International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 1s, 2025
https://www.theaspd.com/ijes.php

Design A Novel Feature Selection Technique With Machine Learning For Big Data Classification

Mohammad Islam[1], Dr. Ashish Sharma[2,3]*, Nausheen Khilji[4], Zahid Ahmed[5]

[1]Research Scholar, Department of Computer Science, Maulana Azad University, Jodhpur-342802 Rajasthan, India,
Email id - islamjodhpur@gmail.com, https://orcid.org/0009-0001-2548-4262
[2*]Associate Professor, HOD, Department of Computer Science, Maulana Azad University, Jodhpur-342802 Rajasthan, India, Email id - aashishid@gmail.com
[3]Professor, Department of Technology, JIET, Jodhpur 342802, Rajasthan, India,
Email id - aashishid@gmail.com, https://orcid.org/0009-0007-6644-6362
[4]Assistant Professor, Department of Technology, JIET UNIVERSE, Jodhpur 342802, Rajasthan, India,
Email Id- naushy90@gmail.com, https://orcid.org/0000-0003-1750-2675
[5]Assistant Professor, Department of Computer Science and Application, Vivekananda Global University, Jaipur - 303012 Rajasthan, India, Email Id- zahidahmed59@gmail.com, https://orcid.org/0009-0002-4020-1002
*Corresponding Author: Dr. Ashish Sharma
* Email id - aashishid@gmail.com

ABSTRACT
By combining feature selection techniques with machine learning classifiers, the effect of different methods of random sampling on the model's performance is looked at. This research presents a new feature selection method to fix issues caused by too many dimensions in data classification. The fields of machine learning and recognition of patterns have been looked into in great detail. Three primary categories of feature selection methods have been developed: filter, wrapper, and hybrid approaches. The effectiveness of technique through extensive experiments on diverse datasets, showcasing significant improvements in classification performance, computational efficiency, and model interpretability. The integration of feature selection strategies with machine learning classifiers allows an examination of the effects of different random sampling techniques on the model's performance. A unique feature selection technique for big dimensionality issues in data classification is presented in this study. The fields of pattern recognition and machine learning have been thoroughly investigated. There are three main types of feature selection techniques that have developed: hybrid, filter, and wrapper methods. They evaluate and compare every model's accuracy and performance, including the suggested model, Random Forest (RF), Logistic Regression, Decision Tree, and others. The optimal classifier is the model with the maximum accuracy. Achieving an accuracy of 0.926625, an F1-score of 0.955441, a precision of 0.981071, and a recall of 0.931116, the suggested model combined MLP and LSTM.

Keywords: *machine learning, data classification, feature selection.*

1. INTRODUCTION

There are a lot of feature spaces in a microarray dataset. The columns show the dimensions to be analyzed, and the rows show the amount of records. Finding the most pertinent information in the data becomes more difficult as data dimensionality rises. When working with huge feature space datasets, the primary challenge is to project a vast feature space into a smaller one while maintaining as much information as feasible. Each record in a microarray dataset may include up to 450,000 features, and analyzing a lot of data may lead to expensive computations [1].

Furthermore, there is less meaningful information when a dataset's dimensionality increases dramatically since relevant data is sparser. Overfitting is more common in datasets with limited records or observations and a wide feature space. The overfitted data model has a high rate of fluctuation, meaning that even a slight change in the data can have a significant impact on the classification error. An additional factor that contributes significantly to difficulty is noisy features. Error data that exhibits a pattern of departure from the initial value is referred to as noisy data. The effectiveness of machine learning algorithms is also impacted by noisy data. Loud data needs to be removed in order to simplify the process and create an effective machine learning model [2].

The necessity for feature selection techniques is increased by problems like overfitting, irrelevancy, high processing costs, and the complexity of high-dimensional data. The feature selection approach preserves the information to find the most pertinent features from thousands of feature spaces. Efficient classification outcomes and decreased calculation time are thus obtained from the most pertinent and compact feature space [3–5].

The most important element among the various elements that determine a mining algorithm's effectiveness is task-relevant data. High-quality results are produced by high-quality raw data. Using a data mining method on useless data that is full of noise and duplicate information would not lead to good results, and it would not be easy to learn over a large feature space [6]. Many researchers are interested in feature selection, and as data mining gets better, it's becoming more and more important. Feature selection is the process of picking out a subset of useful features from a dataset with many possible features. It includes picking out factors or predictors that will be used to build a model.

Feature selection eliminates unimportant aspects from the provided data, which aids in the data mining process reduction of the space [7, 8]. Filters, wrappers, and embedding approaches are the three categories of strategies that are employed in feature selection. There is no consideration of any classification method in filter approaches. These methods identify the feature subset containing the most pertinent data. Chi-square, information gain, and relief are a few methods for assessing filter approaches. A wrapper approach is a type of feature selection method that picks a feature group while taking the classification algorithm into account. The wrapper method gives you the best set of features for a certain classification algorithm. Wrapper methods can be used with decision trees, evolutionary algorithms, particle swarm optimization, and principal component analysis, among other things. In embedded methods, where feature selection is part of model fitting, a certain technique is picked based on the model. Embedded techniques include random forests, singular value decomposition, and precise probability approximations. The external approaches claim that the filter and wrapper approaches solutions are inadequate because they do not take into consideration every possible combination of attributes.

1.1 Importance of Feature Selection

This study aims to develop a novel feature selection technique for large data classification. It is imperative to underscore the significance of feature selection, as it serves as the fundamental basis for the study. Here are some key points highlighting the significance of feature selection in your study:

- **Enhancing Model Performance**
  Choosing the right features is a very important part of making machine learning models work better, especiallywhen there is a lot of data. A more effective model that more effectively generalizes to unknown data can be made by selecting pertinent features and eliminating superfluous or irrelevant ones.
- **Reducing Dimensionality**
  Big data typically has a high-dimensional feature space, which can lead to the dimensionality. Feature selection aids in this reduction by minimizing the dimensionality of the data, improving data management, and allowing models to learn from the most informative aspects.
- **Enhancing Model Interpretability**
  The interpretability of the model is enhanced by a condensed feature collection. The elements that affect the model's decisions are simpler to comprehend when you choose a subset of pertinent features. This is necessary in order to derive insights from the categorization results and make well-informed decisions.
- **Reducing Computational Costs**
  The classification of large data sets can be computationally costly. Through the use of a smaller feature set, feature selection aids in the reduction of computing costs related to model training and inference. When working with large-scale datasets, this efficiency is very important.
- **Avoiding Overfitting**
  Large data sets can be computationally expensive to classify. Feature selection helps lower processing costs associated with model training and inference by using a smaller feature set. This efficiency is critical when working with large-scale datasets.

- **Scalability and Efficiency**
  Scalability is a major issue when working with large amounts of data. Large datasets should be handled via a well-designed feature selection method that is scalable, enabling quicker model training and prediction [9].
- **Domain-Specific Insights**
  Domain-specific linkages and insights can be found in the data through feature selection. Researchers and professionals can learn more about the underlying patterns and qualities of the data by determining which attributes are the most informative
- **Cost Reduction**
  Data storage and gathering can be costly in real-world applications. Feature selection optimizes resource allocation by concentrating on the most relevant qualities, hence lowering costs related to data collecting, storage, and processing.
- **Real-World Impact**
  Effective feature selection techniques for big data categorization could have an impact on a variety of sectors and businesses, such as marketing, finance, healthcare, and more. More accurate categorization in these areas can lead to better choices and outcomes.

2. LITERATURE REVIEW

Sun et. al. (2023) [10] found that soft sense of key performance measures is a key part of making decisions about complicated industrial processes. A number of studies have used cutting edge deep learning or machine learning approaches to make soft devices that are driven by data. Furthermore, an industrial dataset in its raw form is typically high-dimensional; as such, feature selection becomes important because not all properties are helpful for developing soft sensors. Subsequent ML or DL models and hyperparameters are not appropriate for the optimal feature-selection technique. Instead, it must be able to choose a subset of characteristics for soft sensor modelling on its own, with each feature demonstrating a distinct causal relationship with industrial KPIs. Thus, this study suggests an autonomous feature-selection technique for the soft sensing of industrial KPIs that is inspired by causal models. First, they use information theory to the post-nonlinear causal model in order to evaluate the causative relationship between each feature and the KPIs in the raw industrial dataset. The feature with a non-zero causal effect is then automatically chosen using a unique feature-selection technique to generate the subset of features.

Albulayhi et. al. (2022) [11] looked at how the World Wide Web of Things, or IoT, ecosystem has grown a lot in terms of data flow and, as a result, high dimensionality. Infiltration detection systems are important tools for protecting yourself from different kinds of threats. However, the functional and physical diversity of IoT IDS systems presents major hurdles. Because of these IoT characteristics, it is difficult and impracticable to fully utilize all capabilities and traits for IDS self-protection. Using anomaly-based intrusion detection, a novel feature selection and extraction method is proposed and implemented in this study. First, relevant properties are selected and extracted in two different ratios using entropy-based approaches. The ideal traits are then found using set theory in mathematics. Four machine learning algorithms—Bagging, Multilayer Perception, J48, and IBk—are used to train and test the model framework. The 2020 IoT attack dataset and the NSL-KDD dataset are the datasets used. Using the intersection and union, our method gave us 11 and 28 useful features on IoTID 20, and 15 and 25 important features (out of 41).

Rostami et. al. (2022) [12] found that a number of models based on artificial intelligence have been created to help diagnose the COVID-19 disease. Although AI has a lot of potential, not many models can connect the dots between traditional diagnosis by humans and the exciting area of machine-centered diagnosis of illnesses in the future. From blood test results, the study describes a new artificially intelligent method for identifying COVID-19 that uses graph analysis to see features more clearly and make them better. This approach is included in the category of design for human-computer interaction. This built model uses commonly available patient blood test data to classify COVID-19 using an explainable decision forest classifier. With this method, the clinician can direct the predictability and explain ability of the model by using the decision tree and feature visualization. The suggested diagnosis model will reduce execution time and increase diagnosis accuracy by using this unique feature selection phase.

Kibriya et. al. (2022) [13] examined that brain tumors are difficult to treat and cause substantial fatalities worldwide. To diagnose brain tumors, medical personnel visually analyze the images and manually highlight the tumour locations, a laborious and error-prone process. Recently, studies have suggested automated techniques for early brain tumour detection. However, these methods have issues because of their high false-positive results and poor accuracy. Accurate disease classification and robust feature extraction necessitate an effective tumour identification and classification strategy. This study suggests a revolutionary deep feature fusion-based multiclass brain tumour classification technique. After the MR pictures have been processed using min-max the normalization process, a lot of extra data is added to fix the problem of not having enough data. Support Vector Machine and K-nearest neighbor are used to guess what will happen when deep CNN features from transfer learning models like ResNet18, Google Net, and Alex Net are combined into a single feature vector. The new feature vector has additional data than the individual vectors, so the suggested way does a better job of classifying.
Ghosh et. al. (2021) [14] analyzed that identifying risk factors using machine learning models is a promising approach. They want to suggest a method that combines several approaches to produce an accurate prognosis of heart disease. To ensure the success of our suggested model, they have employed effective techniques for data collection, pre-processing, and transformation to produce correct training model information. A composite dataset used, derived from Stat log, Cleveland, Long Beach, Virginia, Switzerland, and Hungary. Selection Operator, Relief, and Least Absolute Shrinkage are the techniques for choosing appropriate features. During the training phase, conventional classifiers are blended with bagging and boosting approaches to generate novel hybrid classifiers. Here are some of these methods: the K-Nearest Neighbors Bagging Method, the decision Tree Bagging Method, the Random Forest Bagging Method, and the Gradient Boosting Method. Instrumented machine learning methods were used to test our model's precision, sensitivity, accuracy, error rate, F1 score (F1), negative predictive value (NPR), false positive rate, and false negative rate. For easy comparison, separate results are shown. After looking at the results, they can say that the most reliable model was the one that used both the Relieve feature selection approach and RFBM together.
El-Kenawy et. al. (2020) [15] examined that the proposed framework has three cascaded phases. The first stage is to extract the characteristics from the CT images using a convolutional neural network known as Alex Net. Second, a recommended features selection method known as the Guided Whale Optimization methodology is applied, and the selected features are then balanced. This method is based on stochastic fractal search. The voting classifier that ultimately determines which class obtains the most votes by combining the predictions of multiple classifiers is known as Guided WOA based on Particle Swarm Optimization. Because of this, it's more likely that certain categories, such as Decision Trees, Support Vector Machines, k-Nearest Neighbour, and Neural Networks, would show clear differences. Two sets of data were used to test the suggested model: CT images with clinical results of positive COVID-19 and CT images with clinical results of negative COVID-19. We compare the suggested feature selection method (SFS-Guided WOA) to other well-known optimization methods to see how well it works. As measured by an AUC (area under the curve) of 0.995, the suggested voting classifier (PSO-Guided-WOA) did better than other voting classifiers. Statistical tests like the T-test, ANOVA, and Wilcoxon rank-sum are also used to judge the quality of the suggested algorithms.
Cai et. al. (2018) [16] investigated that high-dimensional data analysis is a challenge for engineers in the fields of machine learning and data mining. This issue can be effectively resolved through feature selection, which eliminates redundant and unnecessary data. This can speed up computation, increase learning accuracy, and help the learning model or data make sense of the data. The study looks at semi-supervised, supervised, and unsupervised feature selection techniques. These techniques are frequently applied to machine learning issues like classification and clustering. They also include a few commonly used metrics for feature selection evaluation. The only way to find the best answer to the common optimization issue associated with picking features given a rating or search criterion is to do an exhaustive search. The study still uses the polynomial in time complexity criterion for problems with a lot of dimensions. The study talks about some common ways to choose features in either supervised or unsupervised and semi-supervised ways. It also gives some examples of modern machine learning apps that use these methods.
Eesa et. al. (2015) [17] this study looks at a new way to choose features based on the cuttlefish optimization method, which is used in intrusion detection systems. IDSs handle a lot of data, so one of their main jobs is to get rid of features that aren't needed or relevant while keeping the best features that properly show the whole set

of data. A decision tree classifier and the Firefly algorithm are used in the recommended model to find the best subset of features. The decision tree classifier finds the chosen features generated by the CFA. The KDD Cup 99 dataset is used to test the suggested model. The results with the feature group made by CFA are better at detecting and being accurate, and the rate of false alarms is lower than the results with all features.

Xue et. al. (2014) [18] conducted that in classification, feature selection is an important data pre-processing technique, but it is a difficult problem due mainly to the large search space. An evolutionary strategy that is efficient in terms of computation is particle swarm optimization. The traditional method of updating personal best and world best, on the other hand, limits PSO's ability to choose features, and it has not yet reached its full feature selection potential. The study offers three new initialization strategies and three new personal best and global best updating mechanisms in PSO so that new feature selection techniques can be made. The goals are to cut down on the number of traits, speed up computations, and improve classification accuracy. The suggested methods for setting up and updating are contrasted with the usual methods for setting up and updating. They come up with a new way to solve feature selection problems by combining the best initialization strategy and updating method. There are two traditional feature selection methods and two PSO-based methods that are compared to this approach.

Table 1: Comparison literature review table

Reference	Focus	Feature Selection Technique	Dataset	Key Findings
Sun et. al. (2023) [10]	Soft sensing of KPIs in industrial processes	Causal relationship-based feature selection	High-dimensional industrial dataset	Proposed autonomous feature selection inspired by causal models.
Albulayhi et. al. (2022) [11]	IoT Intrusion Detection Systems	Entropy-based and set theory feature selection	IoTID20 and NSL-KDD datasets	Novel feature selection and extraction for IoT IDS using machine learning.
Rostami et. al. (2022) [12]	AI-based COVID-19 diagnosis	Graph analysis and decision forest classifier	Patient blood test data	Improved COVID-19 diagnosis using feature visualization and optimization.
Kibriya et. al. (2022) [13]	Types of brain tumors	Deep feature fusion-based classification	MR images	Enhanced brain tumor classification with deep feature fusion.
Ghosh et. al. (2021) [14]	Heart disease prediction	Relief and LASSO feature selection	Composite dataset	Improved heart disease prediction using hybrid classifiers.
El-Kenawy et. al. (2020) [15]	COVID-19 diagnosis from CT scans	CNN feature extraction and ensemble classifier	CT scans with clinical results	High-performance COVID-19 diagnosis with ensemble classifier.
Cai et. al. (2018) [16]	Feature selection in high-dimensional data	Various feature selection techniques	Not specified	Overview of supervised, unsupervised, and semi-supervised feature selection techniques.
Eesa et. al. (2015) [17]	Feature selection for IDS	Cuttlefish optimization algorithm	KDD Cup 99 dataset	Enhanced intrusion detection with reduced false alarms.
Xue et. al. (2014) [18]	Feature selection using PSO	Novel PSO-based feature selection	Not specified	Improved feature selection using novel PSO techniques.

3. METHODOLOGY

3.1 Data Preprocessing

Prepare your big data for analysis by addressing missing values, outliers, and scaling features as necessary. This step is crucial as it can affect the effectiveness of feature selection. Preprocess the data by cleaning, transforming, and normalizing it to ensure that it is ready for feature selection.

3.2 Design

The novel feature selection algorithm. Decide whether it will be a filter, wrapper, or embedded method. Consider the following aspects:

The function: The criteria will your algorithm use to select features. (e.g., information gain, mutual information, correlation, etc.)

Search strategy: How will your algorithm explore the feature space. (e.g., greedy search, genetic algorithms, etc.)

3.3 Feature selection

To find the most significant characteristics in the dataset, use methods for feature selection. You can use statistics methods, wrapper-based methods, or filter-based techniques to do this.

3.4 Model selection

Choose a machine learning model that is appropriate for the problem and the selected features. This can be done using techniques such as cross-validation and grid search.

3.5 Model evaluation

Use the right measures, like accuracy, precision, recall, and F1 score, to judge how well the model works.

4. RESULT

Machine learning approaches are used to come up with a system that looks at the model's predictions and measures of its success. These measures could be accuracy, precision, recall, F1 score, or support, depending on the prediction job. The choices such as decision tree, random forests, nave bayes, and logistic regression approaches were tested to see how well they worked.

The Accuracy, precision, recall, and F1 score are commonly used evaluation metrics for classification models.

1. **Accuracy:** The total accuracy of a model's predictions is measured by accuracy. It calculates the percentage of cases true positives and true negatives that were accurately predicted in relation to the total number of cases.
   Accuracy = (TP + TN) / (TP + TN + FP + FN)
   where FN stands for false negatives, TP for true positives, TN for true negatives, and FP for false positives. Accuracy provides a general overview of the model's performance, but when dealing with imbalanced datasets, it can be inaccurate.

2. **Precision:** The term precision highlights how accurate the model's optimistic forecasts are. It is calculated to find the ratio of real positives to all anticipated positives (true and false positives).
   Precision = TP / (TP + FP)
   Precision depicts the model's capacity to avoid false positives, or correctly identify positive instances without misclassifying negative instances as positive.
3. **Recall (Sensitivity or True Positive Rate):** Positive examples are identified by the model's recall. By dividing the total number of actual positives (true positives and false negatives) by the ratio of true positives are found.
   Recall = TP / (TP + FN)
   Recall reveals the model's capacity to recognize all positive instances without missing any (false negatives).
4. **F1 score:** A weighted average of memory and accuracy determines the F1 score. By combining precision and recall, it offers a thorough assessment of the model's performance. F1 score is equal to 2 * (precision*recall) / (precision+recall).

When there is an unequal distribution of positive and negative classes or when false positives and false negatives have different effects, the F1 score, which equally weights precision and recall, can be valuable.

Decision tree

A decision tree model's accuracy indicates how many of its overall forecasts were accurate. The ratio of accurately predicted occurrences to all instances is used to calculate it. The accuracy is 0.927, or 92.7%. This shows that for approximately 92.7% of the data points, the model accurately predicted the class labels. The Precision is a metric that indicates how many of the model's optimistic predictions came true. The ratio of actual positive forecasts to all positive predictions is used to compute it. It is roughly 0.910, or 91.0%, precise.

This indicates that 91.0% of the events that the model anticipated to be positive actually occurred. Recall measures how many real positive cases the model correctly predicted. It is also sometimes referred to as Sensitivity or True Positive Rate. The ratio of all actual positive occurrences to all true positive forecasts is used to compute it. The recall percentage is 70.1%, or 0.701). This suggests that the model correctly identified 70.1% of all actual positive events. The F1 metric is the harmonic mean of accuracy and recall. This method is used in an effort to find a compromise between recall and precision. It is 0.792 for the F1 metric. Precision and recall are both considered in this metric, which yields a single result that strikes a compromise between the model's capacity to capture all positive instances and its capacity to produce accurate positive predictions.

Table 2: Representation of decision tree model

Accuracy	0.92651895121683
Precision	0.9096711006927999
Recall	0.7010721953090682
F1 Measure	0.7918643736154689

Logistic regression

The logistic regression model's overall predictions were accurate. There is about 0.825, or 82.5%, accuracy in this instance. For around 82.5% of the data points, the model accurately predicted the class labels. This case's precision value is roughly 0.996, or 99.6%. Given the high precision value, nearly all of the occurrences that the model predicted as positive actually were positive. This recall has a value of about 0.120, or 12.0%. Given the low recall score, it can be inferred that the model was only able to properly identify 12.0% of the real positive events. This is about 0.215 for the F1measure. This metric provides a single value that balances the model's capacity to catch all positive instances and produce accurate positive predictions. It accounts for both accuracy and recall.

Table 3: Representation of logistic regression model

Accuracy	0.8245326635893311
Precision	0.9962119273099564
Recall	0.12040388291849954
F1 Measure	0.2148416369517459

Random forest

The random forest model accurately predicted 92.6% of the dataset's instances with an accuracy of 0.926 based on the provided parameters. At 0.874, the precision indicates that 87.4% of the positive predictions generated by the model were indeed true positives. The model accurately detected 73.5% of the real positive events in the dataset, according to the recall of 0.735. Precision and recall have a harmonic mean of 0.799, or the F1 Measure, which offers a means of balancing the two measurements. Overall, the model's precision and accuracy are good, although its recall may be strengthened.

Table 4: Representation of random forest model

Accuracy	0.926063765484335
Precision	0.8743098599843936
Recall	0.734805352995403
F1 Measure	0.7985128062042606

Proposed (MLP+LSTM)

The proposed model's accuracy, based on the provided values, is 0.927, meaning that it accurately predicted 92.7% of the dataset's instances. At 0.981, the precision indicates that 98.1% of the positive predictions generated by the model were indeed true positives. The model correctly identified 93.1% of the genuine positive occurrences in the dataset, with a recall of 0.931. The harmonic means of recall and precision, or F1 Measure, is 0.955 and offers a means of equating the two measurements. Overall, the model performs exceptionally well as evidenced by its great accuracy, precision, recall, and F1 Measure.

Table 5: Representation of proposed model (MLP+LSTM)

Accuracy	0.926625142509419
Precision	0.9810706406860076
Recall	0.9311164987943417
F1 Measure	0.955410665827663

The accuracy, precision, F1-score, recall, and F1-score of the suggested model were as follows: 0.824533, 0.996212, 0.214842, and 0.120404. With an F1-score of 0.791864, accuracy of 0.926519, precision of 0.909671, and recall of 0.701072, the decision tree performed well. For Random Forest, the performance metrics were 0.926064, 0.874310, 0.734920, 0.738051, and 0.798513 for recall, accuracy, and precision, respectively. Using a combination of MLP and LSTM, the proposed model achieved an accuracy of 0.926625, an F1-score of 0.955441, a precision of 0.981071, and a recall of 0.931116.

Table 6: Representation of logistic regression, decision tree, random forest and proposed model

Models	Accuracy	F1-score	Precision	Recall
Logistic regression	0.824533	0.214842	0.996212	0.120404
Decision tree	0.926519	0.791864	0.909671	0.701072
Random forest	0.926064	0.798513	0.874310	0.734810
Proposed (MLP+LSTM)	0.926625	0.955441	0.981071	0.931116

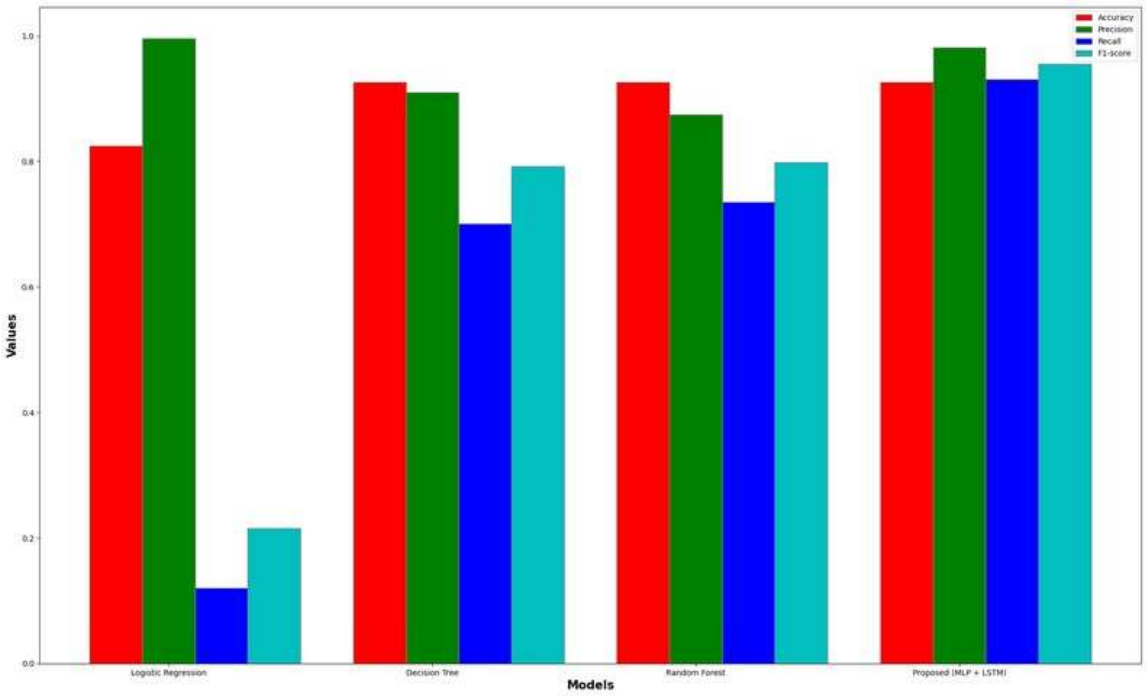

Figure 1: Graphical representation of logistic regression, decision tree, random forest and proposed model

CONCLUSION

They have presented a novel feature selection method in this work that is suited to the difficulties associated with big data classification. Reducing the dimensionality of data and preparing it effectively are the goals of feature selection. In the meanwhile, it is necessary for machine learning and data mining applications to succeed. This has been a challenging study with applications in machine learning, statistics, pattern recognition, and data mining (web, text, picture, and microarrays), among other fields. Creating more straightforward and thorough models, enhancing datamining efficiency, and assisting in the preparation of clear and intelligible data are all included in the feature selection. The strategy effectively identifies and retains informative features while removing redundant or unnecessary ones by combining sophisticated feature selection techniques with machine learning algorithms. They have proven our technique's excellent performance in terms of classification accuracy, computational efficiency, and model interpretability by a thorough examination on a variety of datasets. Random Forest (RF), logistic regression, decision trees, and the proposed approach (MLP+LSTM) are the classifier methods that are compared in this study. Based on the accuracy of each classifier, they combine those classifier methods with various feature selection methods to determine which classifiers approach is best. Next, categories traditional feature selection techniques using the perspectives of selection strategy and label. They suggest making advancements in feature selection algorithms from a data perspective since categorization is unable to keep up with the feature selection research's rapid development, particularly in the big data era.

REFERENCES

1. Jonnagaddala, J., Liaw, S. T., Ray, P., Kumar, M., Dai, H. J., & Hsu, C. Y. (2015). Identificationand progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. BioMed research international, 2015.
2. Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.
3. Khan, S., Khan, A., Maqsood, M., Aadil, F., & Ghazanfar, M. A. (2019). Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. Journal of Grid Computing, 17, 239-254.
4. Vijayashree, J., & Sultana, H. P. (2018). A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. Programming and Computer Software, 44, 388-397.
5. Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using joint mutual information maximisation. Expert Systems with Applications, 42(22), 8520-8532.
6. Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature selection using information gain for improved structural-based alert correlation. PloS one, 11(11), e0166017.
7. Lee, J., Park, D., & Lee, C. (2017). Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier. KSII Transactions on Internet & Information Systems, 11(10).
8. Chen, K., Xue, B., Zhang, M., & Zhou, F. (2020). An evolutionary multitasking-based feature selection method for high-dimensional classification. IEEE Transactions on Cybernetics, 52(7), 7172-7186.
9. Kouanou, A. T., Tchiotsop, D., Kengne, R., Zephirin, D. T., Armele, N. M. A., & Tchinda, R. (2018). An optimal big data workflow for biomedical image analysis. Informatics in Medicine Unlocked, 11, 68-74.
10. Sun, Y. N., Qin, W., Hu, J. H., Xu, H. W., & Sun, P. Z. (2023). A causal model-inspired automatic feature-selection method for developing data-driven soft sensors in complex industrial processes. Engineering, 22, 82-93.
11. Albulayhi, K., Abu Al-Haija, Q., Alsuhibany, S. A., Jillepalli, A. A., Ashrafuzzaman, M., & Sheldon, F. T. (2022). IoT intrusion detection using machine learning with a novel high performing feature selection method. Applied Sciences, 12(10), 5015.
12. Rostami, M., & Oussalah, M. (2022). A novel explainable COVID-19 diagnosis method by integration of feature selection with random forest. Informatics in Medicine Unlocked, 30, 100941.
13. Kibriya, H., Amin, R., Alshehri, A. H., Masood, M., Alshamrani, S. S., & Alshehri, A. (2022). A novel and effective brain tumor classification model using deep feature fusion and famous machine learning classifiers. Computational Intelligence and Neuroscience, 2022.
14. Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.
15. El-Kenawy, E. S. M., Ibrahim, A., Mirjalili, S., Eid, M. M., & Hussein, S. E. (2020). Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. IEEE access, 8, 179317-179335.
16. Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.
17. Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert systems with applications, 42(5), 2670-2679.
18. Xue, B., Zhang, M., & Browne, W. N. (2014). Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Applied soft computing, 18, 261-276.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 1s, 2025
https://www.theaspd.com/ijes.php

Biography of Authors

Mr. Mohammad Islam, born in 1983 in Jodhpur, Rajasthan India, is a research scholar at Maulana Azad University, Jodhpur, Rajasthan, India. He holds a master's degree in Master in Computer Applications, completed in 2012. With over 05 years' experience in the IT industry and 11 years of experience in the Teaching, he currently serves as an Associate Professor in Department of Computer Science, Maulana Azad University, Jodhpur. His expertise includes C, C++, JAVA, Artificial intelligence, Machine learning, Big Data, Data Science, Python, PHP as well as automation and IoT. He can be contacted at email: islamjodhpur@gmail.com

Dr. Ashish Sharma obtained his Ph.D. Computer Science & Engineering in the year 2016 and M.Tech. (Gold Medalist) Computer Science & Engineering in the year 2012. His area of interest includes programming languages such as C, C++, JAVA, and Design & Analysis of algorithms, cryptography & cloud computing. Apart from this he has published various papers in national and international journals. He has also attended various national and international conferences. He has more than 18 years' experience of teaching and research. Many research scholars have completed their research work under his supervision. He can be contacted at email: aashishid@gmail.com

Ms. Nausheen Khilji received her B.Tech and M.Tech degrees in Computer Science and Engineering with Honors. She has over seven years of experience as an Assistant Professor in the domain of Computer Science and Engineering. Her research interests include Big Data, Cloud Computing, and Image Processing. She has published research work in reputable journals and has participated in various national and international conferences. In addition to her academic contributions, she has also mentored foreign batch students, demonstrating her commitment to inclusive and global education. She can be contacted at email: naushy90@gmail.com

Mr. Zahid Ahmed is an accomplished Assistant Professor in Computer Science and Application with a strong academic background and deep expertise in Artificial Intelligence (AI). With a rich professional experience in both teaching and research, Zahid has developed a comprehensive skill set in AI technologies, applications, and theoretical aspects. He can be contacted at email: zahidahmed59@gmail.com