Proceedings of the 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI-2026)
DVD Part Number: CFP26US4-DVD; ISBN: 979-8-3315-5518-4

# Exploring the Impact of Filtering Techniques on the Performance of Different Classifiers for Predicting Diabetes Mellitus

V.Poornima
*Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies* Chennai, Tamil Nadu , India
poornimasudhaagar@gmail.com

K.Hazeena
*Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science & Technology, Chennai. haseenakader67612@gmail.com* Chennai, Tamil Nadu , India
haseenakader67612@gmail.com

A.Angel Cerli
*Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies* Chennai, Tamil Nadu , India
dr.angelcerli@gmail.com

Kamatchy.B
*Department of Computer Science and Information Technology, Vels Institute of science technology and advanced studies,* Chennai, Tamil Nadu , India
kamatchi6282@gmail.com

Sheela.K
*Department of Computer Applications,Vels Institute of science technology and advanced studies* Chennai, Tamil Nadu , Chennai
drksheela.research@gmail.com

**Abstract-- Diabetes is a chronic disorder that affects how many people around the world process their metabolism. The prevalence of diabetes is increasing alarmingly every year. Diabetes can harm various essential organs in the body if it is not managed well. Therefore, it is very important to identify diabetes early and initiate treatment as soon as possible to avoid the disease from causing such complications. This research applied five different techniques in WEKA tools to forecast diabetes based on the input attributes of the dataset. This research used 17 attributes, such as Age, Sex, Polyuria, Polydipsia and other medical terms to assess the likelihood of a patient developing disease. Nowadays Filtering techniques are crucial in various fields and applications, serving to extract relevant information, enhance data quality, and improve overall system performance. The filters—Resample, Obfuscate, and Discretizeare important for preprocessing data for machine learning algorithms. They help to overcome challenges related to data imbalance, privacy protection, and data representation, resulting in enhanced model performance and data usability. These filters are vital for achieving high-quality data preprocessing, feature engineering, and privacy preservation in machine learning workflows. Preprocessed data applied to the classification techniques, namely JRip, SMO, KNN, Logistic and SGD used to examine the diabetes. The performance measurement for this study are the mean absolute error and kappa statistics and the accuracy of correct classification, of the classifier. The results show that Overall, employing these filters in the preprocessing stage assists to optimize data preparation and establishes the basis for successful machine learning outcomes**

***Keyword- HealthCare, Resample, Obfuscate, Discretize, Machine learning Algorithms,Diseases Prediction.***

## I. INTRODUCTION

Diabetes mellitus is a collection of metabolic illnesses characterised by hyperglycemia caused by deficiencies in insulin production, action, or both.The American Diabetes Association's diagnostic criteria are as follows: (1) An HbA1c level of 6.5% or higher indicates glycated hemoglobin. (2) If the fasting blood glucose level exceeds 126 mg/dL, it signifies elevated basal levels.(3) A blood glucose level equal to or surpassing 200 mg/dL, two hours after a 75 g oral glucose tolerance test, suggests impaired glucose tolerance [1]. Type 2 diabetes, in particular, is related with insulin resistance (insulin action deficiency), which occurs when cells respond poorly to insulin, affecting their glucose intake [2].

Diabetes presents a worldwide public health concern. As indicated by the International Diabetes Federation, they estimated the global diabetic population for the year 2019.Diabetes is a worldwide public health problem. In 2019, the International Diabetes Federation reported that 463 million people worldwide had diabetes, with a 51% increase anticipated by 2045. Furthermore, it is predicted that there is one undiagnosed individual for every diagnosed person with diabetes. Medicine is the science and practise of determining illness diagnosis, prognosis, treatment, and prevention. Medicine comprises a wide range of health-care practises that have evolved to preserve and restore health through sickness prevention and treatment [3]. This is one of the most critical areas where data mining approaches may yield considerable benefits. Doctors will be able to forecast diseases more

efficiently with data mining tools, and they will be better equipped to handle possible high-risk patients [4].The majority of diabetes cases can be broadly classified into two primary etiopathogenic categories, as follows:

- Type 1
- Type 2

Type 1 Diabetes: Type 1 diabetes is an autoimmune condition where the body's immune system mistakenly attacks and destroys insulin-producing beta cells in the pancreas. As a result, individuals with Type 1 diabetes cannot produce insulin, a hormone necessary to regulate blood sugar levels. This type usually develops in childhood or adolescence and requires lifelong insulin therapy.

Type 2 Diabetes: Type 2 diabetes is a metabolic disorder where the body becomes resistant to insulin, and the pancreas cannot produce enough insulin to maintain normal blood sugar levels. It is often associated with factors like obesity, sedentary lifestyle, and genetics. Type 2 diabetes typically occurs in adulthood but can also affect younger individuals. Treatment may involve lifestyle changes, medication, and, in some cases, insulin therapy. Controlling diabetes is crucial to prevent complications, enhance quality of life, reduce the risk of associated health problems, and promote overall well-being.

Data mining is a suitable field for applying medicine, as it can deal with large and complex datasets of diseases and their relationships. Data mining can analyze many datasets with many variables that are beyond the capacity of a single analyst or doctor, or even a team of analysts. Data mining starts with defining the problem to be solved. Defining the issue at hand plays a pivotal role in shaping the data mining process and selecting an appropriate modelling technique. Machine learning, a subset of data science, concentrates on algorithms designed to glean insights from data and produce precise predictions. Data mining empowers healthcare institutions to gain insights into patterns related to diseases and more. Through the utilization of data mining methodologies and machine learning algorithms, we bolster the analysis of diabetes, contributing to its mitigation and prevention.

In numerous real-world scenarios, classification stands out as a pivotal decision-making technique. This paper centers around the core objective of categorizing data into diabetic or non-diabetic categories and enhancing classification precision. Contrary to common belief, increasing the volume of samples doesn't invariably translate to heightened classification accuracy in various classification problems. Instances abound where algorithmic performance showcases speed but falters in accurately classifying data. Our model is primarily designed to introduce feature selection filters with the intent of attaining elevated accuracy levels. These filters [5] play a crucial role in assessing the significance of individual features, primarily grounded in their statistical attributes, thereby assigning a numeric score to each feature based on its impact on algorithmic performance. By emphasizing a larger portion of the dataset for training purposes and reducing the dataset size for testing, the potential to augment classification accuracy becomes evident. The overall framework of the proposed system is shown below fig.1:

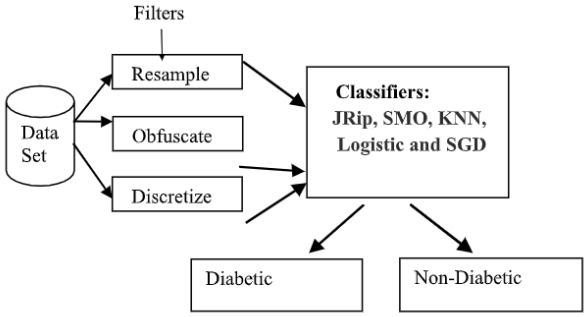


Fig 1: Proposed model

The main task of this research work is partitioned into two stages:

First, for the feature selection three different filtering techniques were used namely Resample , Obfuscate and Discretize to improve the classification accuracy . Next, the refined data undergoes application across five distinct data mining algorithms to facilitate classification, aiming for an optimal blend of accuracy and computational efficiency. Within this study, a range of classification methodologies has been employed to discern between diabetic and non-diabetic datasets. This dataset essentially embodies a binary classification challenge, characterized by its division into diabetic and non-diabetic classes.

## II. LITERATURE SURVEY

Diabetes mellitus is a long-lasting illness that can cause harm to various parts of the body. Detecting infections at an early stage is crucial to prevent them from spreading and potentially saving lives. In this study, researchers aimed to predict diabetes early on using different data mining methods. The researchers utilized a dataset comprising 768 instances sourced from the PIMA Indian Dataset. Their findings demonstrated that the Modified J48 Classifier exhibited superior performance in accuracy compared to alternative approaches [6].Information technology-based healthcare can greatly benefit from Deep Convolutional Neural Networks, which enable personalized medical treatment and reduce the need for invasive procedures [7]. Data scientists often prioritize identifying the most relevant features in a dataset to train learning algorithms effectively [8]. Some features might not contribute to the algorithm's improvement and could even worsen its performance. To address this, Feature Selection (FS) algorithms are utilized to identify the subset of features that enhance model performance while preventing overfitting and expediting training. Additionally, understanding which

features are selected can provide valuable insights into the datasets. Dealing with wide data, where the number of features is extensive, presents a significant challenge. Nevertheless, this approach proves especially advantageous when dealing with substantial datasets, a common scenario in big data applications where the reduction of data size and execution durations hold paramount importance [9]. Existing literature outlines diverse categorizations of Feature Selection (FS) algorithms, with one of the most prevalent and practical classifications organizing features based on their relevance to the learning algorithm [10]. Practical instances in various fields, notably biological and genomics domains, frequently grapple with these precise challenges. For illustration, electroencephalography data have been harnessed for purposes such as epilepsy diagnosis, early identification of type 2 diabetes, and prognosticating mortality among intensive care unit patients [11]. Equally, broad datasets manifest in other domains, including engineering for fault detection, computer security realms pertaining to intrusion detection, and solar radiation estimation [12]. The efficiency of learning algorithms is adversely affected by the existence of an extensive array of features, leading to high dimensionality, thereby amplifying execution times [13]. In the healthcare sector, data mining (DM) approaches are essential for empowering health systems. This study focuses on early prediction of heart disease using various DM approaches. The effectiveness of the study is evaluated using six classifier assessment metrics, and the American Heart Association (AHA) dataset is employed for experimentation [14]. The principal objective of Feature Selection (FS) algorithms revolves around identifying the optimal amalgamation of features that engenders models which are streamlined, expedient, and easier to comprehend. Nonetheless, this pursuit is intricate and recognized as a problem falling within the realm of NP-hard challenges [15]. Rani et al.,[16] present a thorough summary of current studies on the diagnosis of heart disease by examining papers from reliable sources that were released between 2014 and 2022. It lists the difficulties that researchers encounter and suggests possible fixes. The study also makes recommendations for future lines of inquiry into this important area. SVM is commonly used to classify data by identifying the optimal hyperplane between two groups. Working with large datasets, however, might result in a variety of issues, such as laborious and ineffective solutions. Mahdi et al., [17] uses a stochastic gradient descent technique to update the SVM. Two simulation datasets were used to test the novel method, known as the extended stochastic gradient descent SVM (ESGD-SVM).

## III. MATERIALS AND METHODS

Diabetes is considered a chronic medical condition that can lead to various complications. Among the most common microvascular sequelae of diabetes are retinopathy, nephropathy, and neuropathy. It is crucial to employ data mining techniques to identify risk factors associated with each of these conditions in order to prevent their occurrence. To enhance the performance of diabetes detection, a dataset is subjected to three preprocessing filters: Resample, Obfuscate, and Discretize. The preprocessed data is then fed into five different machine learning algorithms for classification. Subsequently, the data will be further collected from the diabetes repository for additional analysis.

### A. Data Mining

For many years, researchers have utilized data mining techniques in the healthcare industry to study the patterns of heart disease. However, with the abundance of data and advanced data analytics tools, a situation known as "data rich but poor information" has emerged. Consequently, even when working with extensive datasets, the scrutinized information might not substantially enhance disease diagnosis or treatment. Within the healthcare domain, numerous scholars have endeavored to enhance data mining methodologies, encompassing pattern analysis, knowledge extraction from databases, knowledge mining, data dredging, and data archaeology.

### B. Classifier Algorithm

A multitude of algorithms are available for classifying disease datasets, encompassing JRip, SMO, K-Nearest Neighbour, Logistics, and Stochastic Gradient Descent. Within this study, an assessment of five Weka classifiers has been conducted using the UCI Diabetes dataset. Presented below are concise overviews of each utilized classifier.

(i) SGD: Stochastic Gradient Descent algorithm

Stochastic gradient descent is a widely used optimization algorithm in the field of machine learning. It is employed to find the most suitable model parameters that lead to the best possible alignment between predicted and actual outputs. It operates through an iterative process to seek the optimum value (minimum/maximum) of an objective function. Gradient descent is widely employed to adjust model parameters and minimize the cost function in machine learning projects. The primary aim of gradient descent involves determining the model parameters that yield optimal accuracy across both the training and test datasets . In this process, the gradient represents a vector indicating the direction of the steepest ascent of the function at a specific point. The algorithm progressively moves in the opposite direction of the gradient, gradually descending towards lower function values until it reaches the function's minimum.

(ii) JRip: JRip is a classification algorithm that falls under the category of Rule-based Induction. It utilizes a propositional rule learner named Repeated Incremental Pruning to Produce Error Reduction (RIPPER). Using WEKA with 10-fold cross-validation and a confidence factor of 0.25, JRip generates a prediction model presented in Figure 3, consisting of two rules. If both IT Elective Grade and IT Professional Grade are rated as Fairly Good, the graduate is predicted to be Not employable; otherwise, the prediction is Employable.

(iii) SMO: SMO (Sequential Minimal Optimization) is an

improved algorithm that customizes training for Support Vector Machines (SVMs). To train a support vector machine, the algorithm addresses the challenge of solving a large quadratic programming optimization (QP) problem. SMO breaks down the large QP problem into smaller QP problems and solves them analytically, making it efficient and capable of handling large training datasets due to its memory adaptability.

(iv) KNN: On the other hand, KNN (K-nearest neighbor) is one of the simplest and earliest classification algorithms. It can be considered a simpler version of the Naive Bayes (NB) classifier, as it doesn't require considering probability values. The 'K' in KNN represents the number of nearest neighbors considered to take a 'vote' for the classification of a sample object. Different values of 'K' can lead to varied classification results for the same sample object.

(v) Logistics: Logistic regression is a statistical technique employed to forecast the probability of a binary outcome by considering one or more independent variables. The algorithm fits a logistic curve to the data, which is an S-shaped curve ranging between 0 and 1, representing the probability of the binary outcome.To make predictions, the independent variables are combined linearly with weights, and then the logistic function is applied to map this linear combination to the probability range. Throughout the training phase, the algorithm identifies the optimal weights that minimize the disparity between predicted probabilities and the factual binary labels within the training dataset. This optimization procedure is commonly executed through methodologies such as gradient descent.

C.Filters:

(i) Resample: Resampling techniques involve modifying the distribution or representation of data points in a dataset. They are commonly used when dealing with imbalanced datasets, where the number of samples in different classes is significantly unequal. The benefits of resampling techniques include improved model performance on imbalanced datasets, mitigating bias towards the majority class, and enabling the model to learn patterns from underrepresented classes.The two main resampling techniques are oversampling and undersampling.

(ii) Obfuscate : Obfuscation techniques are used to transform or mask sensitive or identifiable information in datasets, ensuring privacy and data protection. The purpose is to remove or obscure personally identifiable information (PII) or other sensitive attributes while preserving the useful patterns and relationships in the data. Obfuscation can include

techniques like data anonymization, tokenization, noise addition, or encryption.

(iii) Discretize: Discretization is the process of converting continuous numerical variables into categorical variables or discrete intervals. This transformation can be useful in several scenarios:

- Simplifying complex data: Discretization reduces the complexity of continuous data by grouping values into intervals or bins, making it easier to interpret and analyze.

- Handling non-linear relationships: Some machine learning algorithms may work better with categorical or discrete variables. Discretizing continuous features can help capture non-linear relationships and improve model performance.

- Feature engineering: Discretization can be considered as a form of feature engineering, creating new features that capture specific patterns or characteristics within the data.

## IV. RESULTS AND DISCUSSIONS

In this section, we have presented the results of the performance analysis of various classifiers using filtering techniques and discuss comparative performance of the each classifiers. To evaluate the performance of the proposed model we used the dataset collected from UCI Machine learning repository[18]

TABLE 1:PERFORMANCE ANALYSIS OF FIVE CLASSIFIERS WITHOUT FILTERS

| Evaluator/ classifier | TP | Precision | F-Measure | ROC | PRC | Kappa | MAE |
|---|---|---|---|---|---|---|---|
| JRip | 95 | 95 | 95 | 95.4 | 93.6 | 89.59 | 0.0675 |
| SMO | 92.1 | 92.1 | 92.1 | 91.8 | 88.9 | 83.39 | 0.0788 |
| KNN | 98.1 | 98.1 | 98.1 | 98.4 | 97.8 | 95.96 | 0.0207 |
| Logistic | 92.3 | 92.3 | 92.3 | 96.9 | 96.6 | 83.78 | 0.1114 |
| SGD | 91.9 | 91.9 | 91.9 | 91.5 | 88.6 | 82.94 | 0.0808 |

Table 1 illustrates the comparative study of the classification outcome offered by the five different classifiers on the applied diabetes dataset without applying feature selection methods. The experimental values denoted that the KNN model has observed inferior results with the highest TP, Precision, F-Measure, ROC and PRC curve 98.1%, 98.1%,98.1%,98.4% and 97.8% respectively.

TABLE 2: PERFORMANCE ANALYSIS OF FIVE CLASSIFIERS WITH RESAMPLE

| Evaluator/ classifier | TP | Precision | F-Measure | ROC | PRC | Kappa | MAE |
|---|---|---|---|---|---|---|---|
| JRip | 98.3 | 98.3 | 98.3 | 98.7 | 98.4 | 96.29 | 0.0209 |
| SMO | 96.5 | 96.5 | 96.5 | 96.4 | 95 | 92.58 | 0.0346 |
| KNN | 98.8 | 98.8 | 98.8 | 99.9 | 99.9 | 97.54 | 0.0101 |
| Logistic | 98.8 | 98.8 | 98.8 | 99.3 | 99 | 97.52 | 0.0115 |
| SGD | 98.8 | 98.8 | 98.8 | 98.3 | 98.3 | 97.52 | 0.0115 |

Table 2 illustrates the comparative study of the classification outcome offered by the five different classifiers on the applied diabetes dataset with applying Resample feature selection methods. The experimental values denoted that all five classifiers has raised their accuracy when applied resample feature selection on diabetic dataset. The below table proved that performance metrics of five classifiers got better results than without applying resample.

TABLE 3:PERFORMANCE ANALYSIS OF FIVE CLASSIFIERS WITH OBFUSCATE

| Evaluator/ classifier | TP | Precision | F-Measure | ROC | PRC | Kappa | MAE |
|---|---|---|---|---|---|---|---|
| JRip | 97.3 | 97.3 | 97.3 | 99 | 99 | 94.21 | 0.0349 |
| SMO | 94.2 | 94.2 | 94.2 | 93.7 | 91.6 | 87.59 | 0.0577 |
| KNN | 99.2 | 99.2 | 99.2 | 99.7 | 99.7 | 98.36 | 0.0093 |
| Logistic | 95.8 | 95.8 | 95.8 | 98.8 | 98.8 | 90.9 | 0.0499 |
| SGD | 95.8 | 95.8 | 95.8 | 95.7 | 94 | 90.96 | 0.0423 |

Table 3 illustrates the comparative study of the classification outcome offered by the five different classifiers on the applied diabetes dataset with applying Obfuscate feature selection methods. The experimental values denoted that all five classifiers has raised their accuracy when applied Obfuscate feature selection on diabetic dataset. The below table proved that performance metrics of five classifiers got better results than without applying Obfuscate.

Table 4 illustrates the comparative study of the classification outcome offered by the five different classifiers on the applied diabetes dataset with applying Discretize feature selection methods. The experimental values denoted that all five classifiers has raised their accuracy when applied Discretize feature selection on diabetic dataset. The below table proved that performance metrics of five classifiers got better results than without applying Discretize.

TABLE 4: PERFORMANCE ANALYSIS OF FIVE CLASSIFIERS WITH DISCRETIZE

| Evaluator / classifier | TP | Precision | F-Measure | ROC | PRC | Kappa | MAE |
|---|---|---|---|---|---|---|---|
| JRip | 98.3 | 98.3 | 98.3 | 98.7 | 98.4 | 96.29 | 0.0209 |
| SMO | 98.9 | 98.9 | 98.9 | 99.1 | 98.5 | 97.52 | 0.0113 |
| KNN | 98.8 | 98.9 | 98.8 | 99.9 | 99.9 | 97.54 | 0.0101 |
| Logistic | 98.9 | 98.9 | 98.9 | 99.7 | 99.7 | 97.52 | 0.0111 |
| SGD | 98.9 | 98.9 | 98.9 | 99.1 | 98.5 | 97.52 | 0.0113 |

To analyze the performance of the five classifiers (JRip, SMO, KNN, Logistic, and SGD) under different filtering techniques (No Filters, Resample, Obfuscate, and Discretize),the Kappa statistic can be employed as a gauge to assess the concordance between the anticipated and genuine classes. A heightened Kappa value signifies improved classification efficacy. Overall, KNN, Logistic regression, and JRip show consistent and relatively high performance across different filtering techniques. SMO and SGD exhibit lower performance initially but improve with specific filtering techniques. The choice of the best classifier and filtering technique would depend on the specific dataset and task requirements.

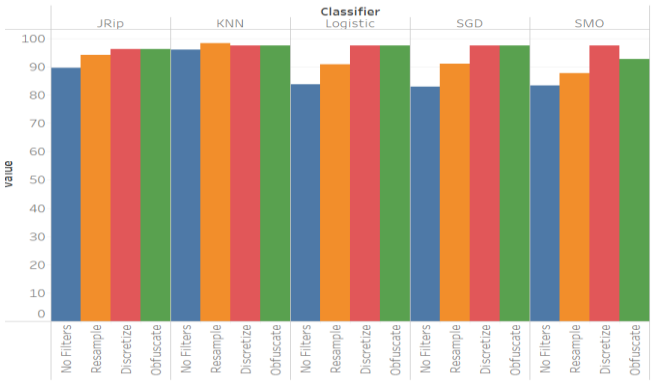


Fig 2: Kappa Analysis of Five classifiers

Figure 2 illustrates the performance of five classifiers (JRip, SMO, KNN, Logistic, and SGD) based on Mean Absolute Error (MAE) under different filtering techniques (No Filters, Resample, Obfuscate, and Discretize). The performance is evaluated by examining the MAE values, where lower MAE values indicate better performance, as they represent smaller deviations between the predicted and actual values. The MAE values for each classifier under the different filtering techniques are analyzed accordingly.

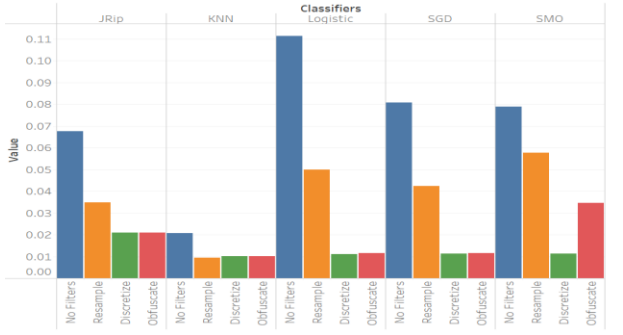


Fig. 3: MAE of Five classifiers

Figure 3 illustrates that, overall, KNN, Logistic Regression, and JRip demonstrate relatively low MAE values across different filtering techniques.. SMO and SGD exhibit higher MAE values initially but show improved performance with specific filtering techniques. The choice of the best classifier and filtering technique would depend on the specific dataset and task requirements, taking into account the importance of minimizing the mean absolute error.The table 5 showed the classification accuracy values for each classifier under the different filtering techniques:

TABLE 5 :CLASSIFICATION ACCURACY OF FIVE CLASSIFIERS

| Classification Accuracy | No Filters | Resample | Obfuscate | Discretize |
|---|---|---|---|---|
| **JRip** | 95.00 | 97.3 | 98.23 | 98.27 |
| **SMO** | 92.11 | 94.23 | 96.53 | 98.88 |
| **KNN** | 98.07 | 99.23 | 99.62 | 98.85 |
| **Logistic** | 90.38 | 95.76 | 98.85 | 98.88 |
| **SGD** | 91.92 | 95.77 | 98.85 | 98.88 |

The above table presents the classification accuracy (%) of five classifiers—JRip, SMO, KNN, Logistic Regression, and SGD—under four different preprocessing techniques: No Filters, Resample, Obfuscate, and Discretize. we can examine the provided accuracy values. KNN had the highest accuracy of 99.23% after resampling the dataset. Logistic, SGD, and JRip showed notable improvements in accuracy, ranging from 95.76% to 97.3%.SMO achieved a moderate accuracy of 94.23%.KNN and Logistic achieved the highest accuracies of 99.62% and 98.85%, respectively, after obfuscating the dataset. JRip, SMO, and SGD showed considerable improvements in accuracy, ranging from 96.53% to 98.85%. SMO achieved the highest accuracy of 98.88% after discretizing the dataset.Logistic, KNN, and SGD achieved accuracies close to 98.88%.JRip had a slightly lower accuracy of 98.27%. Higher accuracy values indicate better performance in correctly classifying instances and the graph is shown in fig 4.

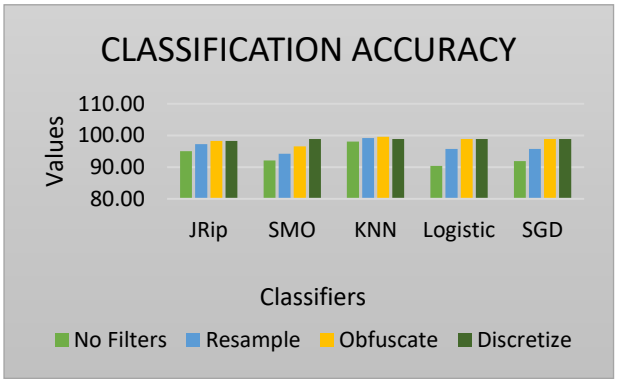


Fig 4: Accuracy of Five classifiers

In assessing classification accuracy across different filtering techniques (No Filters, Resample, Obfuscate, and Discretize) using various classifiers (JRip, SMO, KNN, Logistic, and SGD), the Obfuscate technique outperformed the other filtering methods by demonstrating the highest accuracy. Overall, KNN and Logistic consistently performed well across different conditions, achieving high accuracies. SMO also showed good performance, particularly after obfuscation and discretization. JRip and SGD had slightly lower accuracies compared to the other algorithms but still showed improvements with certain preprocessing techniques. It's crucial to acknowledge that the selection of algorithm and preprocessing methodologies can differ contingent on the distinct dataset and problem domain. Consequently, our findings indicated improved predictive accuracy for diabetes by utilizing filters.

## V. CONCLUSIONS

In conclusion, this research focused on predicting diabetes using five classification techniques (JRip, SMO, KNN, Logistic, and SGD) in the WEKA tool, based on a dataset with 17 medical and demographic attributes. The study emphasized the importance of data preprocessing using three key filters—Resample, Obfuscate, and Discretize—which addressed issues like data imbalance, privacy protection, and data representation. Among the filters, Obfuscate achieved the highest classification accuracy, highlighting its effectiveness in enhancing machine learning outcomes. Overall, the findings demonstrate the value of preprocessing and privacy-preserving techniques in improving early detection of diabetes. This research can support healthcare professionals in making more informed decisions and allocating resources efficiently. Additionally, the methodology can be adapted to other disease prediction models, broadening its impact. Improved automation in medical diagnostics through machine learning could lead to earlier interventions and better patient outcomes. For future work, exploring advanced models such as deep learning and ensemble methods, testing on larger datasets, incorporating real-time and lifestyle data, and applying stronger privacy-preserving methods like differential privacy or federated learning are recommended to further improve prediction accuracy and applicability.

## REFERENCES:

[1] AD Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2020. Diabetes Care. 2019. https://doi.org/10.2337/dc20-S002.
[2] International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation; 2019.
[3] https://en.wikipedia.org/wiki/Medicine
[4] Bisandu, Desmond & Datiri, Dorcas & Onokpasa, Eva & Thomas, Godwin & Haruna, Musa & Aliyu, Aminu. (2019). Diabetes Prediction using Data mining Techniques. International Journal of Innovation Science. 4. 103-111
[5] Bommert A., Sun X., Bischl B., Rahnenführer J., Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data Computational Statistics & Data Analysis, 143 (2020), Article 106839, 10.1016/j.csda.2019.106839
[6] Devi, M. R. (2016). Analysis of various data mining techniques to predict diabetes mellitus. International Journal of Applied Engineering Research, 11(1), 727–730.Google Scholar
[7] Susila, S. Janet Grace, and D. Kavitha. "Corneal Ulcer Feature Extraction and Image Classification using a Deep Convolutional Network and the VGG 16 Model." *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*. IEEE, 2022.
[8] Saeys et al., 2007 Saeys Y., Inza I., Larrañaga P.A review of feature selection techniques in bioinformatics Bioinformatics, 23 (19) (2007), pp. 2507-2517, 10.1093/bioinformatics/btm344
[9] Peralta D., Del Río S., Ramírez-Gallego S., Triguero I., Benitez J.M., Herrera F.Evolutionary feature selection for big data classification: A MapReduce approach Mathematical Problems in Engineering (2015), 10.1155/2015/246139 View article Google Scholar

[10] Zhu Z., Ong Y.-S., Dash M. Markov blanket-embedded genetic algorithm for gene selection Pattern Recognition, 40 (11) (2007), pp. 3236-3248, 10.1016/j.patcog.2007.02.007
[11] Liu and Setiono, 1995 Liu H., Setiono R. Chi2: feature selection and discretization of numeric attributes Proceedings of the international conference on tools with artificial intelligence, IEEE (1995), pp. 388-391, 10.1109/tai.1995.479783
[12] Bernardini et al., 2020 Bernardini M., Romeo L., Misericordia P., Frontoni E. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine IEEE Journal of Biomedical and Health Informatics, 24 (1) (2020), pp. 235-246, 0.1109/JBHI.2019.2899218
[13] Maldonado S., Weber R., Famili F.Feature selection for high-dimensional class-imbalanced data sets using support vector machines Information Sciences, 286 (2014), pp. 228-246, 10.1016/j.ins.2014.07.015
[14] Poornima, V., & Gladis, D. (2021). An Empirical Study of Heart Disease Prediction System Using Various Machine Learning Classification Algorithms. In Proceedings of First International Conference on Mathematical Modeling and Computational Science (pp. 119-129). Springer, Singapore.
[15] Feature extraction: foundations and applications, Vol. 207,Springer (2006)
[16] Rani, Pooja, Rajneesh Kumar, Anurag Jain, Rohit Lamba, Ravi Kumar Sachdeva, Karan Kumar, and Manoj Kumar. "An extensive review of machine learning and deep learning techniques on heart disease classification and prediction." Archives of Computational Methods in Engineering 31, no. 6 (2024): 3331-3349.
[17] Mahdi, Ghadeer, Seror Faeq Mohammed, and Md Kamrul Hasan Khan. "Enhanced Support Vector Machine Methods Using Stochastic Gradient Descent and Its Application to Heart Disease Dataset." Ibn AL-Haitham Journal For Pure and Applied Sciences 37, no. 1 (2024): 412-428.
[18]https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset