Journal of Theoretical and Applied Information Technology
31st August 2021. Vol.99. No 16

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENSEMBLE SELECTION AND COMBINATION BASED ON COST FUNCTION FOR UCI DATASETS

***[1]MOHD KHALID AWANG, [1]MOKHAIRI MAKHTAR, [1]ABD RASID MAMAT**

[1]Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin,

22000 Tembila, Terengganu, Malaysia

*Corresponding author: khalid@unisza.edu.my

ABSTRACT

It is well-known that the classification performance of any single classifier is outperformed by a multiple classifier approach or an ensemble process that incorporates results from different base classifiers. However, even though they have the potential to achieve greater classification precision, their vast number of base classifiers has greatly influenced ensemble methods. In the ensemble process, the selection and combination of appropriate and varied classifiers is a daunting task. In the previous work, we, therefore, suggested a new soft ensemble selection and combination approach (SSSC) to identify the best subset of heterogeneous ensemble team of classifiers and demonstrated the potential of our proposed algorithm to minimise a large number of classifiers while at the same time generating the highest predictive precision for consumer churn data sets. This paper extended the earlier work with the goal of evaluating whether the proposed SSSC model works well with the other UCI repository benchmark data sets. The findings of the experiments demonstrated that the developed model resulted in the improvement of the chosen UCI data sets' prediction accuracy. Based on the results of the experiments, it indicates that the prediction accuracy of the proposed SSSC outperformed other single classifiers and ensemble methods for Liver Disorder, Hepatitis and Breast Cancer data sets. This work has shown that the proposed SSSC is able to search for a minimal number of classifiers in the repository of the ensemble while at the same time enhancing the precision of the classification of the chosen UCI data sets.

Keywords: *Ensemble Selection, Customer Churn Prediction, Ensemble Combination, Soft Set, Ensemble Methods.*

1. INTRODUCTION

Ensemble methods, also known as many classifiers, are a kind of machine learning algorithm technique that involves training a number of base classifiers and combining their output to obtain the most excellent prediction accuracy possible [1]. Combining the predictions of a number of classifiers, such as bagging [2], boosting [3], stacking [4], Bayes optimum classifier [5], rotating forest [6], ensemble selection [7], and hybrid intelligent system [8], maybe a useful approach for improving classification performance.

As a rule, ensemble techniques are divided into two phases: the creation of numerous base classifier models and the combining of those models [7]. When analysing the nature of ensemble methods, which produce a large number of individual classifiers, there seem to be certain drawbacks. One of the most significant risks to the ensemble system is the deployment of a high number of base classifiers. The integration of the whole collection of base classifiers has the effect of decreasing the usefulness of the final decision reached by the ensemble technique. Because of this, selecting and combining the most appropriate set of classifiers is considered to be one of the most difficult challenges in the ensemble learning process.

To address this issue, our previous work provided a novel approach for selecting and combining ensemble classifiers from the pruned ensemble based on soft set theory. The goal of the design method is to use soft set theory to address the issue of selecting less redundant ensemble classifiers and identifying the optimal classifier subset. In earlier work, we used a new soft set-based method to choose and combine classifiers

from the pruned ensemble committee in predicting customer churn in telecoms companies. The suggested model has been shown to be capable of searching for the optimum subset of classifiers while also providing the greatest prediction accuracy for customer churn data sets [9].

This article used three groups of secondary data sets from the UCI benchmark repository to test and demonstrate that our proposed method is acceptable and works well on additional data sets. The suggested SSSC techniques are validated and verified using the chosen data sets. As a result, this article has selected three widely used UCI data sets that have been published and utilised in previous machine learning studies. In earlier studies, these data sets were often used as benchmarks to evaluate the performance of various categorisation methods.

The following is how the remainder of this article is organised. The ensemble's techniques, as well as the ensemble's selection and combining, are all discussed in Section 2. Section 3 discusses the data sets and the suggested ensemble selection and combining technique, as well as the methodological approach. The arguments and findings of research on the proposed approach are described in Section 4. In Section 5, the results of this study are summarised.

2. LITERATURE REVIEWS

2.1 Ensemble Methods

The ensemble approach is founded on the concept of combining ideas from various individuals or basic classifiers while simultaneously attempting to improve outcomes to complement each other [10], [11], [12]. The majority of prior research agree that employing an ensemble approach instead of a single classifier with a precondition increases accuracy and that the base classifiers in the combinations must be accurate and varied [13]. This ensemble technique's idea is comparable to the decision-making approach, in which individuals are urged to create a group conversation with coworkers before making any decisions. Before making any major choices, people often seek out second views. Individual views that may be somewhat different from one another will generally be examined before a choice is made, and then their ideas will be combined to achieve a final conclusion [14], [15]. We prefer to select the recommendations that usually contribute the most toward the ultimate choice based on the premise that not all views are worth considering. Similarly, the final ensemble combination in the ensemble method approach should only include the most relevant and diverse base classifiers. Ensemble selection and the combination is the name of the procedure, which is described in the next section.

2.2 Phases in Ensemble Methods

The ensemble techniques may essentially be split into two major phases: building and combining. [14] suggested including at least two major stages in all ensemble techniques. The ensemble should begin with the development of ensemble classifications. The second step, known as the ensemble integration or combination, is linked to the combination of the predictions of each classification in an ensemble. But some researchers suggest ensemble techniques consisting of three stages [16].The three stages are ensemble building, ensemble cutting and ensemble combination [17].

- Ensemble construction stages produce a set of heterogeneous base learner classifiers that are used to predict the final output using a specified learning method.
- The ensemble pruning phase eliminates certain basic classifiers based on different mathematical methods to enhance the overall accuracy of the ensemble.
- Ensemble selection and combination phase. The filtered learner models are merged to create a single or subset of classifiers during the ensemble selection and combination step, which may yield results that are more accurate than the average of all the individuals' base classifiers.

2.3 Ensemble Construction

An ensemble may be made up of both homogeneous and heterogeneous models. Both of these broad groups, which are homogenous and heterogeneous, respectively, will be explored in more detail in the next section.

2.3.1 Homogenous Ensemble

Homogeneous means that the same learning method is used. In the same learning process, various variables are used to generate

distinct homogeneous models, which are derived from separate executions [2] and boosting are two common techniques for generating homogenous models [18]. Following are some options for creating a homogeneous ensemble:

- Manipulation of the learning algorithm's parameters;
- Injection of randomness into the learning process; or
- Manipulation of the training cases; or
- Manipulation of the input characteristics and classifier outputs.

Bagging

The word "bagging" implies "Bootstrap" [19],[7]. Bootstrap and aggregate are the two fundamental principles of bagging. The combination of independent base classifiers usually leads to a significant reduction in error and the basis classifiers must thus be as autonomous as feasible. Bagging trains every classifier in the ensemble by randomly utilising a portion of training data sets to encourage variation and variety of classifications. The set of data utilised must not overlap data. For example, the random forest method mixes random decision-making trees with this technique to obtain extremely high classification accuracy.

Boosting

The power of the boosting algorithm lies in its capacity to transform poor classifications into strong classifications. The strong classifier is intuitively near to the ideal performance, while the weak classifier is somewhat better than random forecasts. The genesis of this method is based on a fundamental question: may weak and powerful classifiers be combined to produce a flawless result? This notion is extremely significant since the number of weak classifiers typically surpasses the high standard. The boost says that any poor classification may be moved to a strong classification. It is generally simple to acquire a poor student, but it is difficult to get a strong student [18].

2.3.2 Heterogeneous Ensemble

When the classifier relies on several learning techniques on the same data set, the heterogeneous ensemble model is formed [7], [10]. The classifier has diverse views and results of the prediction because of different techniques of learning. This method is one way in which several ensembles are produced, ensuring high results for the ensemble merger. These algorithms each have their own advantages and inconveniences. In comparison with the closest k-neighbour approach, neural networks are powerful for noise for example. The mixture of several classifications can provide greater performance in the categorisation.

More recently, study [20], [21] explored the inclusion of an additional intermediate level that deals with the minimisation of the ensemble size prior to the merging of the ensembles. This step is known as selection of the ensemble. Others call it a subset, a selective ensemble or a thinning ensemble [22]

2.4 Ensemble Selection and Pruning

Various ensemble techniques have been suggested as learning algorithms in data mining to enhance the performance and accuracy of classifiers by previous researchers. There are no dominant ensemble techniques in classification [21]. Most prior research have focused on ensemble creation and ensemble combination in increasing classification accuracy and performance, but have not paid enough attention to ensemble pruning methods. Nonetheless, there is a scarcity of research focused on ensemble pruning techniques [22], [23]

[24] used statistical tests to narrow a collection of heterogeneous ensembles in order to evaluate if the variations in prediction performance between the classifiers in the group are substantial. Only the classifiers with significantly better performance are kept and merged using the weighted sum voting technique. The obvious limitation of these approaches is that they do not account for the variety of classifiers as a whole. However, it is an efficient way to clean out low-performance models in a big group.

[23] suggested an ensemble pruning approach (DivP) for combining various pairwise diversity matrices using a genetic algorithm. The suggested model employs graph algorithms to aid in the development of comparable group classification methods. The DivP approach was evaluated on 21 UCI data sets, and its findings

were compared to five state-of-the-art ensemble pruning methods; it outperformed AGOB, DREP, and GASEN. Additionally, the authors asserted that DivP produces a smaller final ensemble than state-of-the-art techniques.

[25] developed a novel ensemble pruning technique for increasing the efficacy and efficiency of an existing ensemble. The authors define the ensemble pruning issue as a rigorous mathematical programming problem and then reduce it using semi-definite programming (SDP) methods to get an effective approximation solution. Initially, the authors defined the ensemble pruning issue as a quadratic integer programming problem involving the search for a fixed-size subset of classifiers with the greatest divergence and the least misclassification. According to the authors, when evaluated on data from the UCI repository, the SDP-based pruning method outperforms two existing metric-based pruning algorithms. However, one disadvantage is that the method needs a parameter specifying the number of classifiers to keep and runs in polynomial time.

[26] proposed a neural network pruning technique. The researchers have developed the GASEN (Genetic Algorithm-based Selective ENsemble) approach, which asserted that integrating many accessible neural networks may be preferable to assembling them all together. They utilised neural networks as classifiers and 20 distinct datasets in their practical research. Their experimental results showed that the pruned ensemble produced by the GASEN technique outperformed commonly used ensemble methods such as Bagging and Boosting in both regression and classification. The authors emphasised GASEN's advantage since it utilises much fewer components of neural networks but achieve a higher degree of generalisation capacity. The obvious disadvantage is that it is restricted to homogenous ensembles.

[22] developed an ensemble selection technique for constructing the final ensemble from hundreds of base classifier libraries. Different learning methods and parameter settings are used to build base classifiers in library pools. Rather than combining good and poor models in an ensemble, the authors utilised an advanced stepwise selection method to identify a subset of basic classifiers that, when averaged together, provide outstanding performance from a library of models. The findings indicate that ensemble selection regularly finds ensembles that outperform all other models, including models trained using bagging, boosting, and Bayesian models, according to the authors. One of the work's drawbacks is that it solely considers forecast accuracy and ignores other performance indicators.

[27] suggested dynamic pruning of sets as a multiple label classification problem, using set members as labels. By using cross-validation, the multiple label training examples are produced by determining if the members of the set are correct or not in the original training set. The authors stated that using learning algorithms that optimise precision depending on the example in the multi-label classification job improves classification accuracy. The suggested framework was compared to cutting-edge trimming methods. The experiment employed 200 different classifiers, and the results indicated a substantial increase in classification accuracy.

2.3 Ensemble Combination

Ensemble methods normally started with the training of base classifiers as many as possible using numerous learning algorithms and control parameters to construct the pool of ensemble [28]. At the ensemble construction phase, all models or base classifiers are added to the classifier library without considering their performance. Therefore, the classifier's library may consist of strong and weak classifiers. During the ensemble construction, little or no attempt is made to optimise the performance of the individual classifiers.

The ensemble technique, which is based on the idea of combination, is used to attain high generalisation abilities. Following the construction of basic classifiers, the ensemble approach does not seek to identify the greatest single classifier but rather seeks to identify a collection of classifiers that can make the best choices. The outputs of the base-classifiers may be combined in a variety of ways, the most common of which being weighting methods and meta-learning techniques. In situations when the basic classifiers execute the same task and achieve similar performance, weighting techniques are helpful. If some classifiers consistently accurately classify or consistently misclassify a set of examples, meta-

learning techniques are the most appropriate way to use.

2.5 Soft Set Reduction Algorithm as Ensemble Selection Strategy

A new mathematical instrument for dealing with uncertain data, the soft set theory presented by [29]–[31], is a new mathematical tool for dealing with uncertain data [32]. On the basis of soft set theory, the concept of attribute reduction and decision making was proposed by [33]. This method is similar to the [34] rough reduction in that it selects the decision in a given attribute based on the maximal weighted choice value. [31], a novel strategy for attribute reduction in a multi-valued information system based on a soft set was introduced for the first time. When combining and selecting the most relevant classifiers from an ensemble of classifiers, a soft set attribute reduction approach is proposed in this study.

3. METHODOLOGY

3.1 Data Sets

As part of the validation process for the suggested methods, this article has chosen a secondary data set from the UCI Benchmark collection. The data sets are utilised in the validation and verification process of the techniques presented in this paper, which is described in detail below. This study has selected three widely used UCI data sets that have been previously published and applied in other studies as the basis for its investigation. In earlier research, these data sets were often used as benchmarks to evaluate the performance of various classification methods, and they continue to be utilised as such. Table 1 contains a summary of the three UCI benchmark data sets, which is divided into three categories.

Table- 1: The Three UCI Benchmark Data Sets

Data Set	No. of Instances	No. of Features	No. of Classes	Class Distribution
Liver Disorder	345	7	2	145:200
Hepatitis	155	20	2	32:123
Breast Cancer	699	11	2	458:241

3.2 Proposed Soft Ensemble Selection and Combination Method

This study aims to identify the best subset of classifiers from a large number of classifiers. The technique for generating the collection of different classifiers and the ensemble pruning and ensemble combination stages are described in this section. The following steps make up the overall strategy of the suggested ensemble technique in this study:

i. Ensemble Construction.

The study began with the creation of a classifier pool. The first step in creating a successful ensemble technique is to create a varied set of base classifiers in the repository. The pool of classifiers in this study is made up of heterogeneous classifiers created using ten different classification learning methods.

ii. Feature Selection.

This paper proposes a set of feature selection methods that are intended to guarantee that the overall performance of each classifier is as high as feasible.

iii. Ensemble Selection and Pruning.

The soft set algorithm has been suggested as a method of selecting the most relevant and varied base classifiers from a pool of available base classifiers. The soft set pruning method that was used was comparable to other pruning techniques that had been used before. The suggested pruning method is based on a soft set attribute reduction algorithm.

iv. Ensemble Combination.

According to this study, an optimal subset of classifiers is defined as the smallest number of classifiers in an ensemble that produces the highest accuracy in prediction. In this approach, we start with a selection of the best classifiers from the initial pruned ensemble and work our way down to obtaining the optimal subset of ensemble. The most accurate classifier is the one that produces the most accurate predictions. The soft set reduction method should next be applied to the remaining classifiers that were pruned. In order to increase prediction accuracy, the optimised subset of ensemble is merged with the best classifiers currently available,

which are chosen from the new soft set reduction method, if the combination does so. The ensemble combination method is implemented in this article using the simple majority technique. Contrary to popular belief, comparative tests conducted by [35] demonstrated that, while being a simple technique, majority votes are often superior or at least comparable to a more complex classifier combination method. The majority voting method, with its simplicity and high performance, seems to be an intriguing combination that can be used to integrate any kinds of classifiers, regardless of their output, in a straightforward manner. Majority votes have been chosen as the ensemble combination technique for the classifiers in this research as a result of the characteristics described.

It is the goal of this research to identify the optimal ensemble classifiers based on a combination of the value of prediction accuracy (ACC), the sensitivity of the prediction (TPR), and specificity of the prediction (TNR). The ensemble teams that are selected will have a good balance between TPR and TNR while also maintaining the greatest prediction accuracy (ACC). This proposed method is known as Soft Set Ensemble Selection and Combination (SSSC).

3.2 Ensemble Selection and Combination (SSSC) based on Cost Function

A critical component of ensemble techniques is the selection and combining of base classifiers from a pool of ensembles. To construct the classifiers, researchers suggested a variety of methods based on a predefined rule. The selection of the optimum classifiers for ensemble techniques becomes a critical element affecting the performance of the ensemble classifiers. Additionally, ensemble combinations based on a single performance metric, such as accuracy (ACC), tend to generate a subset of biased classifiers. Thus, in order to achieve a balanced and unbiased subset, this article proposes a novel technique for choosing and merging the most suitable classifiers. This section suggested a cost function performance metric that included the value of prediction accuracy (ACC), the prediction's sensitivity (TPR), and the prediction's specificity (SPR) (TNR). The suggested cost function is based on weighted sum methods, which are the traditional method for solving a multi-objective optimisation issue to determine the best classifiers from all base classifiers.

When it comes to predicting performance, accuracy alone isn't always enough. Furthermore, an ensemble technique such as ACC that concentrates on a single performance metric may provide bias classifiers. As a result, additional performance metrics such as TPR and TNR must be combined in order to get a superior generality and performance measure. Users may wish to mix the weights of ACC, TPR, and TNR. To account for all of these variables, the researchers devised a cost function measurement that combines the value of prediction accuracy (ACC), prediction sensitivity (TPR), and prediction specificity (SPR) (TNR). The suggested cost function is based on weighted sum methods, which are the most used method for solving multi-objective optimisation problems. Weights, indicated as w_{c1}, w_{c2}, w_{c3}, are given to each of the cost functions (ACC, TPR, and TNR) in this method, converting the issue to a single objective problem. A linear combination of the aforementioned cost function, which is represented by the following equation, may then be used to get the new ensemble performance.

$$COF_i = (w_{c1} * Acc_i) + (w_{c2} * TPR_i) + (w_{c3} * TNR_i)$$

Where $1 \geq w_{ci} \leq 0 \ \forall_i$ and $\sum_{W_{c3}}^{W_{c1}} = 1$

The primary goal of this cost function is to achieve the best possible combination of these three performance indicators. As a result, the ensemble's chosen teams will have a good balance between TPR and TNR while also maintaining the highest ACC possible. As a result, the experiments are split into three subcategories, which are as follows:

i. Apply cost function with a single performance measure.

$SSSC_1$ focused on ACC, while $SSSC_2$ focused on TPR and $SSSC_3$ focused on TNR. In this experiment, the cost functions are represented by the following equation:

$$SSSC1_i = (1.0 * Acc_i) + (0.0 * TPR_i) + (0.0 * TNR_i)$$

$$SSSC2_i = (0.0 * Acc_i) + (1.0 * TPR_i) + (0.0 * TNR_i)$$

$$SSSC3_i = (0.0 * Acc_i) + (0.0 * TPR_i) + (1.0 * TNR_i)$$

ii. Apply cost function by combining two performance measures.

$SSSC_4$ focused on (ACC and TNR) while $SSSC_5$ focused on (ACC and TPR) and $SSSC_6$ focused on (TPR and TNR). In this experiment, the cost functions are represented by the following equation,

$$SSSC4_i = (0.5 * Acc_i) + (0.0 * TPR_i) + (0.5 * TNR_i)$$

$$SSSC5_i = (0.5 * Acc_i) + (0.5 * TPR_i) + (0.0 * TNR_i)$$

$$SSSC6_i = (0.0 * Acc_i) + (0.5 * TPR_i) + (0.5 * TNR_i)$$

iii. Apply cost function by combining three performance measures.

With the assumption that every performance measures are important, but the TNR is considered the most influential, this approach make a combination to put the highest weight on (TNR). In this experiment, the cost function is allocated an equal weight for (ACC and TPR) and slightly higher weight for TNR.

$$SSSC7_i = (0.3 * Acc_i) + (0.3 * TPR_i) + (0.4 * TNR_i)$$

Based on the above cost functions, the details of the algorithm have been constructed. The next section contains the algorithm's specifics.

A New Soft Set Ensemble Selection and Optimization Algorithm

Input: A Pruned Subset of Classifiers
Output: Optimal Set of Classifiers in Ensemble Team Based on Cost Function

1. ***Start***
2. Convert the decision table containing the Prediction and Actual Output to its minimal representation and create a NEW REDUCT.
3. Define the Cost Function (CoF) based on the combination of Prediction Accuracy weightage ($w_{c1} * Acc_i$), the sensitivity of the prediction ($w_{c2} * TPR_i$) and the specificity of the prediction($w_{c3} * TNR_i$).
4. Select the BEST classifier which has the best performance (CoF) and set it as a BASE classifier
5. Create a matrix of discernibility based on the NEW REDUCT.
6. Convert the discernibility matrix into a discernibility function using the transform function.
7. In order to get the set of the reduct, the absorption rule must be applied to NEW REDUCT.
8. Apply the distributive law to construct the reduct and become a NEW REDUCT
9. FOR ALL possibilities of NEW REDUCT
   9a. Find the BEST classifier from the NEW REDUCT, which has the best performance (CoF) and set it as a COMPLEMENT
   9b. Combine the BASE with the COMPLEMENT to become a NEW BASE
   9c. Apply the Simple Majority Voting technique to get the ACC of the NEW BASE
   9d. IF the ACC is higher than the BASE, continue step 4 to 8.
10. ***End***

4. RESULT AND DISCUSSION

The suggested method will be tested on a data set pertaining to liver disorders in the first experiment. Based on various cost functions, the suggested SSSC for the UCI liver disease data set is shown in Table 2.

Table- 2: SSSC on different Cost Function to UCI Liver Disorder Data Set

Cost Function	COF	ACC	TNR	TPR	Team of Ensemble
$SSSC_1$	0.74	0.75	0.87	0.56	Team 517
$SSSC_2$	1.00	0.60	1.00	0.00	Team 45
$SSSC_3$	0.83	0.55	0.37	0.83	Team 577
$SSSC_4$	0.69	0.74	0.81	0.63	Team 641
$SSSC_5$	0.84	0.71	0.97	0.32	Team 23
$SSSC_6$	0.72	0.75	0.87	0.56	Team 517
$SSSC_7$	0.76	0.75	0.87	0.56	Team 545

The goal of this experiment is to determine whether the suggested technique SSSC

with various cost function values produces different results when applied to the UCI repository's liver disease benchmarks data set. When just TNR is used, the $SSSC_2$(acc=0.0, tnr=1.0, tpr=0.0) may be picked, with OBJ = 1.00, ACC = 0.60, TNR = 1.00, and TPR = 0.00, and the selected ensemble team is Team 45. Despite its ability to forecast the non-risk patient with a perfect score, Team 45 struggles to anticipate the high-risk patient. If the user wishes to concentrate only on the TPR, and the most essential issue is the capacity to forecast high-risk patients with liver disorders, the $SSSC_3$(acc=0.0, tnr=0.0, tpr=1.0) is chosen. Despite having the highest TPR of 0.83 percent, Team 577's prediction accuracy is very poor, with an ACC of 0.55 and a TNR of 0.37 percent.

Also included in Table 2 are the results of combining two performance metrics on a liver disorder benchmarks data set from the University of California, Irvine repository. It can be seen from the findings that when the emphasis is placed on two performance measures, the performance of the other performance measure suffers a small decline. The studies have shown the clear disadvantage of concentrating on just one or two performance metrics.

Table 2 shows that the optimum cost function for the liver disorder data set when evaluating the combined value of ACC, TNR, and TPR is $SSSC_7$ (acc = 0.3, tnr = 0.3, tpr=0.4), with ACC = 0.75, TNR 0.87, and TPR 0.56 with Team 454 as the selected ensemble team. The chosen classifiers focused on the high-risk patient's prediction capacity (TPR) while also considering the high percentage of ACC and TNR.

The suggested SSSC method is then evaluated on the second UCI benchmark data set in the following experiment. Based on various cost functions applied to the UCI hepatitis data set, the suggested SSSC is shown in Table 3.

Table- 3: SSSC on different Cost Function to UCI hepatitis Data Set

Cost Function	COF	ACC	TNR	TPR	Team of Ensemble
$SSSC_1$	0.83	0.83	0.69	0.91	Team 73
$SSSC_2$	0.88	0.81	0.88	0.78	Team 585
$SSSC_3$	1.00	0.71	0.12	1.00	Team 693
$SSSC_4$	0.90	0.83	0.56	0.97	Team 523
$SSSC_5$	0.84	0.81	0.88	0.78	Team 585
$SSSC_6$	0.83	0.81	0.88	0.78	Team 585
$SSSC_7$	0.83	0.83	0.56	0.97	Team 523

In this experiment, the goal is to evaluate whether or not the proposed method SSSC with various cost function values provides a different outcome when compared to the hepatitis benchmarks data set from the University of California at Irvine database (UCI database). When testing on TNR alone, the $SSSC_2$(acc=0.0,tnr=1.0,tpr=0.0) may be chosen since it is focusing on the team that correctly predicts the patient who is not at risk of hepatitis C infection. With a COF of 0.88, ACC of 0.81, TNR of 0.88, a TPR of 0.78, and the TNR of 0.88, the ensemble is a good match for your needs. In contrast, if a user wants to focus only on the TPR and the ability to predict the high-risk patient is the most important consideration, the $SSSC_3$(acc=0.0,tnr=1.0,tpr=0.0) is selected. In spite of the fact that the COF is one, the prediction accuracy is low, with an ACC of 0.71 and a TNR of 0.02; as a consequence, Team 693 is selected, despite the high COF.

According to Table 3, the optimal cost function for the hepatitis data set when evaluating the combined value of ACC, TNR, and TPR is $SSSC_7$ (acc = 0.3, tnr = 0.3, tpr=0.4), with ACC = 0.83, TNR 0.56, and TPR 0.97, and the chosen ensemble team is Team 523. The chosen team of classifiers concentrated primarily on the prediction capacity of the high-risk patient (TPR), while also taking into consideration the high percentages of ACC and TNR.

In the third experiment, we put the suggested method to the test on a data set including breast cancer cases. Table 4 illustrates the proposed SSSC based on several cost functions applied to the UCI breast cancer data set.

Further, we want to test whether the suggested technique SSSC with various cost function values yields a significantly different result from the breast cancer benchmarks data set from the University of California, Irvine (UCI). Using the results in Table 4, it can be shown that multiple performance measures may be obtained by assigning various weights to a particular performance metric. By assigning the greatest amount of weight to a specific metric, the performance of that measure may achieve its

utmost potential.

Table- 4: SSSC on different Cost Function to UCI breast cancer Data Set

Cost Function	COF	ACC	TNR	TPR	Team of Ensemble
$SSSC_1$	0.81	0.81	0.36	0.95	Team 89
$SSSC_2$	0.57	0.75	0.57	0.79	Team 65
$SSSC_3$	1.00	0.75	0.00	1.00	Team 21
$SSSC_4$	0.88	0.81	0.36	0.95	Team 203
$SSSC_5$	0.63	0.75	0.50	0.84	Team 585
$SSSC_6$	0.68	0.74	0.57	0.79	Team 65
$SSSC_7$	0.73	0.81	0.36	0.95	Team 203

When just TPR is used, the $SSSC_3$(acc=0.0,tnr=0.0,tpr=1.0) may be selected, with a COF of 1.00, ACC of 0.75, TNR of 0.00, and TPR of 1.00, and an ensemble of Team 21. Even though Team 21 has the greatest TPR of 1.00 percent, the prediction accuracy is poor, and the TNR is 0.00 percent. On the other hand, $SSSC_2$(acc=0.0,tnr=1.0,tpr=0.0) is chosen if the user just wants to concentrate on the TNR and the ability to predict the non-risk patient is the most essential issue. The TNR is somewhat higher in this instance, at 0.57, but the prediction accuracy is poor, with an ACC of 0.74 percent.

Table 4 also illustrates the results of combining two performance metrics using the UCI repository's breast cancer benchmarks data set. The issue in this case is that when a study focuses on two performance measures, one of them will slightly decline. The disadvantage of concentrating on either a single or both performance metrics may be shown in the findings.

Finally, Team 203 with $SSSC_7$(acc=0.3,tnr=0.3,tpr=0.4) is the best ensemble team, producing the greatest ACC of 0.81, TNR of 0.36, and TPR of 0.95, as shown in Table 4. Team 203 was chosen because it met all of the criteria and was able to maintain the highest ACC.

5. CONCLUSION

This article proposes and tests a novel soft set based ensemble selection and optimisation technique on three benchmark data sets from the UCI repository. The heterogeneous ensemble was created by combining the results of 10 separate classification algorithms. In order to solve this problem, we previously presented a new soft set theory-based method for ensemble classifier selection and combination from the pruned ensemble. Based on soft set theory's dimensionality reduction, this novel method attempts to address the issue of choosing lesser redundant ensemble classifiers and identifying the optimum subset of classifiers. As a result, in prior work, we developed and evaluated a new soft set-based approach for selecting and combining classifiers from the pruned ensemble committee to forecast customer churn in telecommunication businesses.

The suggested model has been shown to be effective of finding the best subset of classifiers while maintaining the highest prediction accuracy. This article selected three sets of secondary data sets from the UCI benchmark repository to test and demonstrate that our proposed method is acceptable and works well on additional data sets. The suggested SSSC techniques are validated and verified using the chosen data sets. As a result, this paper has chosen three well known UCI data sets, which have been published and utilised in previous machine learning research. In earlier research, these data sets were often used as benchmarks to evaluate the performance of different categorisation methods. This paper demonstrated that the suggested model can find the optimum classifier subset and provide the best classification performance for the benchmark data sets.

REFERENCES:

[1] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Inf. Fusion*, vol. 41, pp. 195–216, 2018.

[2] Zhi-Hua Zhou, *Ensemble Methods Foundations and Algorithms*. Cambridge, UK AIMS: Chapman & Hall/CRC, 2014.

[3] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010.

[4] A. Jurek, Y. Bi, S. Wu, and C. Nugent, "A survey of commonly used ensemble-based classification techniques," *Knowl. Eng. Rev.*, vol. 29, no. 5, pp. 551–581, 2013.

[5] K. Dembczy, "Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains," *Proc. 27th Int. Conf. Mach. Learn.*, pp. 279–286, 2010.

[6] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification

algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, 2018.

[7] T. G. Dietterich, "Ensemble methods in machine learning," *Lect. Notes Comput. Sci.*, vol. 1857, pp. 1–15, 2000.

[8] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," *Expert Syst. Appl.*, vol. 42, no. 5, pp. 2785–2797, 2015.

[9] M. K. Awang, M. Makhtar, M. N. A. Rahman, and M. Mat, "A New Customer Churn Prediction Approach Based on Soft Set Ensemble Pruning."

[10] T. G. Dietterich, "Machine-learning research," *AI Mag.*, vol. 18, no. 4, pp. 97–136, 1997.

[11] M. Mohamad, M. Y. M. Saman, and M. S. Hitam, "The use of output combiners in enhancing the performance of large data for ANNs," *IAENG Int. J. Comput. Sci.*, vol. 41, no. 1, pp. 38–47, 2014.

[12] M. Mohamad, M. Y. M. Saman, and N. A. Hamid, "Complexity Approximation of Classification Task for Large Dataset Ensemble Artificial Neural Networks," *Lect. Notes Electr. Eng.*, vol. 520, no. April, pp. 195–202, 2019.

[13] H. Wang, T. M. Khoshgoftaar, and K. Gao, "A comparative study of filter-based feature ranking techniques," *2010 IEEE Int. Conf. Inf. Reuse Integr. IRI 2010*, pp. 43–48, 2010.

[14] J. Ethridge, G. Ditzler, and R. Polikar, "Optimal ν-SVM parameter estimation using multi objective evolutionary algorithms," *2010 IEEE World Congr. Comput. Intell. WCCI 2010 - 2010 IEEE Congr. Evol. Comput. CEC 2010*, 2010.

[15] M. Makhtar, D. C. Neagu, and M. J. Ridley, "Comparing multi-class classifiers: On the similarity of confusion matrices for predictive toxicology applications," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6936 LNCS, pp. 252–261, 2011.

[16] Q. Shen, R. Diao, and P. Su, "Feature Selection Ensemble," *Turing-100*, vol. 10, pp. 289–306, 2012.

[17] M. Bhardwaj and V. Bhatnagar, "Towards an optimally pruned classifier ensemble," *Int. J. Mach. Learn. Cybern.*, vol. 6, no. 5, pp. 699–718, 2015.

[18] Y. Freund and R. R. E. Schapire, "Experiments with a new boosting algorithm," in *International Conference on Machine Learning*, 1996, pp. 148–156.

[19] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 421, pp. 123–140, 1996.

[20] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective fusion of heterogeneous classifiers," *Intell. Data Anal.*, vol. 9, no. 6, pp. 511–525, 2005.

[21] I. Partalas, G. Tsoumakas, and I. Vlahavas, "An ensemble uncertainty aware measure for directed hill climbing ensemble pruning," *Mach. Learn.*, vol. 81, no. 3, pp. 257–282, 2010.

[22] R. Caruana, a Niculescu-Mizil, G. Crew, and a Ksikes, "Ensemble Selection from Librairies of Models," *Icml*, vol. 34, pp. 1–21, 2011.

[23] G. D. C. Cavalcanti, L. S. Oliveira, T. J. M. Moura, and G. V. Carvalho, "Combining diversity measures for ensemble pruning," *Pattern Recognit. Lett.*, vol. 74, pp. 38–45, 2016.

[24] I. Partalas, G. Tsoumakas, and I. Vlahavas, "A Study on Greedy Algorithms for Ensemble Pruning," 2012.

[25] Y. Zhang, S. Burer, and W. N. Street, "Ensemble Pruning Via Semi-definite Programming," *J. Mach. Learn. Res.*, vol. 7, no. Jul, pp. 1315–1338, 2006.

[26] Z. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell. 137*, vol. 137, no. 1–2, pp. 239–263, 2002.

[27] F. Markatopoulou, G. Tsoumakas, and I. Vlahavas, "Dynamic ensemble pruning based on multi-label classification," *Neurocomputing*, vol. 150, no. PB, pp. 501–512, 2015.

[28] A. Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, "Ensemble selection from libraries of models," in *Proceedings of the 21st International Confer- ence on Machine Learning*, 2004, p. 18.

[29] S. BVST, "Soft Set Based Techniques for Mining Uncertain Data," *Int. J. Math. Comput.*, vol. 3, no. 3, pp. 57–64, 2013.

[30] T. Herawan, A. N. M. Rose, and M. Mat Deris, "Soft set theoretic approach for dimensionality reduction," *Commun. Comput. Inf. Sci.*, vol. 64, no. 2, pp. 171–178, 2009.

[31] A. N. M. Rose, M. I. Awang, H. Hassan, A. H. Zakaria, T. Herawan, and M. M. Deris, "Hybrid reduction in soft set decision making," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6838 LNCS, pp. 108–115, 2011.

[32] Z. Kong, L. Wang, Z. Wu, and D. Zou, "A new parameter reduction in fuzzy soft sets," in *Proceedings - 2012 IEEE International Conference on Granular Computing, GrC 2012*, 2012, pp. 730–732.

[33] H. Aktas *et al.*, "Soft sets and soft groups," *Comput. Math. with Appl.*, vol. x, no. 1, pp. 1–11, 2015.

[34] A. Nazari, M. Rose, M. I. Awang, and H. Hassan, "Hybrid Reduction in Soft Set Decision Making," pp. 108–115.

[35] D. Ruta and B. Gabrys, "Classifier selection for majority voting," *Inf. Fusion*, vol. 6, no. 1, pp. 63–81, 2005.