International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Dimensionality Reduction Approach for High Dimensional Data using HGA based Bio Inspired Algorithm

Mr. Ashish Kumar Rastogi1, Prof. (Dr.) Swapnesh Taterh 2 and Dr. B. Suresh Kumar3

Submitted: 25/10/2022 **Revised:** 26/12/2022 **Accepted**: 22/01/2023

Abstract: Data scientists primarily seek to develop innovative methods for analysing data that are both computationally efficient and time efficient, resulting in effective data analytics. When a result, as more data is generated from different sources, the amount of data that may be analysed, manipulated, and visualised grows rapidly. We presented feature selection and evolutionary methods in this study. We also concentrate on data analytic process optimization techniques, i.e., we propose to investigate applications of nature-inspired Hybrid Genetic Algorithm (HGA) algorithms. This incorporates the GA, ACO, and PSO approaches, respectively. Feature selection optimization is a hybrid strategy that optimises chosen characteristics using feature selection techniques and evolutionary algorithms. Prior work iteratively solves this issue to arrive at an appropriate feature subset. Feature selection optimization is a domain-independent method. We employed vast dimensional data to test the suggested model. We extracted the android APK dataset from the permission API used to detect whether the applications are malicious or standard. The initial dataset contains around 351 attributes, and after the application algorithm, it reduces up to 13-16% of attributes-based essentiality and correlation of each attribute. The experimental evaluation has done on the Weka 3.7 environment for validation, and it achieves 90.70% accuracy for NB and 92.5% for SVM.

Keywords: *Dimensionality reduction, optimization algorithms, genetic algorithm, particle swarm optimization, ant colony optimization, classification.*

1. Introduction

The dimensionality of a data refers to the set of input variables or characteristics. Approaches for reducing the number of input parameters in a dataset are known as Dimensionality Reduction (DR). The burden of dimensionality refers to how adding more attribute values makes a predictive analysis job more difficult to estimate. Approaches such as Feature Selection, Matrix Factorization, Manifold Learning, and autoencoder are some of the strategies which can be used to reduce the dimension. Feature selection strategies, utilize scoring or statistical tools to identify which traits to preserve and what to eliminate, are the most popular. Wrapper approach and filter algorithm are the two basic types of feature selection approaches. Wrapper approaches wrap a Machine Learning (ML) technique, training and testing it with several subgroups of input data and choosing the subset that makes the greatest predictive accuracy. A wrapper feature selection technique is an illustration. Filter technique choose the most predicted set of input data using scoring techniques such as association between the characteristic and the target value. Pearson's association and the Chi-Squared testing are two instances. For Dimensionality Reduction (DR), Factorization of matrix n, a linear algebra approach, can be applied. Matrix factorization algorithms, in particular, can then be used to break down dataset matrices into its component pieces. The Eigen Decomposition and Singular Value Decomposition (SVD) are two instances. After that, the sections can be sorted, and a set among those portions can be chosen that better represents the matrix's key structure and can be utilized to describe the data. Principal Components Analysis, called PCA for brief, remains the most popular approach for ranking the parts. Manifold Learning is indeed a high-dimensionality statistics method which can also be used to reduce the dimensionality. These approaches, known as manifold learning, are being used to construct low-dimensional projections of high-dimensional information, which is widely used during visual analytics. The projection is intended to produce a low-dimensional depiction of the data while keeping the data's most important structure and connections. Deep Learning Neural Networks (DLNN) can be built using Autoencoder Techniques to do Dimensionality Reduction (DR). This entails formulating a self-supervised learning

1 Research Scholar Amity University Rajasthan, India
ashish.k.rastogi24@gmail.com
2 Professor Amity Institute of Information Technology Amity University Rajasthan, India swapnesh@hotmail.com
3 Associate Professor Sanjay Ghodawat University Kolhapur, India
sureshkumarbillakurthi@gmail.com

challenge in which a system must accurately recreate the inputs.

The majority of actual data mining applications are distinguished by high-dimensional data with noise, in which all qualities are not equally essential. Since the system takes time to analyze and assess certain type of information, the implementation performance deteriorates. Several of the most widely used data mining techniques are ineffective in deal with that sort of useless and inconsistent data. Even though the data has many dimensions, Decision Tree (DT) techniques are not operationally meaningful. As a result, there is really no best strategy for Dimensionality Reduction (DR) and also no method to problem mapping. The paper provides certain dimensionality reduction approaches as a pretreatment step to address this challenge and improve system throughput. By taking into account all of the relevant aspects, the suggested dimensionality reduction method offers reduced dimensionality depiction without sacrificing any data. Thus, using only dimensionality reduction methods and a nature-inspired approach, a paradigm is proposed to improve predictive performance. The major goal of this study is to increase accuracy by using a hybrid technique for feature selection as well as classification. To improve dataset reliability, the best search greedy oriented and Naive Bayes (NB) technique is presented. As a result, a nature-inspired method was created to improve search capability and dataset prediction performance. The proposed framework makes a comparison to current classification techniques such as J48, sequential minimal optimization, and interpretation. Using the help of feature extraction and feature selection technique, the suggested model attempts to address the issues related to data intricacy in big data sets with undesired attributes. For feature selection, nature inspired technique is applied. Principal Component Analysis (PCA) is used to extract features. J48 is eventually utilized for comparative purposes. In comparison to existing techniques, the suggested system aspires for high efficiency and scalability.

The Proposed system is divided into various units. In Unit II, all related work of dimensionality reduction and feature selection is elaborated in brief. Unit III focuses on proposed system design, where system architecture is described in detail. Algorithm design is described in unit IV. Unit V and Unit VI concentrated on empirical outcome and conclusion respectively.

2. Literature Review

Mandikal Vikram et al. [1] conducted a thorough study of the performance and quality of several dimension reduction approaches in 2019. A thorough review of several methodologies has been provided, with actual datasets serving as benchmarks. This research aims to help data science professionals choose the best appropriate dimensionality reduction strategies focused on the trade-off among systems performance. In the year 2021, Faxian Cao et al. [2] Considering its application areas in information processing, Unstructured Sparse Dimensionality Reduction (USDR) for Hyper Spectral Images (HIS) still faces two problems that restrict its exclusionary effectiveness: Firstly, it can't be used for dimension reduction including both samples of testing and training; secondly, it can't combine spectral information with spatial data for enhancing discriminative performance of hyper spectral images. It expands the first issue to a supervised situation, specifically Dimension Reduction Sparse Reconstruction (DRSR), which may be used for either training or testing datasets. Furthermore, to more enhance the discriminative efficiency of HSIs, a unique technique termed local and global DRSR is suggested to incorporate the spectral as well as spatial data of HSIs. With a uniform measurement matrix, the suggested DRSR calculates the spatial information amongst pixels of HSIs, encompassing the entire instances and associated related positions. The suggested DRSR considerably surpasses various state-of-the-art techniques in experiments. Myasnikov [3] investigated the Uniform Manifold Approximation and Projection (UMAP) approach for hyperspectral information analysis. For very well-known HSI, ideal conditions are obtained and contrasted to the findings of other approaches, such as Principal Component Analysis., Nonlinear Mappin, Linear Embedding Locally, Laplacian Eigenmaps as well as Isomap E.

In 2012, Pouria Fewzee et al. [4] The implementation of 3 contemporary dimension reduction methods to the emotional voice identification issues, mainly greedy attribute extraction, elastic net, as well as Supervised Principal Component Analysis (SPCA), was examined. The two pairs of speech signals were retrieved from the Visual Asset Management (VAM) database, and tests were conducted using every set individually and in conjunction. As per the findings of this research, while considering dimension reduction into account, a larger vector of characteristics does not really allow for better prediction owing to the sub optimality of such dimensionality reduction algorithms. In addition, the grade of the dimensionality reduction technique chosen, in regards of effectiveness and goal region size, is determined by the training objective criteria. The 3 distinct learning jobs (valence, activating, and domination) were investigated in the study, utilizing 3 distinct sets (p, q, & p+q), and it is impossible to say with certainty that one strategy is better compared to everyone else in terms of effectiveness and precision. In 2020, Feiping Nie et al. [5] suggested Semi-supervised Adaptive Local Embedding Learning (SALE), a new locality retained dimensionality reduction conceptual model that understands a local discriminative encoding by establishing a K1-Nearest Neighbour (k1NN) graph on clearly labeled data to examine the internal representation, That is, sub-manifolds from non-Gaussian data points. It explored the geometric features of all items by mapping all data into trained embedding and generating new k2NN graph across all embedding algorithm. As a result, unlabeled data as well as their labeled neighbors can be grouped into the very same sub-manifold to boost integrated data's discriminative potential. Additionally, depending on the proposed SALE paradigm, 2 semi-supervised dimension reduction algorithms using orthogonal as well as whitening requirements were presented. The NP-hard issue in these simulations is solved using an effective alternate incremental optimization approach. Additional testing on a variety of organic and inorganic and actual data sets shows that the approaches outperform the competition when it comes to local structure discovery and categorization.

Jingyu Wang et al. [6] suggested an Unsupervised Adaptive Embedding (UAE) approach for Dimensional Reduction, which would be a linear graph-embedding technique, to overcome these issues in 2021. The affiliation network was then constructed using an adaptable neighbor methodology. Secondly, the affinity graph creation and prediction matrix computation are combined. It takes into account the localized link among specimens and the worldwide characteristics of high-dimensional information, wherein the cleansed data matrix was first presented to eliminate noise in subdomains. The suggested technique's interaction with Local Preserving Projections (LPPs) is however investigated. Furthermore, a new iterative optimal solution is intended to fix the improved method, and its resolution and computing difficulty are investigated. Suchismita Das et al. [7] suggested a broad paradigm for unsupervised fuzzy rule relied Dimensionality Reduction (DR), with a focus on data presentation, in 2021. The following significant properties of this approach are crucial to Dimensionality Reduction (DR) for visual representation: (i) retain neighborhood associations; (ii) successfully handles information on non-linear manifolds, (iii) able to deliver out-of-sample design parameters, (iv) It can refuse data points whenever necessary, and (v) reasonably easily understandable. The Takagi-Sugeno first-order design was adopted. Specialists generally give fuzzy rules, or they are retrieved to use an input – output trained model. There is no output information or expertise accessible. This actually makes things more difficult. The rule values were measured by solving the objective function that maintains within and between geodesic connections (lengths across the manifold) equivalent Euclidean distances upon that projected space. A newer version of the geometric c means clustering technique was developed in this application. The suggested technique is superior to the outcomes of six state-of-the-art information visualization techniques on various organic and actual data sets. The suggested scheme is the only one which consistently performs well across all data sets. The approach is identified as barriers to the starting circumstances. Tests that verify the technique's predictability are conducted. The suggested technique's capacity to refuse data points when they should be rejected is evaluated. The program's scaling problem is also examined. Because of the program's overall character, other objective functions can be utilized to create forecasts that meet various purposes. This is the very first time unsupervised fuzzy rule modeling has been used to manifold training.

In 2015, Lei Xu et al. [8] suggested in 2015 that low-dimensional interpretations of microblog could be learned using deep network-based algorithms. The suggested models make use of semantic relation comes from two kinds of microblog-specific data: retweet relationships and tagging. Deep Learning (DL) techniques outperform classic Dimensionality Reduction (DR) techniques like latent semantic analysis as well as Dirichlet allocation topic method in experiments, and the addition of microblog-specific data can aid in learning enhanced representations. Yongyan Zhang et al. [9] suggested a strategy for enhancing classification performance by integrating DR (dimensionality reduction) techniques and clustering methods in 2018. It is much more efficient in evaluating data's underlying structure. The approaches described in this study include grouping factors based on data resemblance, converting factors into small classes having maximum similarity, and separating factors with low resemblance. Following that, the data's dimensionality was lowered once it was grouped. The data is integrated once it has been decreased in dimension. Ultimately, Support vector machine is used to classify the combined information. The performance of the classifier is used to assess the method's reliability. Gladys M. Hilasaca et al. [10] proposed a framework for comparing DR strategies in 2019. The suggested technique analyzes DR results using quality attributes, and then visualizes the data utilizing biplots. This approach allows clients to see how variable modification affects DR procedures and to find relevant parts of the performance indicators being researched. A novel Dimensional Reduction was generated for this area of interest that incorporates the user-defined parameters, which were confirmed by formative and summative assessments. Chenfeng Guo et al. [11] published an initial work on emotion categorization dimensionality reduction in 2018. Researchers presented and contrasted five popular dimensionality reduction methodologies. Tests on the

Database for Emotion Analysis Using Physiological Signals (DEAP) data revealed that no strategy can consistently beat another, whereas using basic characteristics to do categorization is not really a terrible idea. Hieu Minh Bui et al. [12] used a current pre-trained Deep Neural Network (DNN) and the randomized orthogonal projecting technique to study an additional data DR methodology for image recognition. Because the suggested methodology does not require sequence data to be stored in storage, it is suitable for projects with minimal equipment.

Yanlong Wang et al. [13] effectively implemented Multiple-Point Statistics (MPS) for stochastic modeling by replicating characteristics from training examples to the replicated areas in 2020. Unfortunately, since these characteristics are generally nonlinear in nature, MPS techniques based on linear DR are ineffective for dealing with nonlinear information. Because manifold learning strategies can find the intrinsic properties of high dimensional data by mapping those to a lesser-dimensional manifold, a structure utilizing manifold learning for Dimension Reduction (DR) in MPS is applied for solving. Manifold learning techniques are incorporated with MPS to correctly eliminate the influence of trends from TIs because then the successive modeling can become more precise. Remzi Oten et al. [14] proposed a nonlinear MapReduce framework for hierarchically minimizing the Sammon's cost function. This structure is built on the groups that can be retrieved from the strong dimensional data's minimal-spanning-tree. Because gradient-descent-based approaches are prone to being stranded at local minima in these situations, through use of a Genetic Algorithm (GA) to reduce the objective functions has been examined extensively. Minfeng Zhu et al. [15] introduced Dimensional Reduction Graph, a new graph layout technique that improves the nonlinear DR process with three techniques: close to resembling graph ranges with a sparsity distance matrix, predicting the gradient with the deleterious sampling method, and speeding up the optimization method with such a multi-level layout system. Dimensional reduction graph grows up to large-scale networks containing thousands of vertices and maintains linear complexity for compute and storage usage. Experiments and analyses with state-of-the-art network design approaches show that Dimensional Reduction graph can build aesthetically equivalent layouts in a smaller duration of time and with less storage. GuangHui Yan et al. [16] transformed the instance selection issue into an optimal solution in 2008, attempting to discover the characteristic subset with the highest multifractal while still restrict the amount of characteristics. Individual attribute prioritization was linked with characteristic subgroup assessment for DR in order to prevent extensive search in the vast attribute subset field, and the unsupervised Sequential Forward Fractal Dimensionality Reduction (SFFDR) method was proposed. According to Zhao Shenglin et al. [17], everybody is bombarded by colorful images today, and graphics rendering has now become a hot topic. Everyone carried out various applications such as picture categorization, recognition, and retrieval, among others. Nevertheless, due to the obvious features - nonlinearity, complexity, and big amount — direct processing on those pictures is challenging. Many methods, including as PCA, Artificial Neural Network (ANN), and MCS were established to minimize dimensionality while maintaining innovative features. This research utilizes Sam and Lawrence's Locally Linear Embedding (LLE), one DR method. One nonlinear DR technique is the LLE method. An image recognition test was conducted using the LLE algorithm. The facial images were successfully identified by reducing the 92*112-D-pictures to 6 D using the LLE technique.

Dewa Made Sri Arsa et al. [18] suggested a deep belief network-based DR in 2016. As a case study, hyper spectral photographs were used. The Hyper Spectral Image (HIS) has a lot of features. Several studies have suggested utilizing LDA and PCA in spectroscopic hyper spectral image categorization to minimize hyperspectral picture dimensionality. It has presented a dimension reduction method for HSI categorization using a deep learning model. Deep Belief Networks (DBN) were employed in the suggested scheme. The first Deep Belief Network was used to lower the spectral band width, while the subsequent Deep Belief Network was utilized to collect spectral-spatial features and as a classification method. It evaluated Deep Belief Network and PCA efficiency using the Indian Pines large dataset, which contains 16 categories the results show that DBN did best than PCA in HSI categorization when used as a Dimensionality Reduction technique. D Lakshmi Padmaja et al. [19] conducted a report and comprehensive comparison analysis of research and data mining methodologies and procedures in 2016. The methods' outcomes are evaluated, and a relevant path is determined. It is acknowledged that numerous strategies, including such Sequential Forward Selection (SFS) as well as Random Subset Feature Selection (RSFS) are employed in conjunction with the KNN classification algorithm to reduce the memory space of the research corpora. The study describes these methods by recognizing various Dimensionality Reduction strategies which can be used to enhance accuracy. Jiajia Shu et al. [20] proposed a new Dimensionality Reduction approach in 2014 that combined Kernel Principal Component Analysis (KPCA) with Nonparametric Discriminant Analysis NDA. To reduce the dimension for the small sample dataset, KPCA is employed first, followed by NDA to increase the discriminative power of characteristics in the resultant subspace. The suggested method has been evaluated on the Corel picture collection, which combines and compresses hue, surface, and form data to evaluate recognition rate. Benson S. Y. Lam et al. [21] introduced a new technique for multi-dimensional information categorization in where in noisy signals is dispersed in multiple dimensions of separate categories in 2007. These datasets defy many known DR approaches, which presume that all units have noisy characteristics in almost the same dimension and that all units' trimming operations are performed on the very same parameters. Multi-classifiers are the most common way to solve this issue. Every classification algorithm works with a separate number of parameters and performs DR on its own.

The Machine Learning (ML) techniques, Random forest (RF) as well as Support Vector Machine (SVM) were utilized in the presented design by Sivaranjani S et al. [22] in 2021 to determine the possible odds of being impacted by Diabetes associated Disorders. Following preprocessing the information, step forward as well as backward feature extraction and selection is used to choose characteristics that affect forecasting. After selecting certain characteristics, the Dimensionality Reduction (DR) technique, Principal Component Analysis (PCA) is studied, as well as the predictive accuracy is 83 percent using Randomized Forest method, which is notable when compared to Support Vector Machine (SVM) that has an accuracy of 81.4 percent. S. Keller et al. [23] tackle the categorization of hyperspectral data in 2020, which is equivalent to information collected by the Environmental Mapping and Analysis Program (EnMAP), a hyperspectral satellite operation expected to be released into space soon. Although generated EnMAP information has previously been published, only some few research have concentrated on evaluate the effectiveness of techniques for identifying such EnMAP information. As a result, a competition for categorizing EnMAP information has been launched in order to encourage study into potential monetization tactics. A paradigm comprising Dimensionality Reduction, feature selection, and categorization was based entirely on the data provided. The influence of techniques for Dimensionality Reduction as well as feature selection on the categorization outcomes was investigated using many classifiers for pixel-wise categorization depending on distinct learning theory. Huu-Thanh Duong et al. [24] offer a method for classifying rice seed pictures based on the Histogram of Oriented Gradient characterization and extraction of features in 2019. The test is run on a VNRICE testing set, which demonstrates the efficacy of the suggested method by lowering the amount of extracted features while enhancing precision. The intention of this study, conducted in 2014 by Joshi Snehal K et al. [25], is to have a sufficient understanding of the various Dimensionality Reduction techniques currently available, as well as to incorporate the validity of any of the prescribed techniques, which is dependent on a number of criteria and differing circumstances. Afnan M. Alhassan K et al. [26] look at using feature extraction, dimensionality reduction (DR), and categorization algorithms to forecast and treat chronic illness in 2021. The appropriate choice of characteristics is critical for enhancing the diagnosis tools categorization performance. Furthermore, dimensionality reduction approaches enhance effectiveness of the trained systems. By constructing smart, adaptable, and automatic systems, categorization method gives accurate predicted findings on chronic illness data. In chronic illness identification, parallel and adaptive classifiers are examined, which also are utilized to accelerate the classification method and reduced overall money and effort. The issue of Dimensionality Reduction is handled in 2016 by G. Suresh Reddy et al. [27] by accomplishing component reduction through all the implementation of a new membership value. Singular value decomposition as well as Data gain techniques are used for selecting features, with best features retained. The proposed solution to Dimensionality Reduction is contrasted to feature selection method, and the findings demonstrate that the suggested strategy achieves better Dimensionality Reduction.

FAST is a unique clustering-based attribute selection technique for elevated data described by Shahana AH et al. [28]. The approach comprises removing unnecessary characteristics from a dataset, building a minimal spanning tree, tree splitting, and lastly choosing a set of attributes. Harchli Fidae et al. [29] suggested a new method for reducing the length of high-dimensional information in 2020. The procedure consists of three steps: (1) information translation, (2) data grouping, and (3) data removal. Irrespective of the number of categories or beginning code vectors, the ideal division was discovered in the grouping stage, and then a representation item was chosen from every group to substitute their neighbours in the source data. The idea that items in the very same cluster exchange data gives rise to this concept. K. Pavya et al. [30] investigated the effects of filter-based (F-Score) and wrapper-based feature selection methods on illness recognition and characterization. Principle Component Analysis (PCA) DR (dimensionality reduction) methods are also used in the research. Three criteria were used to evaluate the efficiency: precision, sensitivities, and specificity. The recommended methods were analyzed using four classification model: Multi-Layer Perceptron, Back Propagation Neural Network (BPNN), SVM and Deep Learning Machine. Although both F-Score as well as Recursive Feature Removal enhanced thyroid illness detection accuracy, the wrapper-based approach created the most effectiveness and had the higher precision of 98.14 percent with the Extreme Learning Machine (ELM) classification. In a Quick Serial Visual Presentation scenario, Tian Lan et al. [31] delivered a prepared string of letters to individuals. Offline, Electroencephalogram (EEG) information was acquired and processed. Because of its ease, the Event Related Potential sensor was created using a Linear Discriminant Analysis (LDA) classification algorithm. In greedy wrapper architecture, various dimensionality reduction and feature selection was used. Utilizing the top ten principal components for every channel fared best in experiments, and PCA could be utilized in both offline and online platforms. Dimension reductions as well as selection of feature are two key information preparation approaches in data mining methods, according to Muhammad Abu Bakar Siddik et al. [32] in 2021. The impact of information preprocessing on categorization techniques including Random Forest (RF), Support Vector Machine (SVM), Naive Bayes (NB), KNN, CNN, Decision Tree (DT) as well as Multi-layer Perceptron was demonstrate. Dimension Reduction (DR) was accomplished using PCA as well as Singular Value Decomposition (SVD). For Selection of feature, the Gini Index as well as Entropy was employed. Three corpora were used in the tests. Dimensionality Reduction (DR) and selection of the characteristics had little effect on classification results in the Diabetic Retinopathy Debrecen Database. Entropy-based characteristics extraction has a considerable impact on classification results in the Internet Users Purchasing Intent Data. The Singular Value Decomposition Dimension Reduction (SVD-DR) has effect on classification results for the Single Proton Emission Computed Tomography (SPECTF) Cardiac Data. The Principal component analysis and SVD-DR approaches, and the Entropy-based feature selection procedure, demonstrated increased classification precision than the previously published Backward Feature Removal dimension reduction technique.

In 2008, Zacharias Voulgaris et al. [33] proposed approaches for completing these two objectives individually using a unique notion called Discernibility, with the goal of solving categorization difficulties. The findings of the experiments back up the argument that the offered approaches are a suitable option for dimensionality reduction across a wide range of datasets and classification algorithm. Christos Boutsidis et al. [34] investigated Dimensionality Reduction for k-means classification in 2014. Dimensionality reduction combines two techniques: feature extraction as well as selection. A feature selection-relied k-means clustering technique picks a tiny proportion of the input characteristics before applying k-means clustering to them. A characteristic extraction-relied k-means clustering approach produces a small collection of new synthetic features before applying k-means clustering to them. Recognizing the benefits of k-means clustering and the abundance of heuristic approaches for it, there are no evidentially effective feature selection techniques for k-means clustering. In the research, however, there are 2 evidentially effective feature extraction approaches for k-means cluster formation: 1 randomized projection and another based on Singular Value Decomposition.

To minimize the characteristic dimensionality of brain signals, Jie Zhu et al. [35] presented an Hessian regularization based semi-supervised correntropy algorithm for robust feature selection (HR-SSCRFS) method in 2014. The test experiments demonstrated that the developed approach enhances the effectiveness of retrieving discriminative information from neuroimaging information for Alzheimer's disease categorization significantly. The suggested HR-SSCRFS method will be used to test multimodal neuroimaging information in the future. In 2020, R.Kavitha et al. [36] used the categorization of high-dimensional cardiovascular illness corpora in the pre-processing step of the data mining procedure. This unprocessed data consists of superfluous and contradictory information, which expands the searching area and storage arrays. It is necessary to delete redundant information data in order to obtain classification performance. With appropriate limits, the dimensionality reduction approach has been used to condense high-dimensional information into lesser Dimensional information. A framework is developed for simplified cardiac illness forecast. The platform is built by information extraction utilizing Principal Component Analysis and then computing a computational formula to pick the key features employing appropriate constraints. The suggested research contributed to the application's effectiveness, precision, and quickness. This can be used in knowledge discovery, image analysis, and pattern recognition apps. C. Deisy et al. [37] employed selection of feature to minimize unnecessary and duplicated features, improving prediction performance and lowering categorization processing cost. Fast Correlation Based Feature Selection (FCBC), Multi thread oriented FCBF characteristic extraction, and Decision Dependent – Decision Independent Correlation (DDCDIC) are all discussed in this work. These methods are concerned with feature relevance and bilateral information association for duplication testing in order to increase prediction performance and decrease calculation time. The obtained measurements are compared to the results of the Decision tree method of construction in the Weka tool for lung cancer, Tic 2000 Insurance business data, and breast cancer corpora. According to Luis H. Brito et al. [38], the post-genomic age has increased sales for streamlined operations to discover molecular mechanisms, which can be performed by using machine learning to the protein's features set. This method is known as a feature-based technique, and it's been the subject of various bioinformatics studies. The features of proteins were explored in this study to enhance the findings found in previous studies that employed Support Vector Machine to identify enzyme class. Even during experiment, two methods were tested: one with and one without dimensionality reduction using the statistical method Factor Analysis. These methods outperformed previous research, with F-measure averages of 85.10% and 83.90% in the techniques with and without dimensionality reduction, correspondingly.

Shraddha Sarode et al. [39] published a study in 2015. The World Wide Web now contains a massive amount of material. Web page categorization is one method of data management. Computational complexity is among the web page categorization challenges addressed in this research. The number of words on a web page is referred to as dimensionality. The high dimensionality of online pages makes classification difficult. The major goal of lowering the dimension of internet pages is to increase the classifier's effectiveness. This work uses a rough collection and data gain approach to offer a hybrid model to dimension reduction for web page categorization. To lower the dimension of online pages, feature extraction and dimension reduction approaches are applied. The attribute selection approach is the data gain technique. For dimensionality reduction, the Quick Reduct method based on rough datasets is being used. The naive Bayesian approach is used to classify website pages. For the presented work, substantial findings are achieved and verified. Rachid Benmokhtar et al. [40] discuss the matter of large-scale image representation for object detection and classification in a paper published in 2013. The suggested research focuses on recognition and classification and complexity of object recognition. The paper suggests utilizing Bagging and extraction of features via Support Vector Machine normal to repeatedly pick groups of projection from an outside data. Weights of Support Vector Machine normal in orthogonalized set of projection are used to élite attributes. To enhance the findings and get a more stable choice, the Bagging approach is used. The whole approach scales linearly with size range, allowing it to process massive state-of-the-art image representations. This technique significantly increases classification accuracy for decreasing vector dimension when provided Spatial Fisher Feature vector as data feed, as seen by outcomes on the prominent and hard PASCAL VOC 2007 standard.

3. Propsoed System Design

When using a HGA to tackle a specific issue, there are three important design options to make. Candidate solution representations must be chosen and encoded on the HGA chromosome, an objective function must be specified to evaluate the quality of each candidate solution, and finally, GA run parameters, including which genetic operators to use and their frequencies of operation, must be specified.

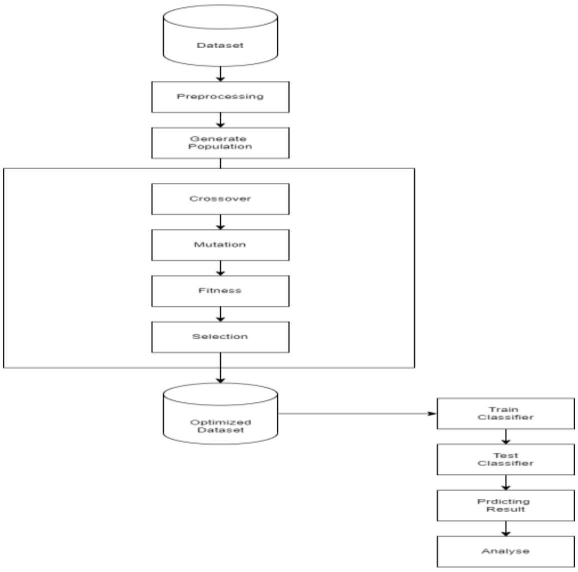

Fig. 1: proposed system architecture

The arrangement of the chromosome was rather simple for the HGA feature extractor. Each feature had a weight vector consisting of a single real value, followed by a masking vector consisting of one or more binary mask bits. When just a single mask bit was utilized, a zero meant that the characteristic should be ignored during classification, while a one suggested that it should be taken into account. Because each of these bits had a significant influence on the chromosome's interpretation, it was often beneficial to utilize more than one masking bit per feature. Because of the significance of these single bits, discontinuities in the objective function might occur, making GA optimization more challenging. Additional masking bits were employed to reduce this impact; a feature was classified only if the majority of the masking bits associated with that feature were set to one. Finally, along with the feature weights, some classifier-specific information was put on the chromosome to be improved. The SVM and NB are classifiers that we used for experimental analysis and show the effectiveness of proposed model.

Algorithm Design

This section describes the algorithms we used to generate the optimized matrix from the input dataset. The iterative genetic algorithm-based approach and ant colony optimization technique have combined to select the best attributes and reduce the dimensionality of a large matrix dataset. The below algorithms work like features extraction and selection techniques in Natural language processing and machine learning methodologies. Algorithm 1,2 and 3 demonstrates the entire procedure and reduction of dimensionality of whole data from throughout the process.

Algorithm 1: Generate random population for testing dataset (pop_gen)

Input: Java program under test data test_set, The initial population size pop_size, Max iteration size Max_itr, selection criteria for GA in %, Produced unique instances uM

Output: Generated Unique testing dataset uTest_data

```
T ← ∅
Population ←∅
Generate_pop ← Initialize_Population(pop_size)
Mutate_Score =0.0
Mutate_Score ← Fitness_Evaluation(Population, uM)
No_of_iteration ← 0
while No_of_iteration < Max_itr or Mutate_Score < max_size
pop ← pop ∪ Initialize_Population(pop_size − Generate_pop ())
MS ← Fitness_Evaluation(pop, uM)
Parent_pop ← select best fit according to % in selection phase
offspring ← Crossover(parent_pop)
Offspring ← mutation(parent_pop)
Pop ← pop ∪ offspring
Mutate_Score ← Fitness_Evaluation(Population, uM)
iteration ← iteration + 1
uTest_data ← pop_pop
return uTest_data
```

Algorithm 2: Calculate the fitness of each chromosome

Input :No. of test cases for calculation of fitness as TC, Unique mutant generated classes uM

Output :Generated Mutation score M_Score

```
1 MSS ← 0.05
```

```
n ← TC.size()
x ← uM.size()
all_killed_Instances ← ∅
foreach t ∈ (TC1 .... TCn) do
if Fitness[tc] == null
killed_Instances[t] ← ∅
TCC[t] ← Current_Method.invoke(S, TC)
foreach m ∈ (uM1... uMx) do
if Current_Method.invoke(m, t) 6= Current_Method.invoke(S, t) then
killed_Instances[t] ← killed_Instances[t] ∪ m;
MSS[t] ← killed_Instances[tc].size() × 100/uM.size()
Fitness_score[t] ← MSS[t] + 1/TCC[t]
Else
All_killed_Instances ← all_killed_Instances ∪ killed_Instances[t];
MS ← all_killed_Instances.size() × 100/uM.size()
return M_Score
```

Algorithm 3 : Generation of Unique attribute set

Input: The generated set of input matrix

Output : Final matrix with dimension reduced as T

```
TC ← Collections.sort(TC);
foreach t ∈ (TC1 ... TCn) do
All_Killed_Instances ← ∅
set New_flag ← False
foreach p ∈ (TC1 ... TCn)
if !TC.get_killed_Instances().contains(p.get_killed_Instances())
&!p.get_killed_Instances().contains(TC.get_killed_Instances()) then
All_Killed_Instances ← TC.get_killed_Instances()
else if p.get_killed_Instances().contains(t.get_killed_Instances()) & p.get_killed_Instances().size() >
TC.get_killed_Instances().size() then
All_Killed_Instances ← p.get_killed_Instances()
if All_Killed_Instances.size()>0 & (All_Killed_Instances.contains(t.get_killed_Instances()) ||
TC.get_killed_Instances().contains(All_Killed_Instances)) then
set New_flag ← True
break;
if New_flag then
remove t from TC
return T
```

The above algorithms provide feature extraction by using proposed optimization techniques. The proposed evolutionary algorithms have been effectively applied to most engineering, science, and even biology and medicine areas, making them strong and resilient. This study gives dimension reduction and large-scale optimization issues all the knowledge they need.

4. Results and Discussions

In this study, we have done a heterogeneous experimental setup to evaluate the proposed work. The open-source environment has been used with JDK 1.8 with NetBeans 8.0. The 2.7 GHz processor has been used with 8 GB RAM. All experiments are independently executed in a similar environment and demonstrated the entire obtained result in table 1.

Table 1: dimensionality reduction with various optimization techniques and classification with NB and SVM for android permissions malware dataset

Method	Input Dataset (MB)	After dimensionality reduction (MB)	NB	SVM
GA	500 MB	460	80.90%	85.30%
ACO		440	82.10%	86.20%
PCO		410	84.50%	85.40%
ACO-GA		398	86.70%	89.80%
Proposed		**360**	**90.70%**	**92.50%**

The above Table 1 describes the dimensionality reduction results from various Optimisation algorithms on the android Malware data set. After the processing of dimensionality reduction, validation has done with supervised classification algorithms that provide accuracy for each by using their bias and support vector machine.

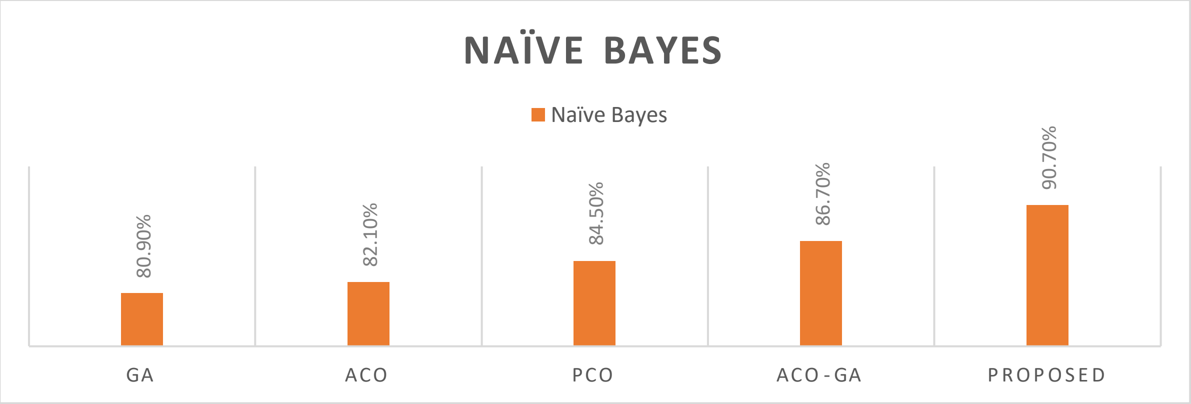

Figure 2: classification accuracy with various dimensionality reduction optimization techniques with Naïve Bayes classifier

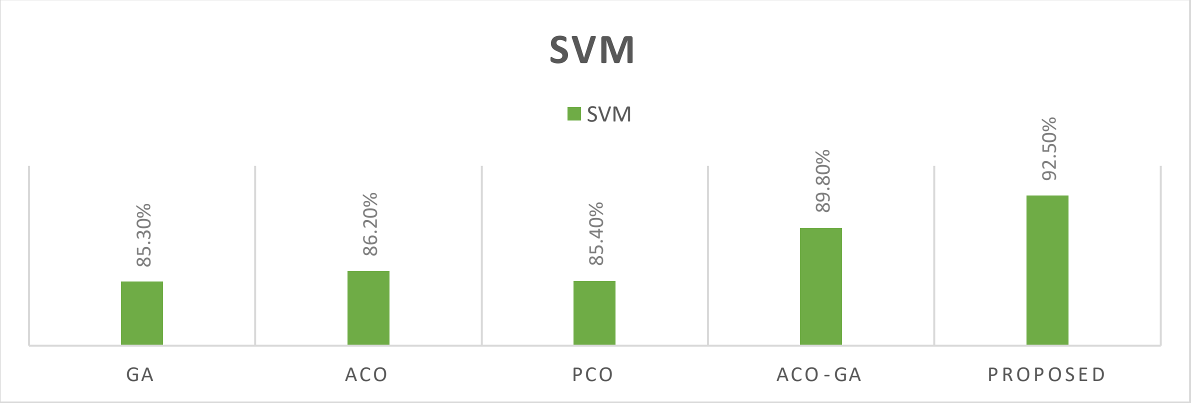

Figure 2: classification accuracy with various dimensionality reduction optimization techniques with Support Vector Machine classifier

The above figure 2 and figure 3 e describe dimensionality reduction techniques with numerous Optimisation algorithms. The genetic algorithm is the primary optimisation technique that provides lower accuracy after validating the supervised classification algorithm. The proposed hybrid algorithm gives higher precision than the traditional optimisation techniques. An extensive experimental analysis and evaluation have been done with Weka 3.7 open-source machine learning environment. The Naive Bayes gives 90.70% accuracy, with SVM producing 92.50% classification accuracy for the android malware dataset.

5. Conclusion

The goal of this study is to illustrate the various characteristics of dimensionality reduction and feature selection techniques. We may deduce that the collection of features formed through dimensionality reduction must be a subset of the original set of features, and vice versa. Actual characteristics are worked on in feature selection, but attributes based on variance are left intact, however in dimensionality reduction, we create new dimensions based on co-variances. Although both approaches reduce the number of features in a dataset, there is a significant distinction between them: feature selection simply picks and eliminates a set of features without changing them, whereas dimensionality reduction turns features into a secondary dimension. The study demonstrates how dimensionality reduction may be used to benefit different types of research, such as genetic research, since the data collected during these studies is massive, and acquiring potential and useful data is deemed vital for the research to continue. The proposed HGA produces 92.50% accuracy with SVM classifier, whch is better than traditionally feature extraction and classification approaches. To implement a various hybrid machine learning algorithm for checking the accuracy with various dataset and validate the space and time complexity etc.

References

[1] Mandikal Vikram, Rakesh Pavan, Navadiya Dhruvikkumar Dineshbhai and Biju Mohan. “Performance Evaluation of Dimensionality Reduction Techniques on High Dimensional Data”, 2019, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI), IEEE

[2] Faxian Cao, Zhijing Yang, Xiaobin Hong and Yongqiang Cheng. “Supervised Dimensionality Reduction of Hyperspectral Imagery Via Local and Global Sparse Representation”, 2021, Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE

[3] E. Myasnikov. “Using UMAP for Dimensionality Reduction of Hyperspectral Data”, 2020, International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), IEEE

[4] Pouria Fewzee and Fakhri Karray. “Dimensionality Reduction for Emotional Speech Recognition”, 2012, International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, IEEE

[5] Feiping Nie, Zheng Wang, Rong Wang and Xuelong Li. “Adaptive Local Embedding Learning for Semi-supervised Dimensionality Reduction”, 2020, IEEE

[6] Jingyu Wang, Fangyuan Xie, Feiping Nie and Xuelong Li. “Unsupervised Adaptive Embedding for Dimensionality Reduction”, 2021, Transaction on Neural Networks and Learning Systems, IEEE

[7] Suchismita Das and Nikhil R. Pal. “Nonlinear Dimensionality Reduction for Data Visualization: An Unsupervised Fuzzy Rule-based Approach”, 2021, Transaction on Fuzzy Systems, IEEE

[8] Lei Xu, Chunxiao Jiang and Yong Ren. “Deep Learning in Exploring Semantic Relatedness for Microblog Dimensionality Reduction”, 2015, Global Conference on Signal and Information Processing (GlobalSIP), IEEE

[9] Yongyan Zhang, Guo Xie, Wenqing Wang, Xiaofan Wang, Fucai Qian, Xulong Du and Jinhua Du. “Distributed dimensionality reduction of industrial data based on clustering”, 2018, IEEE

[10] Gladys M. Hilasaca and Fernando V. Paulovich. “User-guided Dimensionality Reduction Ensembles”, 2019, 23rd International Conference Information Visualisation (IV), IEEE

[11] Chenfeng Guo and Dongrui Wu. “Feature Dimensionality Reduction for Video Affect Classification: A Comparative Study”, 2018, First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE

[12] Hieu Minh Bui, Margaret Lech, Eva Cheng, Katrina Neville, Richardt Wilkinson and Ian S. Burnett. “Randomized Dimensionality Reduction of Deep Network Features for Image Object Recognition”, 2018, Randomized Dimensionality Reduction of Deep Network Features for Image Object Recognition, IEEE

[13] Yanlong Wang, Jinhua Liu and Ting Zhang. “Framework of Multiple-point Statistical Simulation Using Manifold Learning for the Dimensionality Reduction of Patterns”, 2020, 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), IEEE

[14] Remzi Oten and Rui J. P. de Figueiredo. “Topological Dimensionality Determination and Dimensionality

Reduction Based on Minimum Spanning Trees", 1998, IEEE

[15] Minfeng Zhu, Wei Chen, Yuanzhe Hu, Yuxuan Hou, Liangjun Liu and Kaiyuan Zhang. "DRGraph: An Efficient Graph Layout Algorithm for Large-scale Graphs by Dimensionality Reduction", 2021, Transactions on Visualisation and Computer Graphics, IEEE

[16] GuangHui Yan, LiSong Liu, LinNa Du and XiaXia Yang. "Unsupervised Sequential Forward Dimensionality Reduction Based On Fractal", 2008, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE

[17] Zhao Shenglin and Zhu Shan-an. "Face Recognition by LLE Dimensionality Reduction", 2011, Fourth International Conference on Intelligent Computation Technology and Automation, IEEE

[18] Dewa Made Sri Arsa, Grafika Jati, Aprinaldi Jasa Mantau and Ito Wasito. "Dimensionality Reduction Using Deep Belief Network in Big Data Case Study: Hyperspectral Image Classification", 2016, IEEE

[19] D Lakshmi Padmaja and Dr.B Vishnuvardhan. "Comparative Study of Feature Subset Selection Methods for Dimensionality Reduction on Scientific Data", 2016, 6th International Conference on Advance Computing, IEEE

[20] Jiajia Shu, Weiming Liu, Fang Meng and Yichun Zhang. "Multi-feature Image Retrieval by Nonlinear Dimensionality Reduction", 2014, Seventh International Symposium on Computational Intelligence and Design, IEEE

[21] Benson S. Y. Lam and Hong Yan. "A Novel Dimensionality Reduction Method for Pattern Classification", 2007, IEEE

[22] Sivaranjani S, Ananya S, Aravinth J and Karthika R. "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction", 2021, 7th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE

[23] S. Keller, A. C. Braun, S. Hinz and M. Weinmann. "Investigation of the Impact of Dimensionality Reduction and Feature Selection on the Classification of hyper spectral Enmap Data", 2020, IEEE

[24] Huu-Thanh Duong and Vinh Truong Hoang. "Dimensionality Reduction Based on Feature Selection for Rice Varieties Recognition", 2019, 4th International Conference on Information Technology (InCIT), IEEE

[25] Joshi Snehal K. and Sahista Machchhar. "An Evolution and Evaluation of Dimensionality Reduction TechniquesA Comparative Study", 2014, IEEE

[26] Afnan M. Alhassan and Wan Mohd Nazmee Wan Zainon. "Review of Feature Selection, Dimensionality Reduction and Classification for Chronic Disease Diagnosis", 2021, IEEE

[27] G. Suresh Reddy. "Dimensionality Reduction Approach for High Dimensional Text Documents", 2016, IEEE

[28] Shahana A H and Preeja V. "Survey on Feature Subset Selection for High Dimensional Data", 2016, International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE

[29] Harchli Fidae, Es-safi Abdelatif and Ettaouil Mohamed. "Original Approach for Reduction of High Dimensionality In unsupervised learning", 2020, IEEE

[30] K.Pavya and B.Srinivasan. "Feature Selection Algorithms to Improve Thyroid Disease Diagnosis", 2017, International Conference on Innovations in Green Energy and Healthcare Technologies (ICIGEHT'17), IEEE

[31] Tian Lan, Deniz Erdogmus, Lois Black and Jan Van Santen. "A Comparison of Different Dimensionality Reduction and Feature Selection Methods for Single Trial ERP Detection", 2010, 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, IEEE

[32] Muhammad Abu Bakar Siddik, Md. Mahfuzul Islam Mazumder, Rakibul Alam and Musharrat Khan. "Performance Comparison Between Dimension Reduction and Feature Selection Approaches for Data Classification", 2021, Fifth International Conference on Computing Methodologies and Communication (ICCMC, IEEE

[33] Zacharias Voulgaris and George D. Magoulas. "Dimensionality Reduction for Feature and Pattern Selection in Classification Problems", 2008, The Third International Multi-Conference on Computing in the Global Information Technology, IEEE

[34] Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney and Petros Drineas. "Randomized Dimensionality Reduction for k-means Clustering", 2014, Transactions on Information Theory, IEEE

[35] Jie Zhu and Jun Shi. "Hessian Regularization Based Semi-Supervised Dimensionality Reduction for Neuroimaging Data of Alzheimer's Disease", 2014, IEEE

[36] R.Kavitha and E.Kannan. "An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining", 2020, IEEE

[37] C.Deisy, B.Subbulakshmi, Dr. S.Baskar and Dr.N.Ramaraj. "Efficient Dimensionality Reduction Approaches for Feature Selection", 2007,

International Conference on Computational Intelligence and Multimedia Applications, IEEE

[38] Luis H. Brito, Ana L. C. V. Lara, Luis E. Zárate and Cristiane N. Nobre. "Improving the quality of enzyme prediction by using feature selection and dimensionality reduction", 2019, International Joint Conference on Neural Networks, (IJCNN), Budapest, Hungary, IEEE

[39] hraddha Sarode and Jayant Gadge. "Hybrid Dimensionality Reduction Approach for Web Page Classification", 2015, International Conference on Communication, Information & Computing Technology (ICCICT), IEEE

[40] Rachid Benmokhtar, Jonathan Delhumeau and Philippe-Henri Gosselin. "Efficient Supervised Dimensionality Reduction for Image categorization", 2013, IEEE