# Fuzzy Support Vector Machine Based Outlier Detection for Financial Credit Score Prediction System

R. Ramesh[1*] and M. Jeyakarthic[2]

[1*] Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, India. rameshau04@gmail.com, Orcid: https://orcid.org/0000-0002-1121-2223

[2] Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, India. jeya_karthic@yahoo.com, Orcid: https://orcid.org/0000-0001-6822-6004

Received: July 27, 2023; Accepted: October 04, 2023; Published: December 30, 2023

**Abstract**

Economic Credit Scoring (CS), which aids in calculating the credit worth of both individuals and companies, is regarded as one of the greatest study issues in the finance field. In the banking industry, data mining techniques are believed to be helpful since they help designers and developers create appropriate goods or services for customers with the fewest possible risks. Losses and loan cancellations, which are the major sources of hazards in the banking industry, are related to credit risks. A Support Vector Machine based architecture is presented in the current study for the financial credit score prediction system. However, the existing work tends to have increased computational overhead and that requires complete data for attaining required accurate rate. The system known as the Fuzzy Support Vector Machine based Outlier Detection System is introduced in the suggested research study to address this (FSVM-ODS). This study's first grouping of data items utilising a hybrid genetic algorithm with K-Means clustering algorithm is named (HKGA). The dataset must be gradually lowered in size, and calculation time must likewise be decreased. The Enhanced Z-score (EZS) outlier identification (OD) technique was employed in the second step to identify outliers in the dataset. Then, we use a customized beaver searching method to choose the database. For categorization of the datasets, a fuzzy support vector machine is utilised. The whole study project is carried out in the Matlab simulation environment, and it has been shown that the suggested technique achieves a higher outlier identification rate than the current methodology.

**Keywords:** Credit Score Prediction, Outlier Detection, Support Vector Machine, K Means Clustering, Fuzzy.

## 1 Introduction

Financial credit scoring (FCS) is now a prominent study area as a result of the swiftly spreading business financial crises around the world (Rapaccini, M., 2020). Designing a predictions model for predicting the key risks of the business's finances position in previous years is very difficult in any finance firm or organization (Kim, A., 2020). Usually, the FCS creates a binary classification model that is resolved fairly (Huang, H.W., 2021). The output of the classifier model is divided into two categories: those that indicate a firm's failed situation and those that indicate a firm's non-failure situation (Durica, M., 2019). The classification method's contributions are based on statistical results obtained from information about

*Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA),* volume: 14, number: 4 (December), pp. 60-73. DOI: *10.58346/JOWUA.2023.I4.005*

*Corresponding author: Assistant Professor, Department of Computer and Information Science, Annamalai University, Chidambaram, India.

current firms' finances (Masa'deh, R.E., 2018) (Veeramanikandan, V., 2021). The utilization of various data for FCS has led to the presentation of several classifier algorithms at this time (Shao, L., 2021). The FCS techniques that are now accessible are often divided between statistics and artificial intelligence (AI) models. However, the AI-based FCS models have attracted a lot of attention.

One of the major issues in economics that influences how production and need interacts in the marketplace is the predicting of a good's price (Nam, K., 2019) (Selvarani, S., 2021). According to S Rosen, the quantity of commodities purchased and sold as well as their respective pricing define supplies and need in every market. Desire and supplies behave identically in local and national markets as well as in global commerce (Tien, N.H., 2019) (Venkatesh, S., 2020). Therefore, while predicting prices, it is crucial to determine the supplied and need elements incorporated in the model computations (Tien, N.H., 2019). It is vital to establish the parameters of the market under investigation and discover unique price elements, unique to a certain market since pricing procedures vary across various sector marketplaces. Therefore, utilizing empirical data describing the industrial market under discussion, broader conceptual techniques to pricing prediction must be defined (Degiannakis, S., 2018).

The Multi Commodity Exchange (MCX) of India Ltd. now offers a wide range of methodologies for forecasting the values of the multi commodities derivatives (COMDEX) index traded there (Veeramanikandan, V., 2019). The model for predicting stainless-steel prices took into account a number of variables, including the tones of stainless-steel sheets that were delivered to the market, the number of traveler's cars produced, the tones of stainless-steel sheets needed by car plants, the export prices for stainless-sheets (in US dollars per ton), and the money supply.

The primary feature of this study is the introduction of a machine learning-based approach for predicting the risk associated with economic credit scores. This study aims to present the characteristic selecting and preprocessing techniques, as well as the categorization approach.

The following summarises the general structure of this research project: This section provides a thorough overview to the credit rating system. The overview of many relevant research strategies is provided in section 2. The suggested technique is thoroughly discussed in section 3 along with pertinent examples and figures. Section 4 provides a simulated examination of the study effort. Section 5 concludes the study project overall based on the results of the simulations.

## 2 Related Works

Van Vlasselaer et al (2015) suggested APATE, a cutting-edge method to detect fraudulent credit card purchases made in internet businesses. Utilizing the Regency, Frequency, and Monetary (RFM) principles, our method integrates (1) inherent highlights obtained from the attributes of incoming transactions and the client consumption history with (2) network-based features by utilizing the network of credit card holders and merchants and generating a time-dependent symptoms score for each network object.

Jha et al (2012) used transactional aggregating technique to uncover credit card frauds. For the purpose of estimating models and identifying frauds occurrences, scholars combined data to record customer purchasing behaviour prior to each transaction. In order to aggregate transactions and estimate models, authors utilize actual credit card payment information from an international credit card business.

Leu et al (2015) The Secure M-Commerce System (SMCS for short), which enables users to make a secure credit-card transaction for online buying, has been presented as a secure mobile commerce system. In essence, the SMCS connects a trading system's cash flow with its credit card organizations to efficiently safeguard authorized operations from various threats and prevent data leaks.

Zhang et al (2021) created fraudulent detecting solution that makes use of a deep learning architecture and an innovative features extraction method based on homogeneity-oriented behaviour analysis (HOBA). We undertake a comparison analysis based on a real-world database from Kaggle site to evaluate the efficacy of the suggested framework.

Gianini et al (2020) a technique that provides a normalised score to each unique rule, measuring the rule's contribution on the pool's total efficiency, is suggested to be used to the rule selection for the NRT phase. Lucas et al (2020) modeled three various viewpoints on a series of credit card transactions, including (i) whether or not the series involves fraud. (ii) The sequence is determined by securing the cardholder or the payment terminal (iii) It is a sequence of the total expenditure or of the length of time that has passed between the current and prior transactions. Eight sets of sequences from the (training) set of transactions result from configurations of the three binary views. These events are each analyzed using a Hidden Markov Model (HMM).

Jurgovsky et al (2018) Long Short-Term Memory (LSTM) networks were used to include transaction sequences and frame the fraudulent identification issue as a sequence classification challenge. We also use cutting-edge features aggregating techniques, and we present the findings using conventional retrieving measures.

Bahnsen et al (2016) A novel set of features based on extending the transactions aggregating approach and exploring the periodic behaviour of transaction time using the von Mises distributions are proposed. Then, we compare cutting-edge credit card fraud detection models and assess how the various feature sets affect the outcomes using an actual credit card fraud dataset given by a major European card processing business. The findings demonstrate an average gain in savings of 13% when the recommended periodic characteristics are included into the procedures.

de Sá et al (2018) Introducing Fraud-BNC, a tailored Bayesian Network Classifier (BNC) algorithm for a genuine credit card fraud detection issue. A Hyper-Heuristic Evolutionary Algorithm (HHEA), which classifies the information on the BNC techniques into a taxonomy and looks for the most effective combination of these elements for a particular database, was used to automatically create Fraud-BNC. Using a database from PagSeguro, the most widely used online payment service in Brazil, fraud-BNC was automatically created and evaluated alongside two approaches for dealing with cost-sensitive categorization. Results were matched to seven other algorithms, and the method's economic effectiveness and the data categorization issue were taken into consideration.

## 3 Outlier Detection for Financial Credit Score Prediction System

A credit loan's risk assessment is called a credit risk score. It gauges the likelihood of default or delinquency. Prediction modelling employing machine learning techniques is the best method for predicting the probability of default or delinquent. Credit risk scores may be calculated using comparative numeric evaluations or standard likelihood.

In this research work, The Hybrid Genetic Algorithm with K-Means Clustering algorithm is used to first cluster data items (HKGA). The database size has to gradually decrease, which also cuts down on processing time. The Enhanced Z-score (EZS) outlier detection (OD) technique was employed in the second step to identify outliers in the database. And then feature Selection of the dataset is done using modified squirrel search algorithm. For categorization of the information, a fuzzy support vector machine is utilized (Salman, R., 2023).

**Data Clustering Using Hybrid Genetic with K Means Clustering Algorithm**

One of the most popular grouping techniques is the K-means technique, which has been applied in many technological and scientific disciplines. The fact that the k-means method may result in blank groups dependent on the initial centered vectors is one of its main issues. Genetic algorithms (GAs) are adaptive heuristic search algorithms based on natural selection and genetics, which are developmental concepts. This study offers a hybrid k-means method with GAs that effectively solves the vacant clustering issue and groups the data items into groups.

One of the most popular grouping algorithms is the K-means cluster method, which has been employed in many different scientific and technological sectors. The following are the main issues with the K-means algorithm:

- Based on the original center vectors, it can result in blank groups.
- Could lead to suboptimal numbers.
- With a decent level of computing work, it is impossible to find worldwide answers for huge issues.

This study proposes the HKGA, a hybrid genetic algorithm with K-means clustering that effectively removes this flaw.

*Phase 1: K-Means Algorithm*

Step 1: K initial cluster centres$z_1, z_2, z_3, \dots, z_k$arbitrarily selected from the n data$\{x_1, x_2, x_3, \dots, x_n\}$.

Step 2: A point $x_1, i = 1,2,3, \dots, n$ is assigned to cluster $C_j, j \in \{1,2, \dots, k\}$if

$$\|x_i - z_j\| < \|x_i - z_p\|, p = 1,2, \dots, K \ \& \ j \neq p \qquad (1)$$

Step 3: New cluster centres$z_1, z_2, z_3, \dots, z_k$ are computed as follows:

$$z_j^* = \frac{1}{n_i}\sum_{x_i \in C_j} x_i, i = 1,2, \dots, K; \qquad (2)$$

Where $n_i$is the total amount of components in cluster$C_j$.

Step 4: If $z_i^* = z_i, i = 1,2, \dots, K$if so, stop; if not, start again at step 2.

We get an initially centre for each predefined group upon this procedure.

*Phase 2: Genetic algorithm*

Step 1: Population initialization

Each person represents a row-matrix of 1x n observations, and each gene contains the integer [1, K] that denotes the cluster to which this information resides. Take, for instance, 10 observations.$\{x_1, x_2, \dots, x_{10}\}$this has to be matched with the fourth group, k = 4.

Step 2: Evaluation

Determine the desired goal value, and then search for acceptable cluster categories that minimizes the efficiency value. The fitness function for the K clusters in grouping$C_1, C_2, C_3, \dots, C_k$ *is* given by

$$f(C_1, C_2, \dots C_k) = \sum_{i=1}^{k}\sum_{x_j \in C_i}\|x_j - z_i\| \qquad (3)$$

Step 3: Selection

To focus GA search on potential areas of the search space is the goal of selection. In this work, we use a roulette wheel selection method, where individuals from every generation are chosen based on a chance value to survive into the subsequent generations. Following the equation below, the likelihood of variable selection is inversely correlated with its survival rating in the community.

$$p(x) = \frac{f(x) - f_{Min}(\Psi)}{\sum_{x \in \Psi}\{f(x) - f_{Min}(\Psi)\}}; \qquad (4)$$

Where $p(x)$, Possibility that a string x will be selected in a population$\Psi$ and

$$f_{Min}(\Psi) = \text{Min}\{\text{f(x)}|\text{x} \in \Psi\} \qquad (5)$$

Step 4: crossover operator

The crossing is performed on every person in this stage using a customized regular crossing, where the offspring is created by selecting the person with a probability.

Step 5: operating a mutant

The following is the implementation of the modification operation for every person: first, choose two columns at random from the $i^{th}$ individuals, then create two new columns.

Step 6: the best choices identified so far throughout the process, as opposed to GA keeping the best choices identified within the current population.

**Outlier Detection Using Enhanced Z Score Method**

Although many big datasets include data objects that are imperfect owing to noise or outliers, the outlier identification technique is employed to preprocess the original data set. The suggested Enhanced Z-score technique (EZS) outlier identification methodology improves data quality and yields precise findings. HKGA is used to first cluster data items, after which the centroid of each cluster is determined. A data point may be described in terms of how it relates to the mean and normal variation of a set of points using the Z-score, also known as the normal scoring.

A measurement's Z-score is described as

$$Z_i = \frac{x_i - \bar{x}}{sd} \quad (6)$$

where$X_i \sim N\ (\mu, \sigma^2)$,where SD represents standard deviation.

By eliminating the impacts of the data's size and location, Z-scores enable direct comparisons across various databases. The idea behind the Z-score technique of outlier identification is that everything that deviates too much from zero after the data has been scaled and centred (the threshold is often a Z-score of 3 or -3) should be regarded as an outlier.

If the database contains less than 12 items, the Z-score approach will never identify an outlier. This led to the creation of the Enhanced Z-score system (EZS), which is not constrained by the same issues.

The sample mean and sample standard deviation(s), which were utilised as estimators in the preceding Z-Scores issue, may be impacted by a few extreme values or even by a single extreme number. To remedy this issue, the updated Z-Scores use the median and the median of the absolute deviation (MAD) instead of the sample's mean and standard deviation, respectively.

$MAD = median|x - m|$,where m is the median of the samples. The calculation for the Enhanced Z-Score (E) is

$$E = \frac{0.6745(x-m)}{MAD} \quad (7)$$

where$E(MAD)\ =\ 0.675\ \sigma$ for large normal data.

In their simulation using pseudo normal observations for sample sizes of 10, 20, and 40, Iglewicz and Hoaglin (1993) proposed that observations are classified as outliers when |E|>3.5. Similar to the Z-score, the altered Z-score is useful for normal data.

**Feature Selection Using Modified Squirrel Search Algorithm**

Feature selection is done using modified searching method for squirrels. The default method for finding squirrels is based on food foraging nature of the flying squirrels. The squirrel search algorithm's primary

objective is to identify the optimal food resources to ensure flying squirrels to get enough energy. Squirrel search algorithm seems to consume more computational time and predicting the optimal solution required more convergence. This is resolved by introducing the modified squirrel search algorithm in which Gbest parameter is added additionally in standard squirrel search algorithm. In the proposed modified squirrel search algorithm Gbest for every solution will be found based on which position updation will be carried out instead of random position update. The proposed algorithm will find out the optimal features which will ensure the accurate classification rate.

**Algorithm:** Feature selection using modified squirrel search algorithm
**Input:** Input attributes
**Output:** Optimal features
1: Begin
2: Do
3: Initialize the population
Initialize the location
$LS_i = (LS_{i1}, LS_{i2},\ldots,LS_{id})$
where i=1,2,…,n
$LS_{ij}$ = jth location of ith squirrel
n= population

$$d_g = \frac{h_g}{\tan(\varphi) \times sf}$$

Where $d_g$ → Gliding distance
$h_g$ → constant value = 8
$s_f$ → constant value = 18

$$\tan(\varphi) = \frac{D}{L}$$

Where tan $(\varphi)$→ gliding angle
D → drag force
L → Lift force

$$D = \frac{1}{2\rho V^2 S C_D}$$

$$L = \frac{1}{2\rho V^2 S C_L}$$

Where ρ→ constant =1.204 kg $m^{-3}$
V → constant = 5.25 $ms^{-1}$
S → 154 $cm^2$
$C_D$ → constant = 0.6
$C_L$ → random number between 0.675 and 1.5
4: For each individual in population
5: Find the fitness value (Classification accuracy)
6: Find the Gbest value
If (bestFit > Fit)
Update bestFit as Gbest
Else
Update Fit as Gbest
7: Update the position of parameters based on Gbest value
If $r > P_{dp}$

$$LS_i^{t+1} = LS_i^t + d_g \times G_c \times \left(F_h^t - LS_i^t\right)$$

If (fitness of current squirrel < Gbest)

$LS_i = LS_L + U(0,1)$ x Gbest $(LS_u - LS_L)$

Else

Choose random location

End IF

End IF

Where

$LS_L$ and $LS_U$ = lower and upper bounds of squirrel

8: End for

9: If all individuals' position is not updated

10: Go to step 3

11: Else

12: Find the best parameter values based on Gbest value

13: Find the seasonal transitional judgment and Update the position of individual

$$S_c^t = \sqrt{\sum_{k[1}^{D}(F_{ai,k}^t - F_{h,k}^t)^2} \quad i = 1,2,\ldots,Nf_s$$

$$S_{min} = \frac{10e^{-6}}{\frac{t}{(365)^{\frac{T}{2.5}}}}$$

Where T → number of iteration

If $S_c^t < S_{min}$

Winter over and season turns to summer

Else

Season unchanged

End if

If season = summer and present in $f_h$

Stay at the previously updated position

Else if season =summer and present in $f_a$

Update the position as like in below eq

$$LS_{inew}^{t+1} = LS_L + \text{Levy}(n) \times (LS_U - LS_L)$$

Where Levy → random walk model

$$\text{Levy}(x) = 0.01 \times \frac{r_a \times \sigma}{|r_b|^{1/\beta}}$$

Where β→ constant = 1.5

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2^{\left(\frac{\beta-1}{2}\right)}}\right)^{\frac{1}{\beta}}$$

where $\Gamma(x) = (x-1)!$

14: End if

15: Repeat till convergence reached

16: End

The above algorithm will resultant with the optimally selected features for classification.

**Classification Using Fuzzy Support Vector Machine**

In this segment, classifying is complete to forecast whether the credit risks are present or not. In this research work, Multiclass SVM is rummage-sale for the prediction resolution where the prediction decision is prepared by means of the fuzzy decision-making scheme. Thus, SVM classification is prepared on the source of fuzzy decision-making procedure. The innovative model, which differs from obsolete Classification method, takes the notion of fuzzy theory into consideration. The sign function in the prediction step of the categorization technique is likewise replaced with a fuzzy decision-making problem. The suggested method employs the selection values as the self-governing variable of a fuzzy decision-making functional in the forecast section to classify the test data set into several groups, rather than only the sign of which. In real-world situations, when interactions and noisy effect are present at the border of several groupings, this flexible style of decision-making replicates more approaches.

Theoretically, fuzzy decision-making SVM dispensing is based on a database analysis, an SVM process, and a large number of trials. Traditional decision-making theory is included but disregarded in order to consider the dataset to be clean. However, as we already said, not all input points are correctly classified during the SVM classifier distribution era and not all input points may not be accurately separated in many real applications. In other words, whereas certain common input points are simpler to identify, others provide organisational challenges. Deciding values is derived from the predicted choice mechanism of SVM, which is a sign job. These values associated to different input locations identify the next two levels between these points and the ideal distinctive hyper plane, so they may be used as independence factors in the fuzzy decision-making function proposed in this research. We can model fuzzy decision-making SVM dispensing into three phases: SVM straining, choice valued forecasting, and fuzzy decision-making processing. Figure 1 shows fuzzy SVM dispensation.

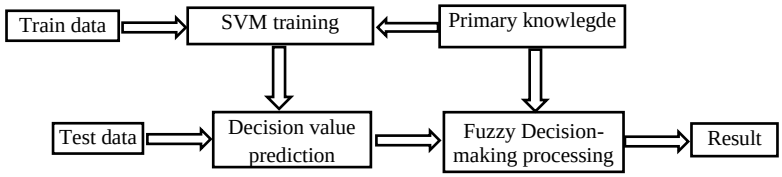


Figure 1: Fuzzy Decision Making SVM Processing

In fuzzy decision-making processing, we develop a fuzzy replica to rebuild SVM's decision-making using fuzzy sets.

***Fuzzy Decision-Making Model***

In traditional ways, sign function is continually cast-off to break test data into classes, thus zero is the sole important threshold. In many cases, class ties aren't that modest. Interaction happens in both groups, particularly near their boundaries. The noise concerns the other class's points more quickly. Due to the impacts, the deciding value is no longer decipherable and zero is ineffective. Although many categorization algorithms aren't good sufficient and certain train databases aren't popular for forecasting, many miscategorised instances occur around threshold. To replace this rigid division, we build a fuzzy border between two fuzzy sets with binary labels. From the previous step, we can retrieve the evaluation number of each input vector.

A fuzzy decision typical might be produced using decision value as an independent variable, assuming that these two fuzzy sets are referred to as A (the values in this set are judged to be projected to -1) and B (the values in this set are thought to be expected to +1). In fuzzy theory, the border between sets A and B is replaced with a grey zone rather of a bright one. Decision values in this area cannot be

categorised as set A or set B in a strict sense. The boundary functions are the choice values as independent variables in this suggested method, and the process values may generate their constancy, which seizes the idea of complexity. Degree 1 in this instance denotes perfect plausibility, whereas Degree 0 denotes utter falsehood. An extremely stretchable duplicate is necessary to accurately reflect the impending real-world scenarios. The borders of sets A and B may be distinguished as trails if the deciding value is denoted by v and changed by explicitly changing the fuzzy decision value to the assortment from 0 to 1:

$$f_A(v) = \frac{arctan(-v.s-d.s)}{\pi} + 0.5 \qquad (8)$$

$$f_B(v) = \frac{arctan(-v.s-d.s)}{\pi} + 0.5 \qquad (9)$$

The above process can be cast-off to forecast the outlier occurrence correctly and early. A simulated platform called Matlab is used to evaluate the study effort elucidated thorough in the subsequent segment.

**Simple Explanation of Algorithms**

In this work dataset from kaggle site is taken for credit score prediction system. This dataset will consist of more features which should be processed and learned for the credit score prediction outcome. Here initially preprocessing is carried out to enhance the input dataset, so that classification can be burden less. In the first step, clustering is done on input dataset, where similar kind of data's will be grouped together based feature distance of dataset. And then outlier detection is carried out where variation among dataset is calculated based z score value and most varied data are considered for the next step to improve the classification accuracy. These two steps are considered as preprocessing.

After preprocessing feature is carried out. Data set will consist both set of attributes that are relevant to credit score prediction and irrelevant to credit score prediction. Those irrelevant features need to be avoided from the classification process to reduce the computation overhead. So this is done in this work using feature selection process.

And finally classification is carried out to find the credit score accurately. Here the feature values pattern will be analyzed. For example, for different data, loan payment, loan pending will be learned. It will find like what is the credit status if loan pending value is more than threshold and what is the credit status if loan payment value is less than threshold. This will be learned using SVM algorithm. Based on this final prediction will be done.

# 4 Results and Discussion

This section analyses the efficiency improvements of the suggested and current research techniques via a numerical assessment of the suggested methodological approach in terms of several efficiency indicators. The suggested research technique is put into practise using the Matlab simulation environment. The following is a list of the effectiveness metrics taken into account in this work: Precision, Sensitivity, Selectivity, and Reliability. The comparison is made between the proposed method Fuzzy Support Vector Machine based Outlier Detection System (FSVM-ODS) and the existing methodologies DNN.

**Accuracy**

It is described as the accuracy with which outliers in the dataset were identified. There are fewer false positives now. The suggested system's accuracy ought to be superior than existing systems like

ADRMine. The true positive, false positive, true negative, and false negative values of the suggested system are used to determine the average reliability. The reliability is determined in a manner similar to this:

$$\text{Accuracy} = \frac{T_p}{(T_p+F_p+F_n)} (10)$$

Table 1: Accuracy Comparison Values

| Method | Decision Tree | SVM | DNN | FSVM-ODS |
|---|---|---|---|---|
| Accuracy | 79 | 84 | 86 | 89 |

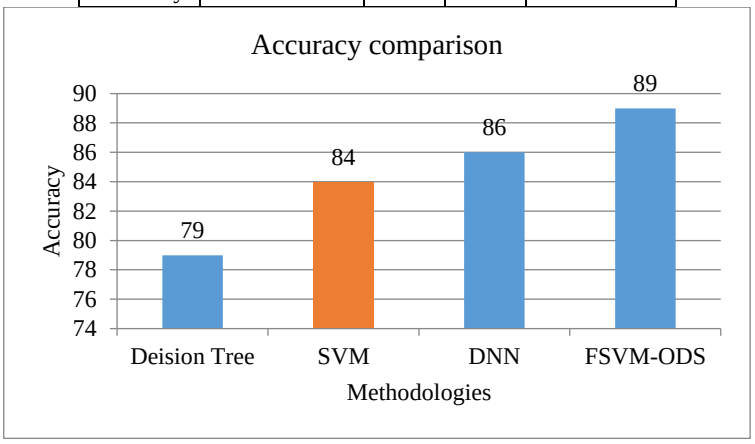


Figure 2: Accuracy Comparison

The comparing of the reliability metrics is shown graphically in Figure 2. This graph demonstrates how superior the suggested research approach is to the ones now in use. FSVM-ODS outperforms DNN in accuracy by 3%.

**Sensitivity**

The percentage of outliers that are correctly identified as positive is known as the true positive rate. The following can be expressed mathematically:

$$\text{Sensitivity} = \frac{\text{number of true positives}}{\text{Number of true positives+Number of false negatives}} \quad (11)$$

Table 2: Sensitivity Comparison Values

| Methods | Decision Tree | SVM | DNN | FSVM-ODS |
|---|---|---|---|---|
| Sensitivity | 81 | 85 | 92 | 96 |

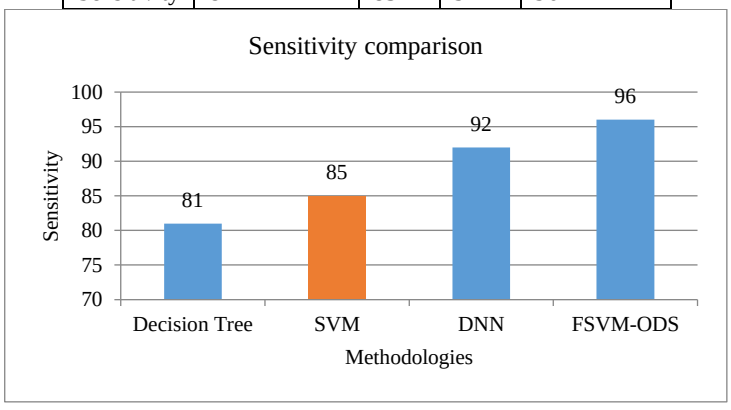


Figure 3: Sensitivity Comparison

The susceptibility metric difference is shown graphically in Figure 3. This graph demonstrates how superior the suggested research approach is to the ones now in use. Comparing FSVM-ODS to the previous DNN, responsiveness is up 4%.

**Specificity**

It is also known as true negative rate, and is measured by how well non-outliers are classified. Mathematically, this may be expressed as follows:

$$\text{Specificity} = \frac{\text{number of true negatives}}{\text{Number of true negatives+Number of false positives}} \quad (12)$$

Table 3: Specificity Values

| Methods | Decision Tree | SVM | DNN | FSVM-ODS |
|---|---|---|---|---|
| Specificity | 65 | 69 | 73 | 79 |

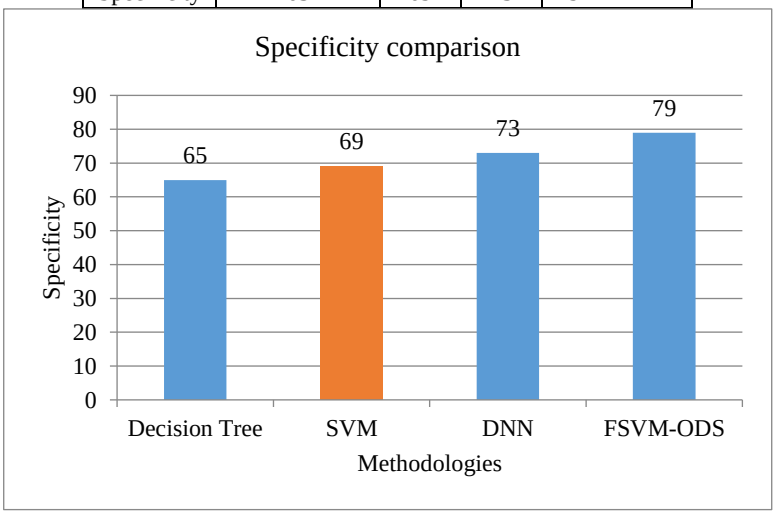


Figure 4: Specificity Comparison

The graphical depiction of the specificity metric comparison is shown in Figure 4. This graph demonstrates that the suggested research approach outperforms the currently used research methods. FSVM-ODS achieves 6% increased specificity than DNN.

**Precision**

It is the percentage of appropriate results from searches

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (13)$$

Table 4: Precision Comparison Values

| | Decision Tree | SVM | DNN | FSVM-ODS |
|---|---|---|---|---|
| Precision | 67 | 72 | 77 | 81 |

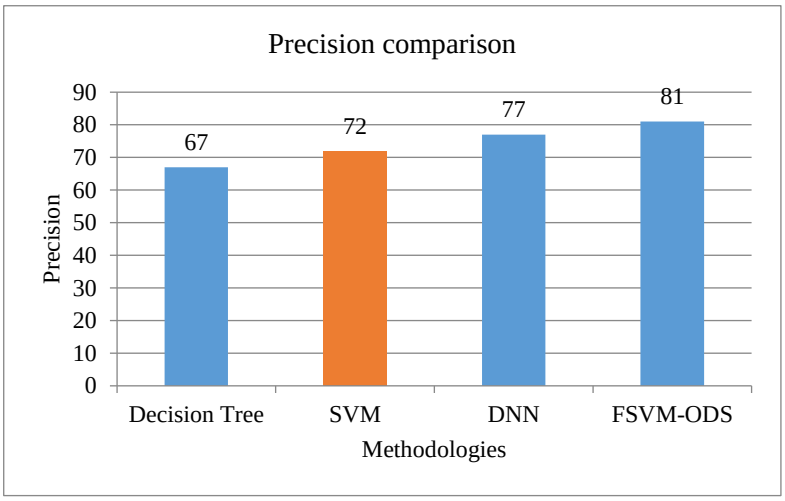


Figure 5: Precision Comparison

The graphical depiction of the accuracy metric comparison is shown in Figure 5. This graph demonstrates how superior the suggested research approach is to the ones now in use. DNN achieves 4% more accuracy than FSVM-ODS.

## 5 Conclusion

We present an approach to data analysis in this ground-breaking study, utilizing cutting-edge methods to improve outlier identification and optimize processing efficiency. Our methodology consisted of two steps: first, data items were clustered using the hybrid genetic algorithm and K-Means clustering technique, or HKGA. We were able to dramatically cut processing time by progressively reducing the dataset size thanks to this creative method. The second phase was identifying outliers in the dataset using the Enhanced Z-score (EZS) outlier identification technique. We included a modified squirrel search approach to further limit our choices, making sure the dataset selected for study was accurate and ideal. A fuzzy support vector machine, a potent machine learning technique, was used to accurately categorize the data. We were able to classify the data efficiently because to this sophisticated algorithm, which also gave us important insights into the intricate patterns included in the data. The powerful Matlab modeling framework, renowned for its analytical power and versatility, served as the study's operating environment throughout. Our findings showed that the outlier identification rates of our suggested strategy were higher than those of the existing methodologies, proving its superiority in data analysis. In addition to its scholarly value, our novel model has significant real-world implications. It may be easily incorporated into real-time banking systems, offering crucial assistance in quickly identifying irregularities and fraudulent activity. It can also be utilized in consumer behavior decision support systems, which help companies make decisions based on thorough and precise data analysis.

## Reference

[1] Bahnsen, A.C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications, 51*, 134-142.

[2] de Sá, A.G., Pereira, A.C., & Pappa, G.L. (2018). A customized classification algorithm for credit card fraud detection. *Engineering Applications of Artificial Intelligence, 72*, 21-29.

[3] Degiannakis, S., Filis, G., & Arora, V. (2018). Oil prices and stock markets: A review of the theory and empirical evidence. *The Energy Journal*, *39*(5), 1-46.

[4] Durica, M., Podhorska, I., & Durana, P. (2019). Business failure prediction using cart-based model: A case of Slovak companies. *Economic & Managerial Spectrum/Ekonomicko-manaz? e´ rske spektrum*, *13*(1), 51-61.

[5] Gianini, G., Fossi, L.G., Mio, C., Caelen, O., Brunie, L., & Damiani, E. (2020). Managing a pool of rules for credit card fraud detection by a Game Theory based approach. *Future Generation Computer Systems*, *102*, 549-561.

[6] Huang, H.W., Hsu, B.W.Y., Lee, C.H., & Tseng, V.S. (2021). Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers. *The Journal of dermatology*, *48*(3), 310-316.

[7] Jha, S., Guillen, M., & Westland, J.C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert systems with applications*, *39*(16), 12650-12657.

[8] Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications*, *100*, 234-245.

[9] Kim, A., Yang, Y., Lessmann, S., Ma, T., Sung, M.C., & Johnson, J.E. (2020). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *European Journal of Operational Research*, *283*(1), 217-234.

[10] Leu, F.Y., Huang, Y.L., & Wang, S.M. (2015). A Secure M-Commerce System based on credit card transaction. *Electronic Commerce Research and Applications*, *14*(5), 351-360.

[11] Lucas, Y., Portier, P.E., Laporte, L., He-Guelton, L., Caelen, O., Granitzer, M., & Calabretto, S. (2020). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. *Future Generation Computer Systems*, *102*, 393-402.

[12] Masa'deh, R.E., Al-Henzab, J., Tarhini, A., & Obeidat, B.Y. (2018). The associations among market orientation, technology orientation, entrepreneurial orientation and organizational performance. *Benchmarking: An International Journal*, *25*(8), 3117-3142.

[13] Nam, K., & Seong, N. (2019). Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. *Decision Support Systems*, *117*, 100-112.

[14] Rapaccini, M., Saccani, N., Kowalkowski, C., Paiola, M., & Adrodegari, F. (2020). Navigating disruptive crises through service-led growth: The impact of COVID-19 on Italian manufacturing firms. *Industrial Marketing Management*, *88*, 225-237.

[15] Salman, R., & Banu, A.A. (2023). DeepQ Residue Analysis of Computer Vision Dataset using Support Vector Machine. *Journal of Internet Services and Information Security (JISIS)*, *13*(1), 78-84.

[16] Selvarani, S., & Jeyakarthic, M. (2021). Rare Itemsets Selector with Association Rules for Revenue Analysis by Association Rare Itemset Rule Mining Approach. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *14*(7), 2335-2344.

[17] Shao, L., Fu, C., You, Y., & Fu, D. (2021). Classification of ASD based on fMRI data with deep learning. *Cognitive Neurodynamics*, *15*(6), 961-974.

[18] Tien, N.H., Phu, P.P., & Chi, D.T.P. (2019). The role of international marketing in international business strategy. *International journal of research in marketing management and sales*, *1*(2), 134-138.

[19] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38-48.

[20] Veeramanikandan, V., & Jeyakarthic, M. (2019). Forecasting of Commodity Future Index using a Hybrid Regression Model based on Support Vector Machine and Grey Wolf Optimization Algorithm. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *10*(10), 2278-3075.

[21] Veeramanikandan, V., & Jeyakarthic, M. (2021). Parameter-tuned deep learning model for credit risk assessment and scoring applications. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *14*(9), 2958-2968.

[22] Venkatesh, S., & Jeyakarthic, M. (2020). Adagrad Optimizer with Elephant Herding Optimization based Hyper Parameter Tuned Bidirectional LSTM for Customer Churn Prediction in IoT Enabled Cloud Environment. *Webology*, *17*(2), *631-651*.

[23] Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences*, *557*, 302-316.

## Authors Biography

R. Ramesh is currently working as an Assistant Professor in Department of Computer and Information Science, Annamalai University, Annamalai Nagar. He received his MCA degree from Periyar University and MPhil degree from Annamalai University, Annamalai Nagar. His current research areas are business intelligence, big data analytics, and natural language processing.

M. Jeyakarthic is currently working as an Assistant Professor in Department of Computer and Information Science, Annamalai University, Annamalai Nagar. He received his MCA and MPhil degrees from Madurai Kamaraj University, Madurai, PhD in Computer Science and Engineering, and MBA in E-Business from Annamalai University, Annamalai Nagar. He had published around 75 papers in international Journals. His current research areas are business intelligence, big data analytics, natural language processing and wireless networks.