International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

Data-driven based Optimal Feature Selection Algorithm using Ensemble Techniques for Classification

Jayshree Ghorpade[1*], Balwant Sonkamble[2]
[1]Research Scholar, P.I.C.T., S.P.P.U., Pune
M.I.T.W.P.U., Pune, Maharashtra, India
jayshree.aherghorpade19@gmail.com
[2]P.I.C.T., S.P.P.U., Pune, Maharashtra, India

Abstract—The shift in paradigm with advanced Machine Learning algorithms will help to face the challenges such as computational power, training time, and algorithmic stability. The individual feature selection techniques, hardly give the appropriate feature subsets, that might be vulnerable to the variations induced at the input data and thus led to wrong conclusions. An expedient technique should be designed for approximating the feature relevance to improve the performance for the data. Unlike the prevailing techniques, the novelty of the proposed Data-driven based Optimal Feature Selection (DOFS) algorithm is the optimal k-value 'k_f' determined by the data for effective feature selection that minimizes the computational complexity and expands the prediction power using the gradient descent method. The experimental analysis of proposed algorithm is demonstarted with ensemble techniques for the non-communicable disease such as diabetes mellitus dataset produces an accuracy of 80.80%, whereas comparative performance analysis for benchmark dataset depicts the improved accuracy of 86.03%.

Keywords- Machine Learning; Ensemble; Genetic Algorithm; Gradient descent; Diabetes.

I. INTRODUCTION

The pandemic has experienced considerable amount of rise in digital data, which is arriving from different sources. Exploring the data with appropriate processing and recovering the precise information at proper timings will help to discover the valuable insights for the processes. One of the major concerns identified for exploring the heterogeneous data is the challenge to handle the heterogeneity and variety of data. It has diverse complex relationships within the multi-source heterogeneous data as related to homogeneous data. Ding et. al. discusses the existing conventional techniques that scarcely extract the correct knowledge with data processing and optimization [2]. As the dimensions of the data rise, the conventional techniques need to be executed frequently for selecting the effective features. A multiple relevant feature ensemble selection (MRFES) algorithm is suggested that depends on the multilayer co-evolutionary consensus MapReduce (MCCM). But, such tasks involve the increase in computational overhead with resource management. Most of the Machine Learning models are explored to solve the real-world feature selection problems, as it is the inseparable part of data processing. The reduced features attempt to increase the class separability of the data for a classification problem. Analyzing this multi-source data intensive scientific real-time applications triggers the formation of new values and insights to identify the innovative hidden patterns and beneficial information [3]. The reasonable use of processed multi-source data not only updates the practical value of the system but also raises the importance of the application with data formatting techniques and system interface designs for public safety management [4].

The supervised feature selection techniques uses the response variable and removes the irrelevant features. The filter techniques explore the relationship of the features with the response variable and select the most relevant features as per their importance. The intrinsic methods form the feature subset during training. Even though, the good features are selected, most of the time, the leading algorithms too malfunction for complex data relations with high dimensions. The use of existing methods demands for increased space and computational power.

The ensemble approach allows the use of multiple classification techniques to enhance the working of the model. The proposed method has a data driven approach that helps to understand and explore the data with significant features. The processing of data to eliminate the extraneous features, handle redundancy, and remove the noisy attributes will assist to avoid erroneous predictions. The Machine Learning model may be generalized with better selection of univariate techniques such as Information Gain (IG) and Chi-square, which removes the unwanted features based on statistical analysis. The appropriate feature selection process improvises the performance of the model with reliable feature extraction. Even small change in input data affects the processing of algorithms. The boosting

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

algorithms like Gradient Boosting and Extreme Gradient Boosting refines the performance of the model by boosting the working of weak learners to strong learners. Random Forest with Decision Trees as its base classifier acts as a meta estimator for various sub-samples of the dataset.

The class imbalance challenge hinders the performance of the model as most of the techniques are biased towards the majority class. Recognizing the relevant features from the difficult data factors is a complex problem. Bader-El-Den et. al. presents an ensemble-based method that deals with the class imbalance problem by oversampling the classification ensemble through growing the number of classifiers that represent the minority class in the ensemble [5]. Zohaib Jan et. al. creates an optimized ensemble classifier that improves the classification accuracy along with a lesser component size. Multiple classifiers are referred to solve the combinatorial problem of optimization by incorporating the evolutionary techniques too [6]. The significance of the proposed research is determining the k-value 'k_f' which updates dynamically with the change in number of features as input to the model. The optimization algorithm such as gradient descent helps to determine the local optimum while fetching the parameters of the Machine Learning models. The iterative processing obtains the gradient of the objective function that minimizes the total iterations during processing and thus manage the variation in the data [7]. Further, the Genetic Algorithm with its stochastic behavior is applied to the top-k features to generate the efficient solution along with the fitness function.

The data with diverse features is an interesting and provocative problem that has given a research direction to have new developments in existing techniques for data processing. The heterogeneous data with disparate attributes or features or characteristics needs to be studied to settle the challenges across data computing. Multiple prediction techniques do not allow the direct analysis of heterogeneous data and hence the data transformations are often required for examining such data. The data intensive computing will explore the advanced processing techniques and solutions, which will help to establish the relation amongst various types of features to leverage the potential power of the data. Section II outlines the previous research work. The research study of the proposed algorithm is represented in section III. The experimental analysis and its illustrations are shown in section IV. Lastly, the concluding remarks are given in the conclusion section.

II. PREVIOUS RESEARCH WORK

Data analytics studies the data at its miniature level to explore meaningful data aspects and establish new insights for the technical and systematic discoveries. It acts as an essential facet of computing, that is driven by data with the diverse features and dimensions of the multi-source data. The learning algorithmic techniques along with emerging technologies are generating exponential data. The value of data diminishes since the abstruse records are difficult to understand, presence of noisy data, unessential data, inconsistencies in data, missing values or data observations, etc. makes the data imperfect for further processing, which may lead to vague inferences. Extracting the good quality data and integrating it poses a challenge for data processing as the dissimilar data comes from different origins, with globalization. Modelling the data with effective algorithmic techniques, guides to enhance the performance & working of the model with cost reduction solutions. The ensemble learning [8], strategically combines the predictions and expected results from different learning algorithms to gain better overall performance than could be acquired from any of the individual component learning algorithms. This approach establishes a prevailing state-of-the-art tactics that tries to attain maximum performance.

StackGenVis [9] referred as 'stacked generalization' is the ensemble-based visualization model that selects effective features with diverse techniques and meta-model. The unwanted under-performing or overpromising models are minimized or eliminated to boost the performance of the model by shrinking the intricacy of the subsequent stack. The expected behavior of model with good results is amended by using ensemble methods. But, to develop the stack of models is an unwieldy trial-and-error procedure. The varied arrangements of data instances along with features pose difficulties for the individual techniques using diverse parameter-models. The systematic analysis of efficient Machine Learning techniques is described by Omar Y. et. al. [10] for handling the feature selection problems with the heterogeneous data. The data-intensive structural designs demand for efficient data modeling with sustainable computational techniques. The data mining techniques with dynamic ensemble selection method for heterogeneous data have been discussed by Barbara [11]. Choosing the right subset of relevant traits or features is crucial for the success of a model.

Additional investigations have spotlighted that the existing algorithms are lacking in stability, where the robustness is affected with updations in the input data. Thus, it is a concern as these relevant feature subsets are examined to achieve knowledge with new understandings and learnings. The computational intelligence should be used to address the challenges of data variety. The real-time data collected from multiple sources has large number of features as compared to the observations and hence raise difficulties during training process of a model. Zhiwen Yu et. al. designed an adaptive semi-supervised classifier ensemble framework (ASCE) that constituted the adaptive feature choosing practice with an adaptive weighting process (AWP) and an auxiliary training set generation process (ATSGP). A group of compact subspaces is generated depending on the chosen attributes whereas the AWP connects each basic selector with a certain weight value

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

[12]. The recent technical advancements are witnessing an increase in the application of Machine Learning modeling and statistical methods those are effective in experimenting the nonlinear relationships in the data. The biological trait referred as gene is one of the fundamental parts. Gene selection acts as a combinatorial search problem that can be steered with appropriate optimization techniques. The Memetic Cellular Genetic Algorithm (MCGA) is suggested for the cancer microarray datasets with relevant feature extraction [13].

Most of the data derived from different origins is heterogeneous in nature with a greater number of attributes or features or traits. The digitized data involves various features with complex relationships. One of such an example is the medical data for disease prediction. The relevant feature subset needs to be extracted with suitable feasibility, so that further it will contribute to appropriate decision making. Overfitting of data occurs when the fitted model gives best results with training data, but fails to have fine generalization for the test data. Other way when the model does not respond properly with train as well as test data, then the underfitting of data is seen in the model. The variation induced at input data level, affects the data processing and model working. The subgroup of unique attributes or independent features needs to be explored for enhancing the working of the model. Processing these relevant features reduces the training time for the model and minimizes the resource utilization. Different Machine Learning algorithms explore applications with enormous features. But, to determine the most relevant features for a predictive model, there is a need of domain knowledge. Thus, the data-intensive model should have some data-driven approach for automatic feature selection process. The relevant feature subset selection helps to explore the data and minimize the computation requirements.

The traditional methods to select the best features from the entire dataset, requires more computations and processing time. The univariate techniques perform the analysis with single variable in the simplest form by determining the significance of each variable or feature or trait with the corresponding response variable. The Chi-square scores gives the most probable features that establish the patterns and relations within the feature set. The Information Gain is the measure with maximum amount of information an attribute relates for the class. The reduction in impurity or entropy results in raising the Information Gain and thus determines the effectiveness of a trait or feature, while classifying the data. The dependencies between the features are analysed and the relational complexities amongst them are studied to speed up the processing of training data.

The significant variables/features possess higher score. The boosting algorithms use the decision-making ability to find the appropriate features with their relevance. The multivariate techniques analyse the comparative importance of a feature, since its use in building key decisions increases. The Extreme Gradient Boosting algorithm prevents the model from overfitting by implementing the regression techniques [(L1 (Lasso) / L2(Ridge)]. Random Forests is one of the most used Machine Learning methods that is known for its robustness and relatively better accuracy. In Random Forests, the impurity decline from each independent variable can be averaged across trees to know relative significance of variable/feature. Random Forest progresses upon the inadequate robustness and trivial execution of decision trees. The Kernel Factory as well as AdaBoost techniques fasten the performance of the single classifier models. Logistic Regression helps to expand the progress of the model with reduction in variance of the expected values.

The organisms' genes tend to progress over a period and produce generations to better adapt to the surroundings for the survival. Genetic technique is a heuristic optimization algorithm inspired by the natural evolution actions. The algorithmic tasks operate on the individuals of a population to yield healthier approximations. The Genetic Algorithm is a technique based on evolutionary algorithm that adapts with the natural genetic process to determine the best optimal solution. The objective function or fitness function such as accuracy scores is an optimization standard, which checks the performance of the model. This function depicts the predictive model's generalization execution that represents the error term on extracted instances. The features with large fitness values are considered to be better. The different operators like crossover, mutation, etc. are vital mechanisms that perform the adaptation process to select the best possible feature subset and generate the genetic diversity [14]. The vigorous amalgamation amongst these operators helps for the successful search of the Genetic Algorithm. The crossover probability tracks the likelihood for the crossover to happen amongst the two predecessors. If its value is less towards ‘0’ then the new generations will be alike their predecessor. Whereas, if its value is more towards ‘1’ then the resultant would be a faster convergence with loss of diversity in the next generations which will hinder the performance of the algorithmic technique. Therefore, an optimal value of the crossover probability needs to be determined. The mutation probability regulates the likelihood for the mutation process that is applied to an individual in a population. Its least value is preferable as there will be very few mutations in each generations. But its greater value will increase the possibility of getting stuck in a suboptimal solution. The dynamic nature of the projected technique allows the linear change in the ratios of mutation and crossover as the search progresses [15]. The skill to capture pertinent information from data has become a relatively crucial issue. Appropriate knowledge must be explored from the underlying information to obtain feasible research outcome [16].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

As the number of features are raised in the problem domains with the advancements, there is a need for profound knowledge of the problem domain revealing the most relevant attributes for further ease of data-modeling. Clustering techniques are used to pick the better features for the set of data instances with a dynamic mode and then these features are applied for the classifications [18]. The performance metrics help to analyse the proper working of the model.

The state-of-art study and research gap identified demands for an efficient technique that will estimate the feature relevance and accord towards the expected predictions. Necessity for the appropriate sampling techniques to persuade even small changes and adapt to the variations in data sub-sets for designing the stable technique. Need of a data-driven approach that will help to generate the effective 'k-value' to determine the best features. Collaborating the performance of multiple base selectors will ensure the robustness of the technique, as single base selectors hardly produce the good results. The ensemble paradigm has been encouraged as a talented framework for enlightening the robustness of the model, where the abstraction of steady feature subgroups is intrinsically challenging [17]. The elementary thought is to cooperatively exploit the strengths of unlike selectors along with overcoming the weaknesses of these selectors.

III. The Proposed Data-driven Optimal FS Algorithm

The processing of a model can be enhanced with appropriate techniques and pertinent features, where every byte of the data should be explored at minute level. The research study analyses and understands the different feature selection techniques that will select the best features and improve the outcome of the model. Ensemble learning includes various classifiers or techniques to select the better features for the model.

A. Pre-processing of Data and Feature Engineering

The irrelevant attributes in the dataset should be separated or dropped so that there's no need for the model to explore and understand lots of features, which may result in overfitting problem. Many of the Machine Learning models deal with heterogeneous data which has numerical or categorical features. The dimensionality reduction is an important task for data processing. Eliminating the features with low variance will minimize the dimensions of the data. Thus, less data with significant relevant features will drop off the complexity of the model and fasten the process of training. Fig.1 depicts the design of the proposed technique to select the best effective features to enrich the performance of the model. Data pre-processing is an important part of the initial stages of processing data, where the imputer methods such as Simple imputer can be used to tackle the missing data values. It explores the data with different statistical techniques. The categorical, nominal and numerical variables are processed with data transformations [19]. Data standardization rescales the data with proper data distribution, whereas normalization helps to have a unit variance that simplifies the data for further experimentation.

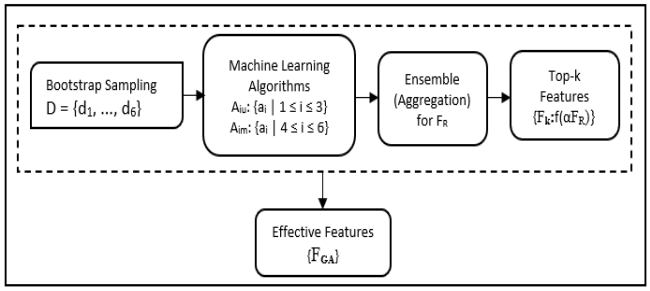

Figure 1. Proposed Technique

The varying ways to select the significant features with ensemble approach will fasten the data processing. The ensemble [20] method helps to model the effectiveness of the required features with diverse algorithms and thus strengthen the working of the model. The multi-source disparate data [21] should be analysed to overcome the issues across data modeling.

B. The Proposed Technique-Algorithm

The projected approach aims at determining and selecting the relevant features which are reliable and robust. The features are ranked as per their importance nevertheless still there is no accord on the number of feature cut-off. Hence, it is a vital modeling parameter to recognize a threshold value of the features that will showcase the most relevant ones and eliminate the redundant features. The study determines the threshold benchmark for feature ranking techniques [22]. The proposed algorithmic technique known as Data-driven based Optimal Feature Selection (DOFS) is an enhancement to 'Ensemble Bootstrap Genetic Algorithm' (EnBGA) [1]. The emphasis of the research enrichment is developing a new basis for the optimal value 'k_f' that can minimise the glitches of stability and threshold of features.

The proposed research method imbibes different feature selectors to process the bootstrap sampled data. Tuning the parameters of the techniques helps to identify the 'best' threshold values. The data transformations and data aggregation techniques generates the data-driven efficient feature subset. The GA algorithm starts randomly with the population in the dataset, performs selection, crossover and modifications for the next generation data. Thus, the effective features are emerged with selection over time by adapting the natural genetic principle. The efficacy of the GA depends on the range of few of the control mechanisms and its parameters such as mutation induced, crossover allowed and population search space.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

- ***Determine data-driven optimal 'k_f'***

The optimal data-driven *k*-value is derived for selecting the top '*k*' features with an extension update to the proposed research work [1] by minimising the objective function with gradient optimisation. Top-k features 'F_k' are determined by processing the input data 'D' having total features 'F_N' with six bootstrap ($d_x \in D$) samples. The reduced features 'F_R' is generated after eliminating the low variance features with $F_R \subseteq F_N$.

The feature scores are evaluated to rank the reduced features 'F_R' as shown in (1) based on the significance score values.

$$F_k = \{ f(\alpha . F_R) \} \quad (1)$$

The top k-features 'F_k' are the subset of reduced features that are determined by exploring the optimization parameter 'α' to rationalize the objective function.

Algorithm: Data-driven optimal value 'k_f'
Input: data 'D'
Output: optimal 'k_f' feature subset as 'F_k'

1: $D \leftarrow \{d_x \mid 0 < x < 7\}$
2: for $d_x \leftarrow d_1$ *to* d_6
$A_{iu} \leftarrow$ univariate algorithms $\{a_i \mid 1 \le i \le 3\}$
$d_{xi} \leftarrow [(X_i), (y_i)]$
$F_{j(s)} \leftarrow$ *[feature-scores* (F_N)*]*
$A_{im} \leftarrow$ multivariate algorithms $\{a_i \mid 4 \le i \le 6\}$
$d_{xi} \leftarrow [(X_i), (y_i)]$
$F_{j(s)} \leftarrow$ *[feature-scores* (F_N)*]*
3: Ranking (selected features) $\leftarrow$ *sorting* $[F_{j(s)}]$
4: Data-driven optimal k-value for F_R
5: $n_p \leftarrow \{n_p \mid 1 < p < F_R\}$
6: $TIs \leftarrow \{ n_p + 2\frac{F_R}{n_p} + 1 \}$
7: $k_p \leftarrow f(\frac{F_R}{n_p})$
8: $klist \leftarrow \{k_1, k_2, \ldots, k_R\}$
9: for each $k_k \in klist$
$S_k \leftarrow$ *score* $[(X_{rank}), (y_{rank})]$
$k_{k_t} \leftarrow max\ [S_k]$
10: Evaluate scores $\leftarrow$ features
{*Range* $[(k_t\text{-}k_p), (k_t\text{+}k_p)]$}
11: Optimal $k_f \leftarrow max$ {*scores in Step-11*}
12: $F_k \leftarrow$ *top-k* optimal feature subset

The data-driven k-interval pace(step) 'k_p' is the single decision variable, which is determined based on the reduced feature sub-set while minimizing the objective function for the total iterations 'TIs' with number of partitions 'n_p' within 'F_R' as shown in (2). An iterative first-order optimisation algorithm known as 'gradient descent' guides to determine a local minimum or maximum of a given function.

$$TIs = \{ n_p + 2\frac{F_R}{n_p} + 1 \} \quad (2)$$

The gradient optimisation as formulated in (3) helps to determine the approximate value of 'k_p'. The 'TIs' with optimal value of 'k_p' will thus reduce the processing time as needed for execution of all the features.

$$k_p = \{ f(\frac{F_R}{n_p}) \} \quad (3)$$

Accordingly, the k-interval list 'klist' is obtained and the scores are evaluated to find the temporary value k_{k_t}.

The k-interval list is more time effective as compared to the sequential single incremental step of the feature sub-set. Further the feature scores are analysed with the features ranging from values $(k_t - k_p)$ *to* $(k_t + k_p)$. Thus, the proposed optimal feature selection algorithm for multi-source heterogeneous data using Machine Learning techniques helps to acquire the final *k*-value and determine the top-k features 'F_k' with optimal 'k_f'.

- ***Extract effective F_{GA}***

Significant features extraction with more relevance to the expected outcome is a combinatorial optimization problem, that can be minimised with stochastic selection technique such as Genetic Algorithm. Extracting the effective features from top 'k' features will assist to improve the performance of the model. The selected top k-features 'F_k' are fed to the next phase of the proposed algorithm for extracting the effective features using Genetic Algorithm (GA) F_{GA}.

$P_{(n)}$ constitute the population {individual $p_i \mid 0 < i \le n$} and $FF_{(n)}$ establishes objective {fitness function $ff_i \mid 0 < i \le n$} function, that evaluates the effective feature subset [$F_{GA} \subseteq F_k$].

IV. Experimental Analysis

Most of the Machine Learning models are applied for medical diagnosis. There are various techniques those are used to analyse the complex relationships [23] of the dataset. The

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

multi-source heterogeneous data is captured and explored to select the best effective features. The proposed enhanced Data-driven based Optimal Feature Selection (DOFS) algorithm with data-driven k-value is studied and implemented for online medical dataset in the form of Electronic Health Records, that represent the electronic information of a patient to improve the quality of care and health safety. This online medical information assist decision-makings & augments the research outcomes [24]. The World Health Organisation has recognised the research work with its significance towards healthcare domain [25]. The dataset consists of multiple sources such as *Medication* data, *Diagnosis* data and *Transcript* data. The *Diagnosis* data is experimented with the univariate & multivariate algorithms to obtain the reduced feature subset 'F_R' and its ranking. The features with least rank have most information to contribute towards the predictions.

The 'k-interval' feature-set analysis for the *Diagnosis* data (International Classification of Diseases, 9th revision) with its scores is represented in Table-I. The high sensitivity gives good true positive ratio. The precision score with 98.87570 implies that the proposed model has correctly classified the observations having high risk in the high-risk category. The data-driven value of 'k_p' is determined to be '4' by using the proposed enhanced technique for the algorithm with gradient descent functionality to minimise the iterations and fasten the data processing. The analysis for selected subset of feature-set is performed using the metrics as accuracy, sensitivity, specificity and precision, with True Positive, False Positive, True Negative, and False Negative parameters of the metrics. These metrics study illustrates that the feature subset with top-k(5) features yields better result.

The accuracy is based on the predicted classes, whereas the Receiver Operating Characteristic (ROC) AUC for the selected features is based on the predicted scores and thus helps for tuning the model parameters.

TABLE I. FEATURE SET ANALYSIS OF F_R

Features#	Analysis			
	Accuracy	***Sensitivity***	***Specificity***	***Precision***
Top5	80.80402	80.87569	33.33333	99.87570
Top9	79.64824	81.35417	32.85714	97.07893
Top13	78.34171	81.90683	32.63889	93.97141
Top17	77.23618	82.25446	31.81818	91.60970
Top21	75.42714	82.55814	30.00000	88.25357
Top25	77.38693	82.94486	35.06494	90.67744
Top28	77.83920	82.95711	36.23853	91.36109

The ROC Area Under Curve (AUC) plot with False Positive Rate (FPR) versus True Positive Rate (TPR) for the analysed feature sets (F_R) of Table-I is shown in Fig. 2.

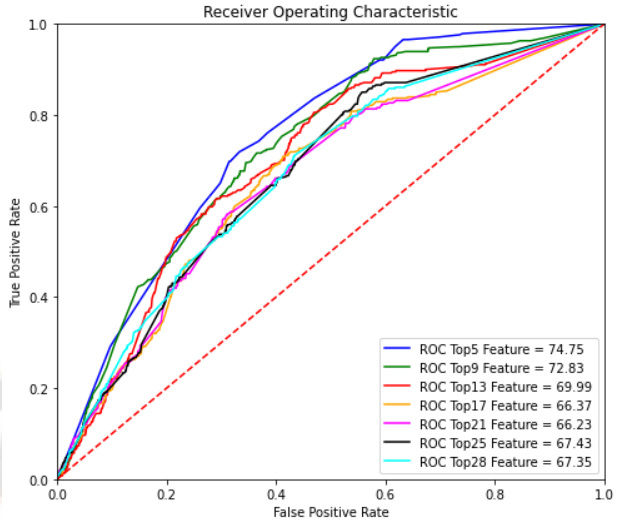

Figure 2. Receiver Operating Characteristic plot (FPR vs TPR) for F_R

Further, the performance analysis for determining optimal *k*-value as discussed in the proposed algorithm is characterised in Table-II and the processing will be executed within the range: (k_t + k_p) as shown in Fig. 3.

TABLE II. FEATURE SET ANALYSIS WITHIN K_P

Features#	Analysis			
	Accuracy	***Sensitivity***	***Specificity***	***Precision***
Top5	80.80402	80.87569	33.33333	99.87570
Top6	80.50251	81.32376	41.46342	98.50839
Top7	80.30150	81.34982	38.77551	98.13549
Top8	79.79899	81.44867	35.21127	97.14108
Top9	79.64824	81.35417	32.85714	97.07893

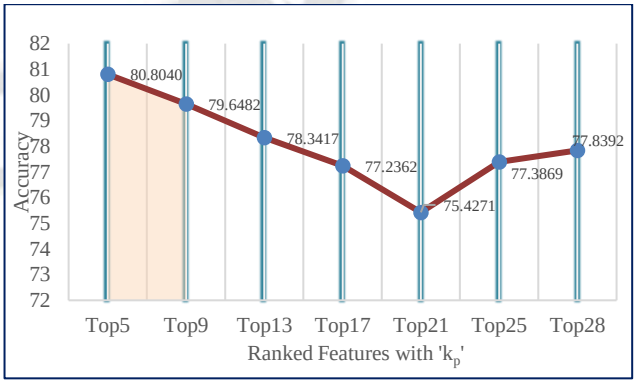

Figure 3. Ranked Features with derived k_p

The more robust ability of the model towards the final data-driven top-k(5) features with 'k_f' is depicted in Fig.4.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

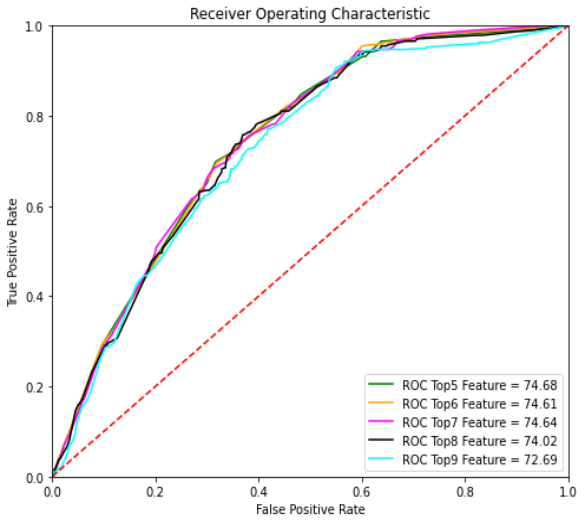

Figure 4. Receiver Operating Characteristic plot (FPR vs TPR) for k_f within (k_t+k_p)

Fig.5 shows the optimal top-k(5) features performance with improved accuracy.

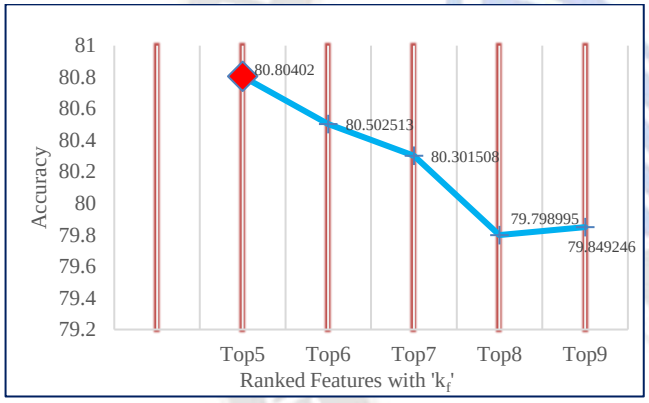

Figure 5. Ranked Features with derived optimal k_f

The proposed algorithm assists with the gradient function to determine the optimal 'k_f' for effective features and sound model performance that plays an important role in the healthcare domain with societal significance. The experimental analysis and implementation of the DOFS algorithm on the *Diagnosis* data, extracted the effective five features. The patients having these disorders/features are more susceptible to have diabetes in future. As per the analysis of the American Diabetes Association (ADA), the cardiovascular disorders and diabetes are the major concerns for the increasing mortality rate in the Covid-19 pandemic [28]. The performance analysis of the final effective features 'F_{GA}' with data driven k-value and the normally selected top-k features is executed and analysed using Gradient Boost (GB) classifier. Thus, the model performance is improved further with Genetic Algorithm along with deriving the most effective and optimal four features GA_k(4).

Also, experimental study for the comparative analysis was performed for the benchmark data, Sonar dataset with sixty features, available on UCI Machine Learning Dataset [26]. Zhang et.al. [27] had used ensemble of classifiers with Genetic Algorithm (GA-Ensemble) for selecting features. The GA-Ensemble algorithm used the Machine Learning techniques like the standard multiple layer perceptron back propagation ANN classifier (GANN), Support Vector Machine classifier (GASVM) to test the feature set with twelve features. But the technique is vulnerable with small changes in data.

TABLE III. COMPARATIVE ANALYSIS OF MODEL PERFORMANCE

Machine Learning Techniques	Methods with Accuracy (%)	
	Ensemble of Classifiers	Proposed DOFS Algorithm
GANN	82.01	84.11
GASVM	79.82	86.59
GA-Ensemble	83.95	86.03

The proposed DOFS algorithm handles this challenge by considering the bootstrap technique to increase the robustness of the model. The reduced feature set simplifies the processing of the model and progresses its predictive power [30]. The experimentation is conducted for comparative analysis and the performance of the proposed system is studied with different classifiers as shown in Table-III. Few of the parameters of Genetic Algorithm are adjusted with initial population size as 100, the maximum features are now kept as the top-k features, mutation probability is set to 0.03, crossover probability is used as 0.7 with cv = 5 for cross-validation.

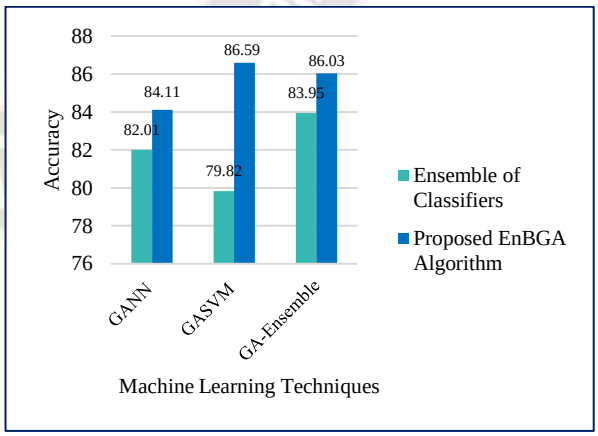

Figure 6. Comparative analysis of proposed Algorithm

An empirical analysis was done by implementing the proposed algorithm with optimal value of 'k_f' that produced the effective feature subset with thirteen features. The visualization

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

graph in Fig.6 designated the gains while comparing the proposed algorithmic technique with the existing methods.

The key findings offer a heterogeneous ensemble model that manages lacunas of the ensemble technique such as overfitting and concept drift, where the individual model finds difficulty in adjusting with the unexpected conditions [29].

V. CONCLUSION

Data driven techniques are exploring the new computing processes for heterogeneous data to enhance the efficacy of the Machine Learning models. Various features of data trigger the revolution of thinking with innovations to have proper predictions by extracting the relevant data. The analysis of the heterogeneous data using proposed algorithm, Data-driven based Optimal Feature Selection (DOFS) with data driven k-value 'k_f' determined by using the gradient optimization helps to select the effective features. The derived optimal value of single decision variable minimizes the objective function for the total iterations. The Receiver Operating Characteristics assists with the robustness of the model. The forecasting ability of the model is enhanced with new cognizance to produce actionable prognostics. In Future, the parameters used in performance metrics of Genetic Algorithm can be hyper-tuned and tested for variants of fitness function. The metaheuristic algorithm can perform better, when adapted with appropriate values of mutation and crossover operators. The variations in input data can be studied with different classifiers.

REFERENCES

[1] Ghorpade-Aher J., Sonkamble B., 2021, "Effective Feature Selection Using Ensemble Techniques and Genetic Algorithm," International Congress on Information and Communication Technology, *Lecture Notes in Networks and Systems, vol 236. Springer*, Singapore, doi.org/10.1007/978-981-16-2380-6_32.
* World *Health Organization*, Article in English | Scopus|ID:covidwho-1460284, Aug'2022, https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/pt /covidwho-1460284?lang=en

[2] W. Ding, C. Lin and W. Pedrycz, 2020, "Multiple Relevant Feature Ensemble Selection Based on Multilayer Co-Evolutionary Consensus MapReduce," *IEEE Transactions on Cybernetics*, vol. 50, no. 2, pp. 425-439.

[3] J. Ghorpade and B. Sonkamble, 2020, "Predictive Analysis of Heterogeneous Data Techniques & Tools," 5th *International Conference on Computer and Communication Systems*, Shanghai, China, pp. 40-44, World Cat: OCLC Number-1175635436, IEEE Press, Piscataway, NJ, US.

[4] X. Yu and Q. Wu, 2021, "Multi-source Heterogeneous Data Association Technology to Build Public Safety Big Data Integration Research," *International Conference on Big Data Economy and Information Management (BDEIM)*, doi: 10.1109/BDEIM52318.2020.00012, pp. 17-20.

[5] M. Bader-El-Den, E. Teitei and T. Perry, 2019, "Biased Random Forest For Dealing With the Class Imbalance Problem," in IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 7, pp. 2163-2172, doi: 10.1109/TNNLS.2018.2878400.

[6] M. Z. Jan, J. C. Munoz and M. A. Ali, 2020,"A novel method for creating an optimized ensemble classifier by introducing cluster size reduction and diversity," in *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2020.3025173.

[7] X. Wang, L. Yan and Q. Zhang, 2021, "Research on the Application of Gradient Descent Algorithm in Machine Learning," 2021 *International Conference on Computer Network, Electronic and Automation (ICCNEA)*, pp. 11-15, doi: 10.1109/ICCNEA53019.2021.00014.

[8] Panagiotis Pintelas and Ioannis E. Livieris, 2020, "Special Issue on Ensemble Learning and Applications", *MDPI Scopus journal, Algorithms'20*, 13, 140; doi:10.3390/a13060140, pp.1-4.

[9] A. Chatzimparmpas, R. M. Martins, K. Kucher and A. Kerren, 2021, "StackGenVis: Alignment of Data, Algorithms, and Models for Stacking Ensemble Learning Using Performance Metrics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1547-1557.

[10] Omar Y.Al-Jarrah, Paul D.Yoob, Sami Muhaidat, George K. Karagiannidis, KamalTaha, 2015, "Efficient Machine Learning for Big Data: A Review," *Elsevier, Science Direct, Big Data Research*, vol. 2, pp. 87-93.

[11] Barbara Pes., 2020, "Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains, " Neural Computing & Applications, vol. 32, pp.5951–5973, https://doi.org/10.1007/s00521-019-04082-3.

[12] Z. Yu et al., 2019, "Adaptive Semi-Supervised Classifier Ensemble for High Dimensional Data Classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 2, pp. 366-379.

[13] M. G. Rojas, A. C. Olivera, J. A. Carballido and P. J. Vidal, Nov'2020, "A Memetic Cellular Genetic Algorithm for Cancer Data Microarray Feature Selection," IEEE Latin America Transactions, vol. 18, no. 11, pp.1874-1883, doi: 10.1109/TLA.2020.9398628.

[14] Hassanat A, Almohammadi K, Alkafaween E, Abunawas E, Hammouri A, Prasath VBS, 2019, "Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach. Information," vol.10, no.12:390, pp. 1-36, https://doi.org/10.3390/info10120390

[15] Swati Swayamsiddha, 2020, "Bio-inspired algorithms: principles, implementation, and applications to wireless communication," Nature-Inspired Computation and Swarm Intelligence, pp. 49-63, Algorithms, Theory and Applications, ISBN 9780128197141

[16] D. B. Rawat, R. Doku and M. Garuba, Dec'2021, "Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security," in IEEE Transactions on Services Computing, vol. 14, no. 6, 1, pp.2055-2072, doi: 10.1109/TSC.2019.2907247.

[17] Jie Wang, Jing Xu, Chengan Zhao, Yan Peng Hongpeng Wang, 2019, "An ensemble feature selection method for high-dimensional data based on sort aggregation', *Systems Science Control Engineering, IEEE Access*, vol. 7, no.2, pp. 32-39.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 4
DOI: https://doi.org/10.17762/ijritcc.v11i4.6378
***Article Received**: 29 January 2023 **Revised**: 26 February 2023 **Accepted**: 01 March 2023*

[18] R. d. O. Nunes, C. A. Dantas, A. M. P. Canuto and J. C. Xavier-Junior, 2016, "An unsupervised-based dynamic feature selection for classification tasks," IEEE, International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, pp.4213-4220.

[19] Feng Xia, Wei Wang, TeshomeMegersa Bekele, Huan Liu, 2017, "Big Scholarly Data: A Survey," *IEEE transactions on BigData*, vol. 3, no. 1, pp. 18-35.

[20] Ali M, Ali SI, Kim D, Hur T, Bang J, Lee S, Kang BH, Hussain M, 2018, "uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features,' *PLoS* One, 2018 Aug 28,13(8):e0202705, PMID: 30153294; PMCID: PMC6112679.

[21] Lidong Wang, Randy Jones, 2017, "Big Data Analytics for Disparate Data," *American Journal of Intelligent Systems*, vol. 7, no. 2, pp. 39-46.

[22] Malhotra, Ruchika & Sharma, Anjali, 2021, "Threshold benchmarking for feature ranking techniques," Bulletin of Electrical Engineering and Informatics, vol.10. no.2, pp.1063-1070, doi:10.11591/eei.v10i2.2752.

[23] B.Seijo-PardoI. Porto-DiazV. Bol on-Canedo A. Alonso-Betanzos, 2017, "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Elsevier*, *Knowledge-Based Systems*, vol.118, pp. 124-139.

[24] Pimentel A, Carreiro AV, Ribeiro RT, Gamboa H., Jun'2018, "Screening diabetes mellitus 2 based on electronic health records using temporal features," *Health Informatics Journal*, vol.24, issue.2, doi: 10.1177/1460458216663023, PMID: 27566751, pp.194-205

[25] Alkundi A and Momoh R., 2020, "COVID-19 infection and diabetes mellitus, "*Journal of Diabetes, Metabolic Disorder Control*, DOI: 10.15406/jdmdc.2020.07.00212, vol. 7, no.4, pp.119-120.

[26] *UCI* Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/connectionist+bench (sonar,+mines+vs.+rocks)

[27] Zhang, Zili & Yang, Pengyi, 2019, "An Ensemble of Classifiers with Genetic Algorithm Based Feature Selection," *IEEE Intelligent Informatics Bulletin*, Vol-9, pp. 18-24.

[28] Jayshree Ghorpade-Aher, Balwant Sonkamble, Dec'2022, "A Machine Learning Algorithm for Multi-Source Heterogeneous Data with Block-Wise Missing Information", *IJCSE, Engg Journals Publications*, ISSN: 0976-5166, Vol. 13, No. 6, pp.1893-1904.

[29] Sharma, N., Dev, J., Mangla, M. et al., 2021, "A Heterogeneous Ensemble Forecasting Model for Disease Prediction," *Springer*, New Gener. Computing., pp.1-15.

[30] Y. Zhao and R. Duangsoithong, 2020, "Empirical Analysis using Feature Selection and Bootstrap Data for Small Sample Size Problems," IEEE 16th *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, Chonburi, Thailand, pp. 814-817.