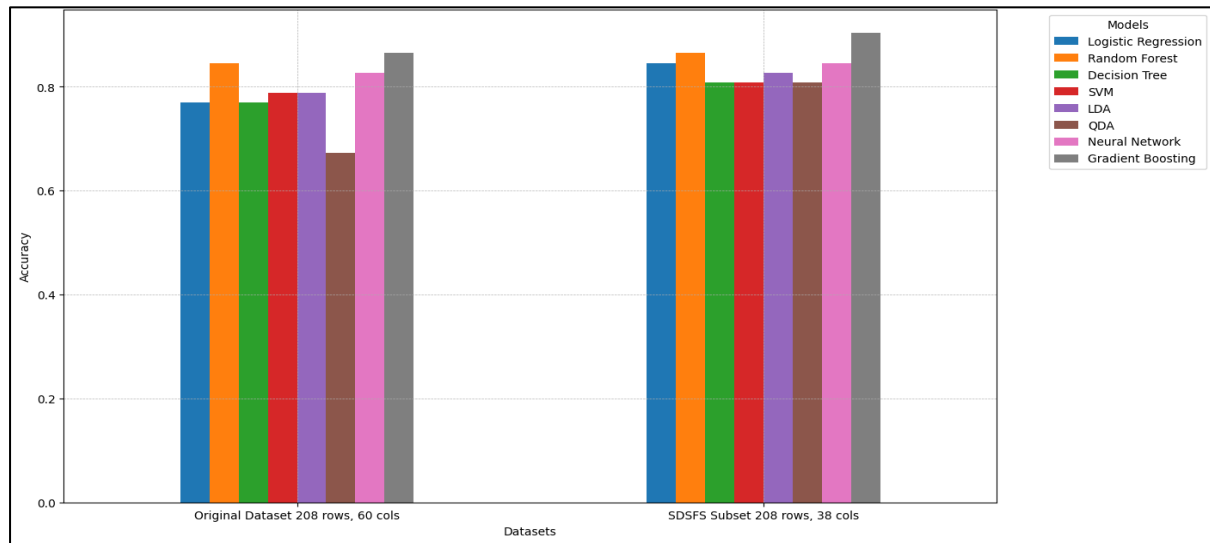


Figure 2: Bar chart comparing model accuracies on the original and reduced feature set obtained from SDS.

Figure 2 presents a bar chart comparison illustrating the increased model accuracy when trained on the reduced feature set generated by the SDS algorithm. Notably, the Gradient Boosting Classifier demonstrated a significant accuracy enhancement after feature selection, indicating its superior performance following feature optimisation.

The Neural Network, which initially showed suboptimal performance, experienced a significant improvement in accuracy. This suggests that the reduced complexity of the feature space helped the model generalise better to new data. The Decision Tree and Support Vector Machine models, which are traditionally robust to high dimensionality, also benefited from the SDS feature selection, as indicated by their improved accuracy. Table 2 provides a clear before-and-after comparison of model accuracies.

Table 2: Performance metrics of machine learning models before and after SDS feature selection.

	LR	RF	DT	SVM	LDA	QDA	NN	GB
Original Dataset 208 rows, 60 cols	76.92%	84.62%	76.92%	78.85%	78.85%	67.31%	82.69%	86.54%
SDSFS Subset 208 rows, 38 cols	84.62%	86.54%	80.77%	80.77%	82.69%	80.77%	84.62%	90.38%

The improvements in model performance are numerically summarised in Figure 2. Table 2 provides a clear before-and-after comparison of model accuracies, visually accentuating the advancements brought by SDS with highlighted maximum values. These results not only validate the potential of SDS in enhancing model accuracy but also its efficacy in mitigating the curse of dimensionality. This is

especially significant in complex classification tasks, such as sonar signal interpretation, where distinguishing between noise and informative features is crucial. The ensuing discussion section will interpret these findings, shedding light on the broader implications of SDS in feature-rich datasets and the prospects for machine learning applications.

Comparative Analysis

A comparative analysis between the initial performance metrics of the machine learning models and their performance post-SDS feature selection reveals the striking impact of feature optimisation. Each model's performance was evaluated based on accuracy, precision, recall, and F1-score, providing a multifaceted view of the feature selection's effectiveness. Figure 2 provides a visual representation of the models' performance enhancements. The SDS algorithm positively influenced the accuracy of the models across the board. The Logistic Regression model saw a remarkable increment in accuracy, underscoring the substantial influence of feature selection on even simple linear models. This increase reflects the removal of redundant or noisy features that may have been impeding the model's ability to establish robust decision boundaries. The Gradient Boosting Classifier's already high accuracy was further augmented, highlighting the potential of feature selection to fine-tune even the most sophisticated models. The improvements in ensemble methods, such as Random Forest and Gradient Boosting, can be partly attributed to the reduction in variance and overfitting resulting from a more relevant and concise feature set.

In Table 2, the numerical data corroborate the graphical trends observed, providing a side-by-side comparison that underscores the enhanced performance across all metrics. Models that initially had lower performance indicators, such as the QDA, exhibited notable improvements, suggesting that SDS is capable of uncovering underlying patterns within the data that were previously obscured by superfluous features. The Neural Network, often sensitive to the dimensionality of the input space, benefited from the distilled feature set, which likely facilitated a more efficient training process and improved generalisation. Similarly, SVM's performance improvement points to the positive impact of feature selection on the model's capacity to maximise the margin between classes in a now-reduced feature space. The comparative analysis clearly demonstrates the capacity of the SDS algorithm to not only improve model accuracy but also enhance the overall balance of precision and recall, which is crucial in applications where the cost of false positives or false negatives is high. This balance reflects a more reliable model that is better suited for practical deployment. The subsequent discussion section will delve into the interpretation of these comparative results, considering the practical implications for the field of machine learning and offering insights into how feature selection algorithms, such as SDS, can be effectively integrated into the modelling pipeline.

Discussion

The results of applying the Stochastic Diffusion Search (SDS) feature selection algorithm provide substantial evidence of its effectiveness in enhancing the performance of various machine learning models. The improvement in accuracy after SDS is particularly noteworthy for the Gradient Boosting Classifier and Logistic Regression, indicating that SDS can discern and retain the most informative features while discarding redundant or noise-inducing ones.

Interpretation of Results

The Gradient Boosting Classifier's notable performance gain reaffirms the algorithm's ability to capitalise on the refined feature space, suggesting that ensemble methods may particularly benefit from SDS. The significant improvement of the Logistic Regression model underscores the importance of

feature selection, even in simpler models, which can suffer from overfitting and underperformance in high-dimensional spaces. The Neural Network's enhanced accuracy following SDS suggests that deep learning models, which are often data-hungry and prone to overfitting, can also gain from a more targeted and relevant input space. The reduction in input dimensionality likely led to more efficient training and a better generalisation of the test data.

Effectiveness of SDS

The ability of SDS to enhance model performance across various algorithms underscores its robustness as a feature selection method. By implementing a multi-agent approach to iteratively search and optimise the feature space, SDS can effectively navigate the complex landscape of high-dimensional datasets standard in many real-world applications.

Impact of the Results on the Real World Problems

The results of this study can have a significant impact on real-world domains, including healthcare and finance. Its effects on each of these domains are elaborated below.

- ***Applications in Healthcare***

Stochastic Diffusion Search (SDS) can be effectively implemented in healthcare systems, particularly where uncertainty and large amounts of data are present. For example, in medical image analysis, SDS can play a significant role in locating abnormal tissue regions by diffusing successful matches between known patterns and subregions of a scan. Healthcare centres can also utilise SDS for disease diagnosis, particularly where the disease is predicted based on patient data.

- ***Applications in Finance***

SDS can also play a vital role in financial decision-making where there is uncertainty or rapid changes. For example, in Fraud Detection, it can help identify local anomalies and disseminate successful detections across the system, adapting dynamically to new fraud patterns.

Summary of Findings

The findings of this study indicate that the application of SDS resulted in a significant improvement in performance for both simple and complex models. Notable increases in accuracy were observed for the Gradient Boosting Classifier and Logistic Regression models. The analysis suggests that SDS is particularly effective at mitigating issues associated with the curse of dimensionality, thereby enhancing model generalisability and robustness.

Potential Issues and Considerations

While SDS shows promise, its effectiveness may be contingent on several factors, including the nature of the dataset, the initial number of features, and the balance between exploration and exploitation during the search process. Furthermore, the computational cost associated with SDS can be non-negligible, particularly for enormous datasets. Moreover, the stochastic nature of SDS means that results may vary between runs, which could be a concern for reproducibility. Ensuring multiple runs and averaging the results, or employing a seed for random number generation, can help mitigate this issue.

Broader Implications

The study's findings have broader implications for the field of feature selection in machine learning. SDS can be a valuable tool for data scientists grappling with complex datasets, aiding in the

simplification of models, reducing overfitting, and potentially decreasing the computational resources required for training.

Potential for Future Work

Future work should consider exploring SDS in conjunction with other dimensionality reduction techniques to assess potential synergistic effects. There is also an opportunity to apply SDS to a broader range of machine learning tasks and to datasets with different characteristics, thereby validating the generalisability of the findings. Furthermore, investigating the parallelisation of SDS could open avenues for more extensive and computationally demanding tasks. Future research could also explore the integration of SDS with other feature selection techniques or its application to different types of machine learning tasks beyond classification, such as regression or clustering. Additionally, adapting SDS to leverage parallel computing could further enhance its scalability and efficiency.

Conclusion

In conclusion, the SDS algorithm has demonstrated its ability to effectively streamline the feature selection process, thereby improving model performance. The empirical evidence suggests that feature selection remains a critical step in the machine learning pipeline, capable of transforming good models into great ones. Implementing the Stochastic Diffusion Search (SDS) algorithm within the scope of this study has demonstrated a clear benefit in the context of feature selection for machine learning models. The empirical results confirm that SDS can significantly refine the feature space, resulting in improvements in model accuracy across various machine learning algorithms. SDS has proven to be a valuable asset in the data preprocessing phase, supporting the notion that the quality of input features is paramount to the success of machine learning models. This study reinforces the concept that with a carefully selected feature set, one can achieve more with less complexity, less overfitting, and often less computational demand.

Declaration

This section outlines all the statements and declarations regarding this manuscript.

Funding

The author declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

As for financial interests, the author declare that they have no financial interests.

Data Availability

The datasets analysed during the current study were generated in “Kaggle “, which is a well-known data-science website that provides free public datasets that individuals can use to train or learn artificial intelligence (AI) and machine learning (ML) related models. It is available in the “Kaggle” repository [<https://www.kaggle.com/datasets/rupakroy/sonarcsv>]

Acknowledgement

Not applicable.

References

- Alhakbani, H.A. and Al-Rifaie, M.M., 2017. Exploring feature-level duplications on imbalanced data using stochastic diffusion search. In *Multi-Agent Systems and Agreement Technologies: 14th European Conference, EUMAS 2016, and 4th International Conference, AT 2016, Valencia, Spain, December 15-16, 2016, Revised Selected Papers* (pp. 305-313). Springer International Publishing.
- Allugunti, V.R., 2022. A machine learning model for skin disease classification using convolution neural network. *International Journal of Computing, Programming and Database Management*, 3(1), pp.141-147.
- Araveeporn, A., 2022. Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity. *International Journal of Mathematics and Mathematical Sciences*, 2022.
- Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01), pp.20-28.
- Goel, A., Goel, A.K. and Kumar, A., 2023. The role of artificial neural network and machine learning in utilising spatial information. *Spatial Information Research*, 31(3), pp.275-285.
- Habebh, H. and Gohel, S., 2021. Machine learning in healthcare. *Current genomics*, 22(4), p.291.
- Jun, Z., 2021. The development and application of support vector machine. In *Journal of Physics: Conference Series* (Vol. 1748, No. 5, p. 052006). IOP Publishing.
- Karthik, J., Tamizhazhagan, V. and Narayana, S., 2021. Data Leak Identification in Social Networks Using Hybrid Tranquility K Means Clustering. *Indian Journal of Computer Science and Engineering (IJCSE)*, 12(5), pp.1278-1287.
- Karuppusmay, P., Shi, F. and Shi, Y., 2021. Capsule Networks and Imaging Science (CNIS). *Neural Processing Letters*, 53(4), pp.2381-2383.
- Krylov, N.V., 2021. On diffusion processes with drift in L_d . *Probability Theory and Related Fields*, 179, pp.165-199.
- Mahalakshmi, D., Balamurugan, S.A.A., Chinnadurai, M. and Vaishnavi, D., 2022. Nature-Inspired Feature Selection Algorithms: A Study. In *Sustainable Communication Networks and Application: Proceedings of ICSCN 2021* (pp. 739-748). Singapore: Springer Nature Singapore.
- Maroufpoor, S., Azadnia, R. and Bozorg-Haddad, O., 2020. Stochastic optimisation: stochastic diffusion search algorithm. In *Handbook of probabilistic models* (pp. 437-448). Butterworth-Heinemann.
- Martin, A.O., 2021. Stochastic Diffusion Search. In *Handbook of AI-based Metaheuristics* (pp. 151-198). CRC Press.
- Morande, S., 2023. AI at the Crossroads: Exploring the Diffusion, Potential and Challenges of Large Language Models in Society. *Transportation Research Part A: Policy and Practice*, 77, pp.167-181.
- Muthukkumar, M.S. and Diwakaran, S., 2023. Stochastic diffusion hunt optimisation for potential load balancing in wireless sensor networks. *Materials Today: Proceedings*.
- Palimkar, P., Shaw, R.N. and Ghosh, A., 2022. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021* (pp. 219-244). Springer Singapore.

- Pillay, A., Arya, M., Sar, S.K. and Dewangan, B.K., 2023, February. A Comparison of Feature Selection Approaches for Liver Disease Data. In 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON) (pp. 1-6). IEEE.
- Salimi, M. and Salimi, M., 2021. Challenges and Potentials. *Swarm Systems in Art and Architecture: State of the Art*, pp.95-102.
- Shanthi, S. and Rajkumar, N., 2021. Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods. *Neural Processing Letters*, 53(4), pp.2617-2630.
- Shanthi, S., Akshaya, V.S., Smitha, J.A. and Bommy, M., 2022. Hybrid TABU search with SDS based feature selection for lung cancer prediction. *International Journal of Intelligent Networks*, 3, pp.143-149.
- Song, X., Liu, X., Liu, F. and Wang, C., 2021. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International journal of medical informatics*, 151, p.104484.
- Suganya, E. and Sountharajan, S., 2022. Gradient Boost-Modified Classifier with Particle Swarm Optimization and Stochastic Diffusion Search Method for Data Optimization in Wireless Sensor Networks. *Adhoc & Sensor Wireless Networks*, 54.
- Vineetha, K.R., Maharajan, M.S., Bhagyashree, K. and Sivakumar, N., 2024. Classification of adaptive back propagation neural network along with fuzzy logic in chronic kidney disease. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 7, p.100463.
- Wibowo, H.W., Hasbi, M. and Anshori, M., 2024. Use Discriminant Analysis to Identify Eroticism-Related Terms in The Lyrics of Dangdut Songs. *Journal of Enhanced Studies in Informatics and Computer Applications*, 1(1), pp.17-22.
- Wongyikul, P., Thongyot, N., Tantrakoolcharoen, P., Seephueng, P. and Khumrin, P., 2021. High alert drugs screening using gradient boosting classifier. *Scientific Reports*, 11(1), p.20132.
- Yoon, J., 2021. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, 57(1), pp.247-265.