Fraud Detection in Health Insurance Using a Combination of Feature Subset Selection Based on Squirrel Optimization Algorithm and Nearest Neighbors Algorithm Methods

Kian Parnian [1], Farid Sorouri[2], Alireza Najafi Souha[2], Aidin Molazadeh[3], Sahar Mahdavi[2]

1- Department of Computer Engineering, Moghadas Ardabili Institute of Higher Education, Ardabil, Iran
Email: parnian.kian@gmail.com (corresponding author)
2- Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran
Email: fa.sorouri@gmail.com
Email:a.najafi@lauardabil.ac.ir
Email: sr.mahdavi73@gmail.com
3-National Industrial Group, Tehran, Iran
Email: Dr.molazadeh1@gmail.com

Received 1 Apr 2021 ***Revised*** 3 May 2021 ***Accepted*** 31 May 2021

ABSTRACT:
Insurance fraud is one of the most expensive financial crimes in economics. Deliberate fraud in insurance companies that leads to the illegal payment of insurance benefits to an individual or group is known as insurance fraud. Today, large amounts of data are stored in real-world databases on insurance transactions, and this amount continues to grow rapidly. Therefore, there is a need for semi-automated methods to detect hidden knowledge in such databases. On the other hand, due to the increasing variety of fraudulent methods in health insurance, the features and characteristics of fraud samples compared to normal samples have increased widely, only a part of these features will be useful to build a fraud detection model. Therefore, in this study, to detect fraud in health insurance, a combination of feature subset selection based on squirrel optimization algorithm and nearest neighbor classification has been used. In the proposed method, educational data set obtained from Ardabil's Social Security Insurance Organization will be used to determine the patterns of fraud detection in health insurance and the test data set prepared from this data set to evaluate the model. According to the results of the implementation of the proposed method, it can be seen that the proposed method, by eliminating irrelevant features and finding useful features, has been able to obtain good results for classifying and predicting fraud in the Social Security Organization database and it has and it has a higher accuracy rate than other data mining classification algorithms in this field.

KEYWORDS: Fraud Detection, Squirrel Optimization Algorithm, Nearest Neighbor Algorithm, Social Security Insurance

1. INTRODUCTION

Today's demands for services require global access all the time. Fintech represents the use of IT solutions in business models to provide better financial services to customers. However, this term is still very much debated. In fact, fintech is a term for a wide range of technologies that dynamically interact in a common infrastructure. The term also means the constant coexistence of technology and finance. Companies that use this business model offer benefits such as easier use and cheaper and safer transactions [1]. Fintech services have become more attractive to both customers and providers. This fact is further confirmed by the steady increase in fintech investments over the past few years. In the future, this technology may perform better and even replace traditional financial institutions [2].

In practice, fintech relies on a variety of payment methods, such as credit cards and financial transactions, which include digital currencies. The second one is based on blockchain technology, which provides a direct link to financial institutions. In fintech, financial transactions represent repetitive processes in which sensitive information is exchanged between two peers. Existing artificial intelligence technologies often complement fintech services by implementing these processes in an automated and secure manner. In addition to capabilities, business models used for fintech must ensure information security. Since the fintech business model relies on existing information technology (IT) infrastructure, financial activities can be exploited. In general, fraudulent actions target specific weaknesses of financial activities including credit cards, financial transactions and blockchain technology. Malicious activity is carried out by criminals or multiple criminals and can have severe consequences. In fact, only a minority of organizations implement any anti-fraud mechanism. After that, only a small minority of fraud victims will fully recover. Identification of such malicious actions is a major technical challenge for companies and organizations [3, 4].

The term "fraud" is to earn money from an organization or individual without necessarily having a direct effect on the law. In this competitive world, fraud becomes a critical problem if it is normal and prevention policies are not followed. Fraud detection, if done manually using a process such as screening or verification, will be a normal process that does not lead to prevention. Thus, automation is the only way to do this efficiently [5, 6, 7]. It can be effectively done by using artificial intelligence to perform predictive analysis [8, 9, 10] in all areas.

Fraud is one of the challenges that the insurance company has been facing for a long time and it constitutes a significant part of the damages caused to it. However, insurance fraud has become a serious problem worldwide [11]. Fake health insurance can take many forms and can cause great damage to insurance companies. The customer who tries to use health insurance with fake names are the owners of the organization who can use the information and insure the claim if this has never happened (fake claims) or as much as necessary. To exaggerate the fact (claim), there is all kinds of fraud in insurance companies that lead to financial casualty insurance. Because the extent of insurance fraud can be investigated, the use of insurance fraud is generally controversial [12]. In the United States, the annual failure of insurance fraud ranges from $ 40 billion (Federal Bureau of Investigation in 2015) to about $ 80 billion (Alliance Against Insurance Fraud in 2015). Studies in the U.S. Property & Casualty Insurance show that there are 10% of all reported fraud claims. 2015 National Insurance Crime Bureau in 2015 estimated the fraud practices costs at € 4 billion in damages for Germany. The survey shows that 4% of all German households have committed insurance fraud in the last 5 years, and another 7% of respondents have information about insurance fraud behavior among their acquaintances.

In the insurance industry, the use of anti-fraud technology and Special Investigation Units to detect insurance fraud is very common. These systems rely on both visible claims and information from the owners of the organizations. Remarkable features of the owner of the organization can be an interesting feature that accompanies you. These features are related to the ethical and social considerations of medicare owners and thus promotion of fraud. To improve such systems, it is better to have more influence, which may be important in increasing the need to promote frauds [13]. Why examining your data and checking your features can be an important step in detecting early fraud by insurance companies. Data mining is a science that has solutions for using important information and valuable models to identify and detect defrauders in health insurance.

For this reason, fintechs and the insurance industries use intelligent methods from the field of Machine Learning (ML) to detect suspected fraud patterns. ML includes anomaly detection techniques that automatically detect and classify suspicious data from financial networks. Methods such as learning algorithms, statistical models and artificial neural networks are used to generate models from a data set. In the next step, the resulting representation is observed in order to extract appropriate techniques and policies to prevent fraud.

The continuation of the article is organized as follows. In section 2, related works will be described. Details of the proposed method will be presented in the third section. In the fourth section, the implementation and evaluation of the proposed method will be stated. In the fifth section, the conclusion of the article will be presented.

2. RELATED WORKS

In 2021, Haque et al. formulated the problem of fraud detection on the minimum definitive claim data consisting of medical diagnostic codes and procedures. In this paper, a solution to detect the problem of fraudulent claims has been presented using a new agent learning method, which translates diagnostic and methodological codes into a mixture of clinical codes. Suffixes of the mixture of clinical codes have also been investigated using short -term and long -term memory networks and principal component analysis . Experimental results show promising results in identifying fake records [14]. In 2020, Kunickaitė et al. used machine learning (decision trees, hybrid classification, random and reinforcement forests) to detect health insurance fraud. The performance of the model has been evaluated using accuracy, error rate, recall and accuracy. The best results have been obtained using the hybrid classification method [15]. In further research, the analysis of the application of deep learning models [16, 19] and anomaly detection methods [20, 23] will be useful.

In 2020, Ramandi et al. identified processes, key factors, the effects of fraud in complementary health insurance, obstacles and challenges in these processes by comparative study of successful experiences of leading countries in the field of anti-fraud campaign of complementary health insurance and also interviewing experts in this field. Finally, solutions were provided to prevent and control this phenomenon. The interviews were uploaded in text format in MAXQDA software and then were analyzed [24]. In addition, fraud detection was also widely used in network applications [25, 34].

In 2020, Shamita et al. clarified a framework for identifying fraud by learning faster and identifying maximum cases of fraud. Common problems, such as data heterogeneity and unbalanced classification of classes, were also discussed in this paper. As part of developing an efficient framework for detecting fraud, we used several learner and optimization techniques. This framework has been evaluated with a set of claims data obtained from CMS Medicare. Finally, they concluded that the use of multilayer perceptron, a leading neural network with genetic algorithm optimization, has helped to increase results and achieve higher accuracy. Principal component analysis was also used to select the most important variables. The use of principal component analysis and other appropriate pre-processing techniques has also helped reduce training time, resulting in efficiency and speed [35].

3. SUGESTED METHOD

A typical fraud detection system consists of several layers of control [36], each of which can be performed automatically or under human supervision. Part of the automated layer includes machine learning algorithms that generate predictive models based on labeled interactions. Over the past decade, intensive machine learning research to detect fraud in financial and credit companies has led to the development of supervised, unsupervised and semi-supervised learning techniques [37, 39]. In this study, a combination of feature subset selection based on squirrel optimization algorithm and nearest neighbor classification has been used to detect fraud in health insurance. Feature selection in fraud detection methods can increase the accuracy of fraud detection [40]. Therefore, it is one of the basic steps in this field. The purpose of selecting features is to eliminate the difficulty of classification and increase the accuracy of classification by selecting the relevant features. In feature selection tasks, regardless of training cost or number of features, the best combination of features for optimal classification performance is found. As a result, the solution to the feature optimization problem is a set of optimal features as a solution in which the amount of the proportionality of each two-component vector solution is the number of features and the amount of classification error [41]. Using the feature selection problem as a minimization problem, the goal is to minimize the number of irrelevant features and minimize the classification error. In the following, we will formulate the proposed method.

3.1 PROBLEM FORMULATION

In this research, the feature selection problem is considered as an optimization problem that is solved using the squirrel optimization algorithm. The objectives of this optimization problem are classified into two general groups, one is to reduce the number of features and the other is to reduce the amount of classification error. Accordingly, the number of data set features, independently, indicates the size of the problem or the number of trees that a squirrel can fly on. It is in the range of [0,1] in the binary search space. In the final solution, the most suitable features are selected as oak tree. Oak trees are the best food source for squirrels; so finding these trees is very useful for squirrels to fly.

In the decision search space, x1 is a real positive number that indicates the error rate, and x2 is a real positive number that indicates the number of attributes. The F function results in a balanced set of decision vectors that both reduce the error rate and the number of specifications to be denoted by [(F (x2, F (x1)]. If the evaluation function prefers the least amount of error regardless of the number of attributes, naturally the number of the selected features may still be high. On the other hand, if the objective function focuses on having a minimum number of attributes, the classification error may not be minimal. Therefore, in the problem of optimizing the choice of features, both objectives must be considered to achieve an optimal solution consisting of a set of optimal features. In the proposed method, the main trees are set with the features in the data set. Thus, each solution is a matrix whose values are equal to the number of features, and each element points to a feature in the data set.

Given that the proposed method was implemented in one dimension and the goal is to select or not to select features, so the initial matrix has a row column and d (to the number of main features). Therefore, the length of a vector is equal to the number of features of the relevant data set. Each vector in the squirrel optimization algorithm is a set of features in the data set that a number of components of this vector may randomly have 0 and 1. A value with value of zero specifies a feature that has not been selected, and a value with the value of one indicates selection of a feature associated with that data. In the proposed method, in order to select the features for the initial vectors, a random function with a constant threshold value is used. The performance of this function is in such a way that if the random probability is less than the value of the function threshold, a value of zero is assigned to the relevant value to this feature, otherwise the value 1 is recorded for this feature and the intended feature will be evaluated based on the objective functions. After selecting the initial population, the initial location and flight of each tree is determined based on the nature of the squirrel optimization algorithm using the evaluation functions. In this method, the location of each tree is considered as the selected feature of the data set features and the flight distance of each tree is considered as the convergence speed to high classification rate and reduction of classification error. Features that have the most value of the evaluation function are the output of the initial feature selection step. The best position and flight results are saved at this stage and the position of the other trees is updated accordingly. This process continues until the final answer is reached, which creates a balance between the objectives.

3.2 PROPOSED FITNESS FUNCTION

As mentioned, the proposed method uses the squirrel optimization algorithm to select a subset of fraud-related features in insurance companies. In this method, the objectives are combined and finally two types of general features in the form of minimization is achieved. The evaluation of clear subsets is done based on the two main objectives of reducing the number of features and the amount of classification error. In order to evaluate the initial population and select the expert community and find the particles with the highest weight, the fitness function is expressed as the following equation:

$$\mathbf{Minimize F(x)} = \mathbf{f_1(x)} = \frac{\mathbf{L}}{\mathbf{A}}\mathbf{L} \in \mathbf{A}.\mathbf{A} \in \mathbb{R}^+ + \mathbf{f_2(x)} = \mathbf{1} - \frac{\mathbf{FP+FN}}{\mathbf{P+N}} \ .\ (\mathbf{P}+\mathbf{N}) \in \mathbb{R}^+ \quad \mathbf{(1)}$$

In this relation, L is the number of features selected from the data set and A is the total number of features. In order to evaluate the error rate of each particle according to the selected features at each stage, the confusion matrix measure uses the true positive (TP) to indicate ordinary insured, which has been correctly modified by the model. True Positive (TP) is used to show the ordinary insurer that is properly identified by the model. False Positive (FP) is a fraudulent insurer while the system mistakenly displays them as ordinary insurers. True Negative (TN) is a sample classified as a fraudulent insurer in the proposed method and is in fact fraudulent. True Negative TN is a sample classified as a fraudulent insurer in the proposed method and is in fact ordinary fraudulent. In Equation 1, P is equal to the sum of TP + FN and N is equal to FP + TN. The first objective function f_(1) (x) is related to the ratio of selected features to the total features of the data set, while the second objective is for evaluating the degree of classification error.

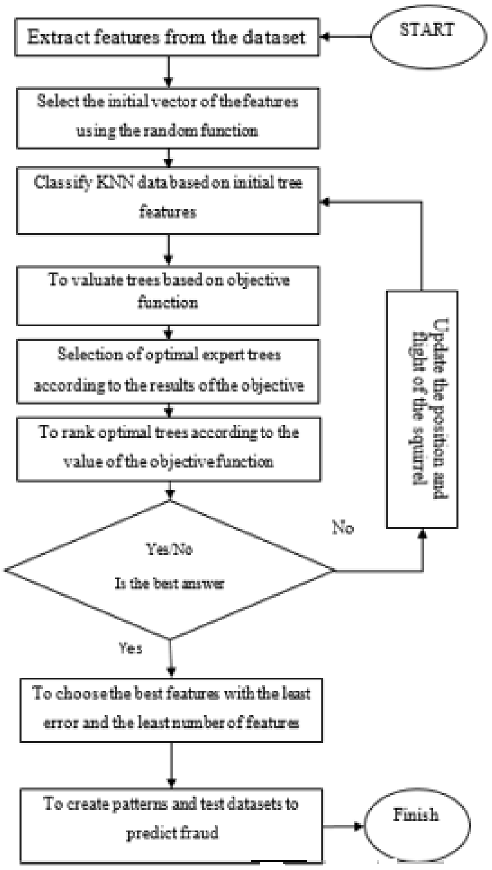
START
Extract features from the dataset
Select the initial vector of the features using the random function
Classify KNN data based on initial tree features
To valuate trees based on objective function
Selection of optimal expert trees according to the results of the objective
To rank optimal trees according to the value of the objective function
Update the position and flight of the squirrel
Yes/No
Is the best answer
No
Yes
To choose the best features with the least error and the least number of features
To create patterns and test datasets to predict fraud
Finish

Fig.1. Diagram

In the proposed method, the features selected at each stage are used to evaluate the KNN algorithm for classifying training samples to determine the degree of classification error for each particle and to obtain the degree of classification error. According to the number of selected features, the best particles in each stage are selected based on the degree of optimality and ranked in each stage. By converting these features obtained from each solution into feature vectors, KNN will try to differentiate between normal and fraudulent insured and create a safe margin between the two classes. Finally, the lowest error rate and the lowest number of features selected from the training data set are identified as the best solution and the features extracted as the classification pattern. Fig.1 shows the flow diagram of the proposed method.

4. TO IMPLEMENT THE PROPOSED METHOD

As mentioned, the proposed method was designed to provide a system for detecting fraud in social security based on a combination of feature selection based on the squirrel optimization algorithm and the nearest neighbor classification algorithm. In the proposed method, data on customer behavior in social security insurance were collected, which are classified into two groups: common insurers and fraudulent insurers. Therefore, in the continuation of this chapter of the research, first the proposed method preprocesses the data obtained from the database of the Social Security Insurance Organization using the squirrel optimization algorithm. Then, to implement the proposed method, the feature subset selected from the preprocessing step is classified using the nearest neighbor classification algorithm. Finally, we evaluate the proposed method using the class matching of the experimental sample set and the actual class of these samples and compare the proposed method with other classifications.

4.1DATA PREPROCESSING

As mentioned, the data obtained from the database of the Social Security Organization has different numerical, classified and nominal features. The use of all this data, in addition to increasing the complexity of the system, also reduces the accuracy of classification. Samples are labeled. Therefore, those features that have little effect on the classification of labeled samples should be removed. Therefore, we first select the data set features using the squirrel optimization algorithm. In the present data set, features that cannot have a significant effect on the classification of labeled samples and the determination of fraud detection patterns in the Social Security Organization are identified and eliminated during the heuristic steps in the squirrel optimization algorithm. Finally, the features that are effective in classifying fraudulent cases in the Social Security Insurance database remain as the output of the preprocessing phase.

Social Security Insurance information has a variety of values, all of which do not play an effective role in determining patterns of fraud detection in the Social Security Insurance database. Therefore, removing these features can play an important role in increasing the accuracy of fraud detection.

Features selection is one of the most challenging tasks in machine learning. If a set contains n number of features, a total of n^2 subsets is possible. In the optimal mode, the purpose of selecting features is to select the best subset of them. When n is inclined to a large number, it will be very difficult to select a feature subset because it is not possible to evaluate the performance of the model in each subset. Hence, various methods have been proposed to select the effective feature with the complexity of logical calculations. Extensive search, greedy search, random search, etc. are such techniques that have been used to find the best subset to use feature selection problems. But most methods suffer from early convergence, extreme complexity, and high computational cost. Therefore, meta-heuristic algorithms are considered very important for this type of situation. They are the most efficient and effective techniques and are able to find the best subset of features by maintaining the accuracy of the model.

Accordingly, in this study, the feature selection method has been considered based on the squirrel optimization algorithm in order to detect important features in the data set of the Social Security Insurance Organization. Squirrel optimization algorithm is one of the new meta-heuristic methods in which problem solutions are collected as branches on which squirrels can fly, and among the useful solutions, the almost optimal solution is selected as the best branch on which the squirrel can jump. Accordingly, in the proposed method, we will use this algorithm for the feature selection problem. As mentioned in the previous chapter, one of the most important issues in meta-heuristic optimization algorithms is the encoding of initial solutions. Given that the squirrel optimization algorithm is known as a meta-heuristic search algorithm, so this algorithm is based on a set of initial solutions. The squirrel optimization algorithm selects useful ideas and optimal solutions among the initial solutions, which are defined as existing ideas and problem solutions, and discards useless solutions. The basis of useful ideas is selected in the initial stage. Fig.2 suggests an example of the initial population in the squirrel optimization algorithmas shown in Figure 2, the initial population in the squirrel optimization algorithm is defined as a binary matrix. Each row of this matrix is considered as a solution or branch of the tree, and each element of each row is considered as a feature corresponding to the matrix column in the data set. Naturally, the number of machines is equal to the number of features in the data set and is equal to 16 features. Each of these values with a value of 1 indicates the selection of the attribute corresponding to the index of this column in the feature set. If the value of an object is 0, it means that the attribute of the column index of this element in the feature sets has not been selected in the proposed solution.

.

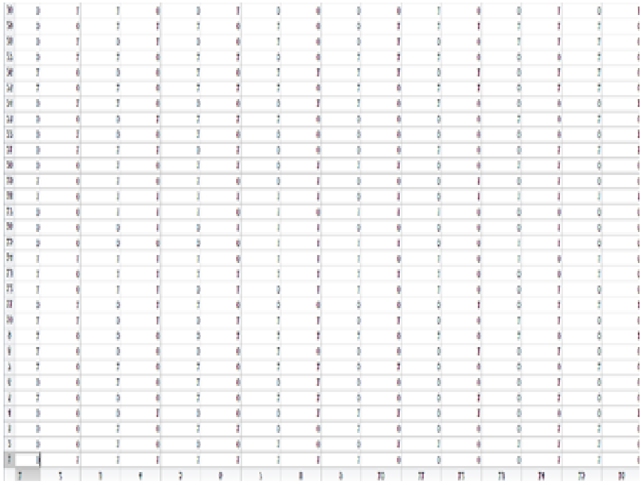

Fig.2. An example of the initial solutions in the squirrel optimization algorithm

The squirrel optimization algorithm starts by evaluating the initial solution as close branches and calculates the values of the solution fitness function at each stage. In the proposed method, the nearest neighbor algorithm has been used to calculate the fitness function of the solutions in the brainstorming algorithm. Therefore, according to the subset of features selected by each solution, the available data are classified according to the same features by the nearest neighbor algorithm and the error value of the solution is calculated. This error value for each solution is combined as a component of the function proportional to the ratio of the number of features of the solution to the total features of the data set. Finally, the value of the function is obtained proportional to the combination of the error value and the number of the solution features. The solution with the values of the minimum fitness function is remained and the rest of the solutions are eliminated. Then, the heuristic search stage begins according to the optimal solutions mentioned in the previous chapter to produce new solutions based on the optimal solutions extracted from the initial population as new branches. Therefore, in the next stage, the new solutions are evaluated as new branches and the same process continues until the stop conditions in the squirrel optimization algorithm are reached.

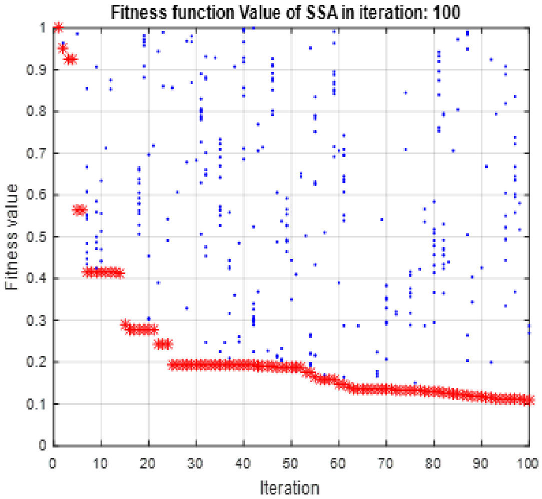

Fig.3. Fitness function values during the implementation of the squirrel optimizationalgorithm.

Given that in the proposed method, the gender of the fitness function is of the minimization type and the goal is to reduce the number of features and reduce the error in the data set of the Social Security Organization, so the value of the lower fitness function for each solution indicates the optimality of this method. Is it a solution and causes the selection of the solution as the optimal solution by the squirrel optimization algorithm. At the first stage, the squirrel optimization algorithm considers the value of the fitness function of all solutions. With the implementation of the algorithm in each stage, the solutions that have the least value of fitness function are considered as optimal solutions and heuristic search. A step to improve this solution begins. The figure 4shows the values of fitness performance of the optimal solutions when implementing the squirrel optimization algorithm.

As shown in Figure 3, the proper performance value to solve the solution starts from 1 in the squirrel optimization algorithm and gradually decreases in each step by finding optimal solution until it is repeated in 100 steps. The optimal solution is obtained with the value of the fitness function of 0, 1103. Accordingly, in the stages of the squirrel optimization algorithm, the optimal solutions are gradually developed to find the solution with the least amount of error and the least number of features. Table 1shows the remained features in the main dataset.

Table.1. Features selected by squirrel optimization algorithm

Feature name	Insurance type	Occupation type	Activity description	Activity code	Occupation title	Contract type code	Insurance premium rate	Number of children more than 4
Feature number	1	4	6	7	8	9	15	16

As shown in Table 1, out of 16 features in the main data set, only 8 features have been selected as effective features in detecting fraud in the Social Security Organization.The obtained feature subset has the lowest number of features and the lowest number of errors in each training data in the existing data set. The value of the optimal solution accuracy for the selected features using the nearest neighbor algorithm is 97.73 %.

4.2 EVALUATION OF THE PROPOSED METHOD

After implementing the proposed method based on feature selection based on squirrel optimization algorithm and nearest neighbor classification, new fraud samples can be predicted in the data set of the Social Security Insurance Organization as a proposed model. Using the feature selection method and the nearest neighbor algorithm, it is possible to examine patterns in which new samples in the Social Security Organization database are examined based on the selected features in the previous stage to identify the fraudulent or healthy person. Identify the insurer. For this purpose, to determine the accuracy of the proposed method for predicting new fraud samples based on the proposed method and features selected by the squirrel optimization algorithm, the evaluation takes place using the matching of the class of new fraud sample sets and the actual class of these samples. Therefore, to evaluate the proposed method, compare the predicted class for the new fraud samples and use it with the actual class of the turbulence matrix samples with the four parameters including positive true (TP), false positive (FP), true negative (TN), and false negative (FN) as follows:

TP: A sample that is classified as ordinary insurer in the proposed method and in fact he/she is an ordinary insurer.

FP: A sample that is classified as ordinary insurer in the proposed method but is in fact a fraudulent insurer.

TN: A sample that is classified as fraudulent insurer in the proposed method and in fact he is a fraudulent insurer.

FN: A sample that is classified as fraudulent insurer in the proposed method but is in fact a ordinary insurer.

Based on the fact that the turbulence matrix is a standard method for measuring classification performance in two-class data, the evaluation measure derived from the confusion matrix include the following cases:

Since the confusion matrix is a standard method for measuring classification performance in two-class data, the evaluation measure obtained from the confusion matrix include accuracy, recall, precision, and the F-measure defined in the following equations:

$$\text{Accurace}=(TP + TN)/(TP+TN+FP+FN) \qquad (2)$$

$$\text{Recall} =TP/(TP+FN) \qquad (3)$$

$$\text{Precision}=TP/(TP+FP) \qquad (4)$$

$$\text{Fmeasure}=(2*\text{Precision}*\text{recall})/(\text{Precision}+\text{Recall}) \qquad (5)$$

Evaluation measure derived from confusion matrix variables are used as a tool to measure the quality of the proposed method and compare it with other methods. Therefore, in this dissertation, we first compare the combination of feature selection method based on squirrel optimization algorithm and nearest neighbor algorithm with nearest neighbor algorithm, neural network algorithm and Bayes algorithm without using feature selection method. Figure 4 shows a comparison of the proposed method with the classification algorithm in terms of accuracy of fraud detection in new fraud samples.

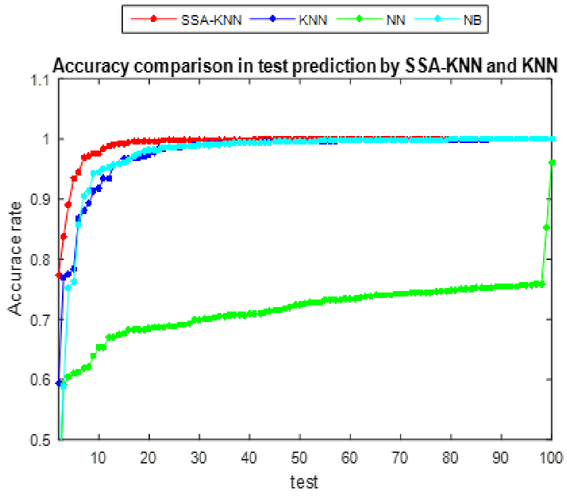

Fig.4. Accuracy comparison of the proposed method with classification algorithms

As shown in Figure 4, the accuracy diagram for the proposed method, which is based on a combination of feature selection based on the squirrel optimization algorithm and the nearest neighbor algorithm, has been drawn compared to other classifiers without using the feature selection method to predict fraud samples in the data set of the Social Security Insurance Organization. According to Figure 4, it can be seen that in general, the proposed method has an earlier tendency towards the optimal point of accuracy measure. Therefore, it can be said that the proposed method is more accurate than the classification algorithms without using the feature selection method. Figure 5 shows the comparison of the proposed method with the classification algorithm in terms of fraud detection recall in new fraud samples.

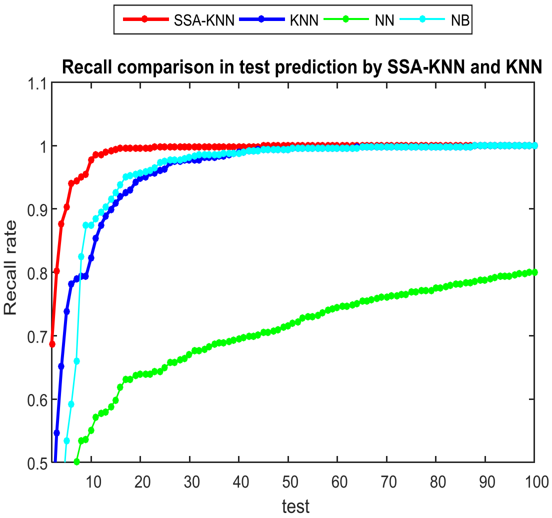

Fig.5. Recall comparison of the proposed method with classification algorithms

As shown in Fig.6, the recall measure diagram for the proposed method, which is based on a combination of feature selection based on the squirrel optimization algorithm and the nearest neighbor algorithm, has been drawn compared to other classifiers without using the feature selection method to predict fraud samples in the data set of the Social Security Insurance Organization. According to Fig.5, it can be seen that in general, the proposed method has an earlier tendency towards the optimal point of recall measure. Therefore, it can be said that the recall measure in the proposed method is more than the classification algorithms without using the feature selection a method. Figure 6 shows the comparison of the proposed method with the classification algorithm in terms of fraud detection precision in new fraud samples.

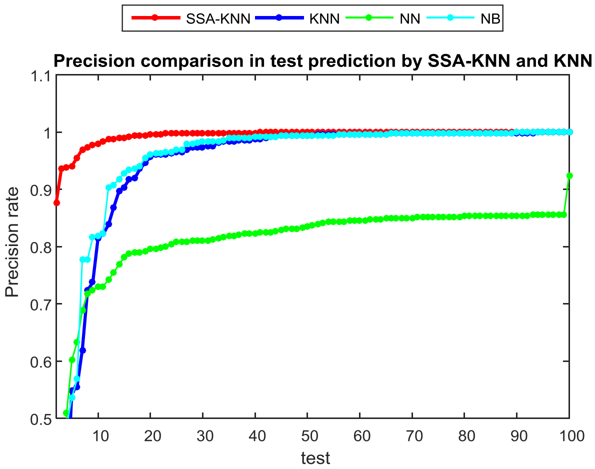

Fig.6. Precision comparison of the proposed method with classification algorithms

As shown in Figure 6, the precision measure diagram for feature selection methods based on the squirrel optimization algorithm and the nearest neighbor algorithm and other classification algorithms has been obtained without using the feature selection method to predict new fraud samples. According to Figure 6, it can be seen that the proposed method, relying on feature selection based on squirrel optimization algorithm and combining it with the nearest neighbor classifier to predict new fraud samples, has a higher precision measure than classification algorithms without using the feature selection method. Also, Figure 7shows the comparison of the proposed method with classification algorithms in terms of F-measure for detecting fraud in new fraud samples.

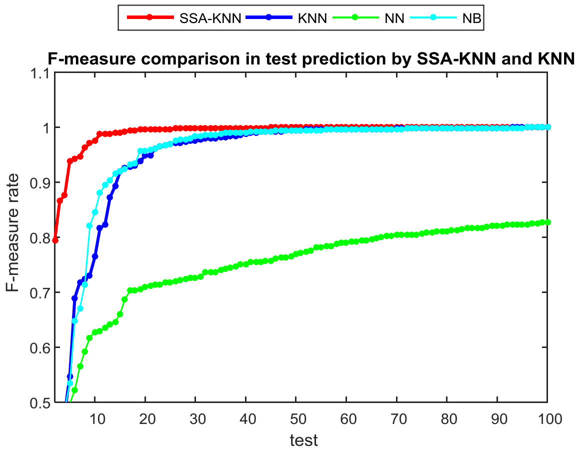

Fig.7. F-measure comparison of the proposed method with classification algorithms

As can be seen in Figure 7, F-measure obtained for the proposed methods and other classification methods. As can be seen in this figure, the proposed method has a higher F-measure rate than other classifications without using the feature selection method.

4.3COMPARISON OF THE PROPOSED METHOD WITH DATA MINING CLASSIFICATION ALGORITHMS

After implementing and evaluating the proposed method based on the evaluation measure extracted from the confusion matrix on the test data in the data set of the Social Security Insurance Organization, it is time to compare the proposed method with other data mining classification algorithms in terms of evaluation measures in order to validate and determine the performance improvement of the proposed method in predicting new fraud samples compared to other previous methods. Given that the data used in this study was obtained from the Social Security Organization, the proposed method used this data to develop the model and this data has not been used in other publications in the field of fraud detection in insurance, for this purpose, to compare the proposed method with other methods, other algorithms were implemented on this data set and the values of evaluation measures were extracted to be compared with the proposed method.

Table 2 shows the comparison of the proposed method with other classification algorithms in terms of evaluation measure for new fraud samples.

	Accuracy	Recall	Precision	F-measure
SSA_KNN	98.8	98.54	98.97	98.72
KNN	96.64	94.71	93.6	93.54
NN	70.7	68.98	79.98	73.75
NB	96.41	93.96	93.82	93.77

Table 2. Comparison of the proposed method with classification algorithms in terms of evaluation measure

As can be seen in Table 2, the proposed method, based on the feature selection method based on the squirrel optimization algorithm, not only found the best agent features among all the features available in the Social Security Organization database, but also decreased the classification error of training samples and predicting new fraud samples. It shows significant improvement over other data mining classification algorithms.

5.COMPARISON OF THE PROPOSED METHOD WITH DATA MINING CLASSIFICATION ALGORITHMS

In this study, a combination of feature subset selection based on squirrel optimization algorithm and nearest neighbor classification has been used to detect health insurance fraud. To evaluate the model, the proposed method used the educational data set obtained from Ardabil' Social Security Organization for determining the patterns of fraud detection in health insurance and the test data set prepared from this data set. It can be concluded that the presence of additional and irrelevant features of the class label among the data sets in machine learning and classification models in data mining reduces the negative effects and classification precision of educational records, especially test records, recently added to the model. In other words, it can be said that the existence of irrelevant features changes the focus of model on achieving accurate patterns not to distinguish ordinary insurers from fraudulent insurers and reduce the optimal performance of the proposed model. According to the results of the implementation of proposed method, it is observed that the proposed method shows good results for classifying and predicting fraud in the database of the Social Security Organization by removing irrelevant features and finding useful features, and it has a higher accuracy compared to other data mining classification algorithms in this area.

REFERENCES

1. Stojanović B, Božić J, Hofer-Schmitz K, Nahrgang K, Weber A, Badii A, Sundaram M, Jordan E, Runevic J. Follow the trail: machine learning for fraud detection in Fintech applications. Sensors. 2021 Jan;21(5):1594.
2. Roszkowska, P. (2020). Fintech in financial reporting and audit for fraud prevention and safeguarding equity investments. Journal of Accounting & Organizational Change.
3. Mohajer, A., Bavaghar, M., Saboor, R., & Payandeh, A. (2013, August). Secure dominating set-based routing protocol in MANET: Using reputation. In 2013 10th International ISC Conference on Information Security and Cryptology (ISCISC) (pp. 1-7). IEEE.
4. Mohajer, A., Mazoochi, M., Niasar, F. A., Ghadikolayi, A. A., & Nabipour, M. (2013, June). Network Coding-Based QoS and Security for Dynamic Interference-Limited Networks. In International Conference on Computer Networks (pp. 277-289). Springer, Berlin, Heidelberg.
5. Lavanya S, Kumar SM, Kumar PM. Machine Learning Based Approaches for Healthcare Fraud Detection: A Comparative Analysis. Annals of the Romanian Society for Cell Biology. 2021 Apr 2:8644-54.
6. Mohajer, A., Bavaghar, M., & Farrokhi, H. (2020). Reliability and mobility load balancing in next generation self-organized networks: Using stochastic learning automata. Wireless Personal Communications, 114(3), 2389-2415.

7. Mohajer, A., Barari, M., & Zarrabi, H. (2016). Big Data-based Self Optimization Networking in Multi Carrier Mobile Networks. Bulletin de la Société Royale des Sciences de Liège, 85, 392-408.
8. Dami S, Barforoush AA, Shirazi H. News events prediction using Markov logic networks. Journal of Information Science. 2018 Feb;44(1):91-109.
9. Dami S. News Events Prediction Based on Casual Inference in First-Order Logic (FOL). Journal of Soft Computing and Information Technology. 2016 Dec 21;5(4):11-25.
10. Javid, S., & Mirzaei, A. (2021). Presenting a Reliable Routing Approach in IoT Healthcare Using the Multiobjective-Based Multiagent Approach. Wireless Communications and Mobile Computing, 2021.
11. Fiederling K, Schiller J, von Bieberstein F. Can we trust consumers' survey answers when dealing with insurance fraud?. Schmalenbach Business Review. 2018 May;70(2):111-47.
12. Haque ME, Tozal ME. Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts. IEEE Transactions on Services Computing. 2021 Jan 12.
13. Lacruz, F., & Saniie, J. (2021, May). Applications of Machine Learning in Fintech Credit Card Fraud Detection. In 2021 IEEE International Conference on Electro Information Technology (EIT) (pp. 1-6). IEEE.
14. Dong, Manqing, Lina Yao, Xianzhi Wang, Boualem Benatallah, Chaoran Huang, and Xiaodong Ning. "Opinion fraud detection via neural autoencoder decision forest." Pattern Recognition Letters 132 (2020): 21-29.
15. 7. Kunickaitė R, Zdanavičiūtė M, Krilavičius T. Fraud detection in health insurance using ensemble learning methods. InCEUR Workshop proceedings [electronic resource]: IVUS 2020, Information society and university studies, Kaunas, Lithuania, 23 April, 2020: proceedings. Aachen: CEUR-WS, 2020, Vol. 2698 2020.
16. Dami S, Yahaghizadeh M. Predicting cardiovascular events with deep learning approach in the context of the internet of things. Neural Computing and Applications. 2021 Jan 3:1-8.
17. Dami S, Esterabi M. Predicting stock returns of Tehran exchange using LSTM neural network and feature engineering technique. Multimedia Tools and Applications. 2021 May;80(13):19947-70.
18. Rezaei A, Dami S, Daneshjoo P. Multi-document extractive text summarization via deep learning approach. In2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI) 2019 (pp. 680-685). IEEE.
19. Mirzaei A, Souha AN. Towards Optimal Configuration in MEC Neural Networks: Deep Learning-Based Optimal Resource Allocation. Wireless Personal Communications. 2021 Jul 5:1-23.
20. Dami S, Shirazi H, Hoseini SM. A Data Mining Model for Anomaly Detection of Satellite Launch Vehicle. Adst Journal. 2013 Jan 1.
21. Dami S, Yahaghizadeh M. Efficient event prediction in an IOT environment based on LDA model and support vector machine. In2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) 2018 Feb (pp. 135-138). IEEE.
22. Emami H, Dami S, Shirazi H. K-Harmonic Means Data Clustering With Imperialist Competitive Algorithm. University Politehnica of Bucharest-Scientific Bulletin, Series C: Electrical Engineering and Computer Science. 2015 Feb;77(7).
23. Farhang, M., Mohajer, A., Zobeyravi, O., & Rahimzadegan, A. Adaptive Spectrum Sensing Algorithm Based on Noise Variance Estimation for Cognitive Radio Applications.
24. Ramandi S, Niakan L, Rajaee Harandi S, Asheghi H. Fraud detection in supplementary health insurance and ways to compete. Iran Health Insurance Organization. 2020 Oct 10;3(3):178-87.
25. Rahimi AM, Ziaeddini A, Gonglee S. A novel approach to efficient resource allocation in load-balanced cellular networks using hierarchical DRL. Journal of Ambient Intelligence and Humanized Computing. 2021 Apr 13:1-5.
26. Somarin AM, Barari M, Zarrabi H. Big data based self-optimization networking in next generation mobile networks. Wireless Personal Communications. 2018 Aug;101(3):1499-518.
27. Zhang S, Madadkhani M, Shafieezadeh M, Mirzaei A. A novel approach to optimize power consumption in orchard WSN: Efficient opportunistic routing. Wireless Personal Communications. 2019 Oct;108(3):1611-34.
28. Mirzaei A, Barari M, Zarrabi H. Efficient resource management for non-orthogonal multiple access: A novel approach towards green hetnets. Intelligent Data Analysis. 2019 Jan 1;23(2):425-47.
29. Mirzaei A, Zandiyan S, Ziaeddini A. Cooperative Virtual Connectivity Control in Uplink Small Cell Network: Towards Optimal Resource Allocation. Wireless Personal Communications. 2021 Apr 23:1-25.
30. Mirzaei A, Barari M, Zarrabi H. An Optimal Load Balanced Resource Allocation Scheme for Heterogeneous Wireless Networks based on Big

Data Technology. arXiv preprint arXiv:2101.02666. 2021 Jan 7.

31. Mohajer, A., Yousefvand, M., Ghalenoo, E. N., Mirzaei, P., & Zamani, A. (2014). Novel approach to sub-graph selection over coded wireless networks with QoS constraints. IETE Journal of Research, 60(3), 203-210.
32. Barari M, Zarrabi H, Somarin AM. A New Scheme for Resource Allocation in Heterogeneous Wireless Networks based on Big Data. Bulletin de la Société Royale des Sciences de Liège. 2016 Jan 1;85:340-7.
33. Mohajer, A., Barari, M., & Zarrabi, H. An Enhanced multi-dimensional Adaptive Handover Algorithm for Mobile Networks.
34. Mohajer, A., Barari, M., & Zarrabi, H. An Efficient Resource Allocation Mechanism including Load-aware Handover Decision.
35. Ilango V. A time-efficient model for detecting fraudulent health insurance claims using Artificial neural networks. In2020 International Conference on System, Computation, Automation and Networking (ICSCAN) 2020 Jul 3 (pp. 1-6). IEEE.
36. Bavaghar, M., Mohajer, A., & Taghavi Motlagh, S. (2020). Energy Efficient Clustering Algorithm for Wireless Sensor Networks. Journal of Information Systems and Telecommunication (JIST), 4(28), 238.
37. Elshaar, S., & Sadaoui, S. (2020). Semi-supervised classification of fraud data in commercial auctions. Applied Artificial Intelligence, 34(1), 47-63.
38. Dzakiyullah, N. R., Pramuntadi, A., & Fauziyyah, A. K. (2021). Semi-Supervised Classification on Credit Card Fraud Detection using AutoEncoders. Journal of Applied Data Sciences, 2(1), 01-07.
39. Melo-Acosta, G. E., Duitama-Muñoz, F., & Arias-Londoño, J. D. (2017, August). Fraud detection in big data using supervised and semi-supervised learning techniques. In 2017 IEEE Colombian conference on communications and computing (COLCOM) (pp. 1-6). IEEE.
40. Pouramirarsalani, A., Khalilian, M., & Nikravanshalmani, A. (2017). Fraud detection in E-banking by using the hybrid feature selection and evolutionary algorithms. International Journal of Computer Science and Network Security, 17(8), 271-279.
41. Zhang, X., Han, Y., Xu, W., & Wang, Q. (2019). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. Information Sciences.