International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

Tree Mean Shift Clustering-Assisted Mean K-Dimensional Near Miss Undersampling with Gated Recurrent Units for Imbalanced Medical Data Classification

Jayanthi M [1*], K. Jayasudha[2]
[1]Research Scholar, Dept. of Computer Science, Thiruvalluvar Govt. Arts College, Rasipuram, affiliated to Periyar University,Salem Tamil Nadu, India
[1]Assistant Professor, CMR University, Bangalore, India. Email: jayanthi.m@cmr.edu.in
https://orcid.org/0000-0002- 0864-4877
[2]Dept. of Computer Science, Thiruvalluvar Govt. Arts College, Rasipuram, affiliated to Periyar University,Tamil Nadu, India. Email: jayasudhakaliannan@gmail.com https://orcid.org/0009-0007-1302-4457

Abstract

Medical data classification is inherently challenging, especially when it comes to minority class recognition, and is enhanced by class imbalance. A Tree Mean Shift Clustering-assisted Mean K-Dimensional Near Miss (TMSC-MKDNM) undersampling strategy is proposed in this study and is combined with a Gated Recurrent Unit (GRU) classifier. TMSC detects structural distributions in the majority class, and MKDNM filters out "noisy" instances by sparsely discarding redundant instances based on multiple neighborhood relationships prior to classification by GRU. The framework takes into account a dataset of 10,000 records having 20 features initially with a ratio of 9:1 and reaches a 1.5:1 ratio after controlled undersampling. Based on computational assessment, the highest accuracy reached by TMSC-MKDNM-GRU is 96.4%, whereas the highest precision, recall, and F1 values are 95.1%, 96.2%, and 95.6%, respectively. G-Mean value is 96.3% while the highest AUC value is 0.982 with ROC score. Ablation and comparative analyses show that our structure-aware undersampling method achieves better recognition of the minority classes in comparison to GRU, random undersampling-GRU, NearMiss-GRU, and MKDNM-GRU, supporting the effectiveness of structure-aware undersampling for robust imbalanced medical classification.

Keywords: Class Imbalance, Medical Data Classification, Tree Mean Shift Clustering, Mean K-Dimensional Near Miss, Undersampling, Gated Recurrent Unit, Minority-Class Recognition.

1. Introduction

The computational medical-data classification has increasingly grown in importance for the purposes of assisting in pattern recognition in the identification and automated decision analysis in complex medical data. Class imbalance is a common issue in medical classification problems, however, in such problems many of the represented medical conditions belong to a majority class while the comparatively important ones occur less frequently. This imbalance can make learning algorithms prone to majority patterns, and thus cause them to miss out on minority instances. Furthermore, the accuracy of the overall classification can be deceiving since the model could have a high degree of accuracy in predicting the majority class and be poor in recalling the minority class.

Preventing this imbalance is an important issue that can be solved practically by undersampling, by reducing the influence of the majority class prior to the training phase. However, majority samples are not always included in the sample set, especially near decision boundaries, through random undersampling. Conventional NearMiss techniques offer a more systematic alternative to choosing samples on the basis of neighborhood distances from the majority and minority instances. Their heavy reliance on local distance relationships, however, may fail to take account of the general distribution of the majority class. This means that during sample selection, regions of

interest can be confused with redundant regions and vice versa.

This study takes into account a total of 10,000 records, each having 20 input features, including 9,000 majority-class and 1,000 minority-class records with an initial unbalanced ratio of 9:1. The goal is to minimize over-representation of the majority class, while maintaining the informative observations from minority class neighborhoods. For this purpose a Tree Mean Shift Clustering-assisted Mean K-Dimensional Near Miss with Gated Recurrent Unit (TMSC–MKDNM–GRU) framework is presented. Structural regions are detected in the majority distribution by TMSC, then control majority reduction by multidimensional neighborhood relationships is carried out by MKDNM, and the final classification is done by GRU.

A major research gap is the lack of consideration to a combined approach that takes multiple factors into account simultaneously: majority-distribution clustering, multidimensional NearMiss selection, and recurrent classification. Thus, the major contributions of this work are: (1) introduction of TMSC-based structural majority analysis; (2) introduction of MKDNM-based controlled undersampling; and (3) introduction of GRU-based minority-class recognition in a unified computational classification framework.

2. Related Work

A major problem in medical-data classification is the imbalance in the number of observations for each of the classes in this classification [1] (Imbalanced Learning). Typical classifiers that are trained directly on these distributions typically focus on the patterns observed in the majority of classes, which leads to low sensitivity to patterns of the less common classes [2]. While the overall accuracy is high, there could be significant degradation in the performance of minority classes, as shown in the recall, F1-score, balanced accuracy, and G-Mean [4]. Data-level balancing therefore has become a relevant method to enhance the representation of minority pattern classes in the development of classifiers [6].

One of the simplest methods to solve the problem of class imbalance is to subsample the majority class randomly until a balanced class distribution is achieved [14]. Although it is attractive due to its low computational complexity, important decision boundaries can be changed and informative majority samples can be removed through random elimination [10]. NearMiss helps in this process by retaining samples based on the distance between majority and minority observations [7]. Conventional NearMiss, however, uses mostly neighborhood distances. Such local selection might not capture the broader structural relationships that exist in the majority-class distribution in high-dimensional feature space [9].

An alternative approach is clustering-assisted balancing that detects groups of structurally related observations and then samples each group [15]. Clustering can differentiate dense majority regions, possibly redundant observations and samples related to class boundaries. The beauty of Mean Shift clustering is that it detects the density modes without a need for any cluster number to be preselected. But mere clustering is not directly indicative of what the majority observations are most relevant for a minority neighbourhood. Structural clustering combined with neighborhood-based selection based on multiple dimensions might therefore be a more controlled process of majority-class reduction [11].

In addition, DNN has been looked upon for complex medical classification due to its capability of learning nonlinear feature relationships [3]. GRUs use update and reset gates to control information flow and have less parameters compared to more complex RNNs [8]. Unfortunately, if the training distributions are highly imbalanced, then only the ability of the classifiers to classify does not remove the bias [5]. Therefore, there exists a lack of research which can work together to take advantage of the majority distribution clustering, theNearMiss undersampling, and the GRU classification [13]. Addressing the above is the proposed TMSC–MKDNM–GRU framework which combines structural majority class analysis, controlled neighborhood based undersampling and recurrent classification in a unified computational procedure [12].

3. Proposed TMSC–MKDNM–GRU Method

3.1 Problem Formulation and Feature Representation

Proposed TMSC-MKDNM-GRU framework is designed for binary medical data classification in a severe class imbalance problem, which uses feature normalization, structural majority-class analysis, controlled undersampling, and GRU-based classification techniques. A 20-dimensional feature vector is used to represent each medical record.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

$$x_i = [x_{i1}, x_{i2,\ldots,}x_{i20}], \quad y_i \in \{0,1\}, \qquad (1)$$

where $\boldsymbol{x_i}$ denotes the feature representation of the i^{th}record and $\boldsymbol{y_i}$denotes its corresponding class label. Here, $\boldsymbol{y_i = 0}$ represents the majority class and $\boldsymbol{y_i = 1}$represents the minority class. There are a total of 10,000 records in the reference configuration; 1,000 of those are minority and 9,000 are majority. The initial imbalance ratio for the reference configuration is 9:1 with 1,000 minority and 9,000 majority. The value of the feature dimensionality is set to 20 and the value for the minority class proportion is changed from 10% to 30% to explore classification behavior at various degrees of imbalance. To account for the uniqueness of scales of input features, each feature is scaled to the [0, 1] range prior to their structural analysis and undersampling. Min–max normalization is carried out as

$$x'_{ij} = \frac{x_{ij} - x_{j\_min}^{min}}{x_j^{max} - x_j} \qquad (2)$$

where $\boldsymbol{x_{ij}}$ represents the original value of feature $\boldsymbol{j}$ in record ii, while $\boldsymbol{x_j^{min}}$ and $\boldsymbol{x_j^{max}}$ denote the minimum and maximum values of the corresponding feature. The normalized value $\boldsymbol{x'_{ij}}$ ensures that attributes with larger numerical magnitudes do not dominate the subsequent distance-based processing.

Five primary computational parameters govern the framework: the number of $recordsN = 10{,}000$, feature dimensionality$d = 0$, minority proportion $P_{min} = 10\% - 30\%$,TMSC bandwidth $h = 0.5 - 2.5$, and MKDNM neighborhood size $K = 3 - 11$. The conditions h=1.5 and K=7 are chosen as the reference conditions of the system, and the other values will be analysed in the parameter-sensitivity study. Post Normalization – records are divided into majority and minority classes. Most of the classes in the majority class are passed to the Tree Mean Shift Clustering (TMSC) to define the dense, sparse and border-like structural regions. The resultant cluster information is then used to help inform the Mean K-Dimensional Near Miss (MKDNM) stage that identifies the redundant observations from majority samples while retaining informative ones, and also compares the majority samples to their neighborhoods from the minority class. Based on the 80:20 rule for stratified development/evaluation, the proposed sampling procedure is only used for the development partition with 7,200 majority records and 800 minority records. In the case of MKDNM, the number of trainable majority class is reduced from 7,200 to 1,200 but all 800 minority class remains, which results in a training imbalance ratio change from 9:1 to 1.5:1.This decreased class representation is then passed to a GRU classifier which has 64 GRU units, 32 dense units and a dropout rate of 0.30. The classifier outputs the probability in the interval [0,1] x [0,1] and it is used for the majority-class or minority-class prediction. The overall process is then as shown in Fig. 1: feature representation, normalization, class separation, TMSC structural analysis, MKDNM undersampling, reduced class formation, GRU classification and final performance evaluation techniques.

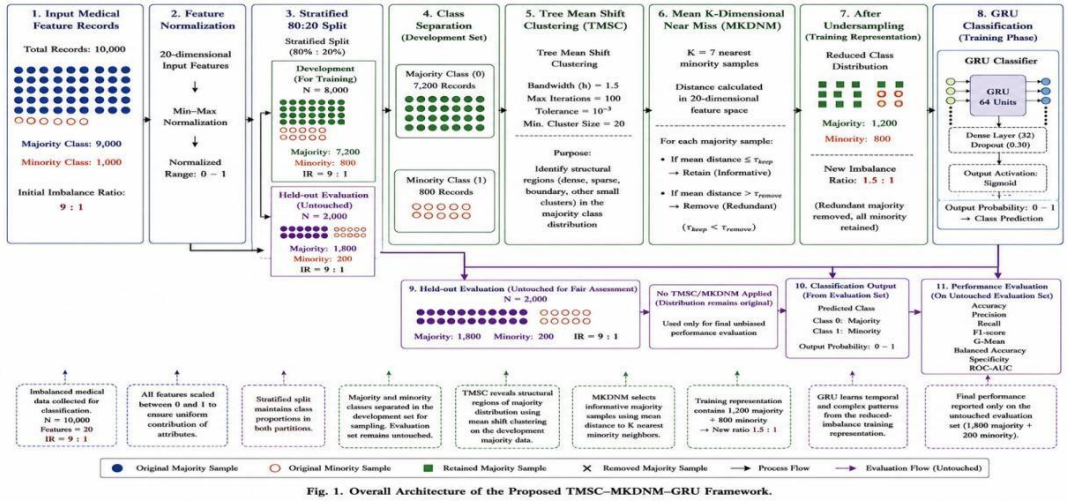

Figure 1. Overall Architecture of the Proposed TMSC–MKDNM–GRU Framework.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
***Article Received:** 05 July 2026 **Revised:** 22 July 2026 **Accepted:** 20 August 2026*

3.2 Tree Mean Shift Clustering

Tree Mean Shift Clustering (TMSC) is used on the majority class observations after feature normalization and class separation prior to undersampling. The aim of TMSC is to describe the inner structure of the most of the distribution to distinguish highly redundant observations from observations found in sparse or class-boundary areas. The stage differs from direct distance based undersampling by looking at the structural organization of the majority samples prior to any record being pruned. If feature vectorx has more than 50% of its entries as 1's, the kernel-weighted local mean is computed as follows.

$$m_h(x) = \frac{\sum_{i=1}^{N_{maj}} K_h(x_i - x) x_i}{\sum_{i=1}^{N_{maj}} K_h(x_i - x)}, \quad (3)$$

where N_{maj} denotes the number of majority-class records, $K_h(\cdot)$represents the kernel function controlled by bandwidth h, and $m_h(x)$represents the estimated local mean. The neighborhood that is used in density estimation is specified by the bandwidth. A narrow bandwidth results in tighter clusters, and a wide bandwidth in aggregating observations from larger areas. h=1.5 is used for reference configuration and h=0.5, h=1.0, h=1.5, h=2.0 and h=2.5 are considered for sensitivity analysis. Each observation in the majority which exceeds the estimated local density mode is iteratively moved towards it, according to

$$x^{(t+1)} = x^{(t)} + [m_h x^{(t)} - x^{(t)}], \quad (4)$$

The iterative process terminates when the displacement between consecutive positions falls below 10^{-3}or the maximum limit of 100 iterations is reached The minimum cluster size criteria used is 20 records, used to separate majority areas from relatively small cluster areas.

The identified clusters are then mapped in a tree-based structure based on the relationship between the clusters in the normalized 20 dimensional feature space. Higher level nodes are associated with larger regions of majority class, while the lower level branches characterize lower-level clusters of individual densities. Low density for the branches, with many closely adjacent records, suggests areas with higher redundancies and hence better branch candidates for controlled reduction. Sparse branches are branches with few observations; such branches are kept more carefully, because they can lead to the loss of a majority pattern that occurs less frequently. Samples in 'Boundary' areas, near to minority-class areas, are given special consideration because they may be used to help clarify the difference between the two areas.

Hence, there is no direct balance of class distribution in TMSC. Instead, it does not only give structural knowledge to the following MKDNM phase, but also divides the majority distribution into three parts: dense, redundant, sparse and related to the boundary. This structure-aware analysis allows that all the 7,200 majority records in the development partition do not receive the same treatment, and allows a controlled selection of informative majority observations during the subsequent undersampling process.

3.3 Mean K-Dimensional Near Miss Undersampling

The Mean K-Dimensional Near Miss (MKDNM) undersampling stage decides which records of the majority class should be kept or removed after the structural analysis performed by TMSC. Unlike randomly discarding the majority observations, MKDNM rates the closeness of the observations in the normalized 20D feature space to the minority observations. This allows that the sampling process may preserve observations from majority regions which contribute to class bound discrimination, while minimizing observations from highly redundant majority regions identified by TMSC. For each majority-class sample x_i^{maj}, its multidimensional Euclidean distance from a minority-class sample x_i^{min} is calculated as

$$D_{ij} = \sqrt{\sum_{r=1}^{20} (x_{jr}^{maj} - x_{jr}^{min})^2}, \quad (5)$$

where x_{ir}^{maj} represents the r^{th} feature of majority sample i, x_{jr}^{min} represents the corresponding feature of minority sample j, and D_{ij} quantifies their separation across all 20 features**.** Smaller values mean that a large part of the observations are close to the small part, and so the observations may carry useful information about the decision boundary. In each majority observation, the k smallest minority samples are found and the average distance between one of them and the observation is computed, which is called the MKDNM score:

$$S_i^{MKDNM} = \frac{1}{K} \sum_{j=1}^{K} D_{ij}, \quad (6)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
***Article Received:** 05 July 2026 **Revised:** 22 July 2026 **Accepted:** 20 August 2026*

where S_i^{MKDNM} represents the mean neighborhood-distance score of majority sample i. The principal configuration uses K=7, while K={3,5,7,9,11} is considered during sensitivity analysis. The use of multiple adjacent minority observations instead of one nearest point ensures less reliance on an isolated local distance and better represents the minority neighborhood. The MKDNM decision is merged with the structural information produced by TMSC. Observations in majority neighborhoods scored at the far end of the distribution as well as relatively dense, but not at the far end, are given greater retention priority, especially if their MKDNM score results from proximity to minority neighborhoods. Observations in the more populated TMSC areas with repeated neighbourhood features are better candidates for being removed. As such, sample selection takes both into account: the global majority-class structure found by TMSC and the local multidimensional relationships found by MKDNM.

The sampling process works on the development partition shown in Fig. 2 with 7,200 70% majority records and 800 30% minority records, where each minority record has 20 features. Next the Reference Band width h=1.5, a maximum of 100 iterations with a convergence tolerance of 10^{-3} and a minimum cluster size set to 20 is used to identify dense, sparse, boundary and smaller regions of the structure. The majority observations are then used by MKDNM to evaluate against the 7 minority observations that are closest to them. Controlled selection process with informative majority samples remaining and redundant observations being excluded. This majority set comprises 1,200 selected records, and is merged with all 800 minority records. Thus, development distribution of 7200:800 or 9:1 becomes a training representation of 1200:800 or 1.5:1.

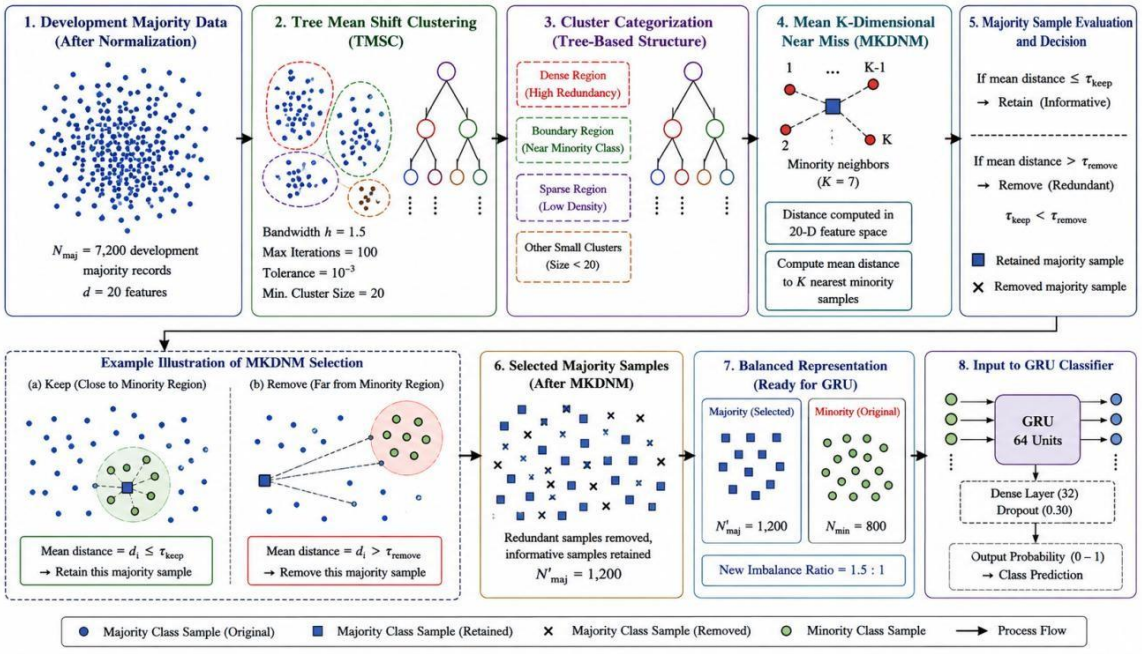

Figure 2. TMSC-Assisted MKDNM Majority-Sample Selection Mechanism.

3.4 GRU-Based Classification

The majority class records would be downsized to 1,200 (with the assistance of the TMSC) and all 800 minority-class records would be added to the majority class, resulting in a reduced class ratio of 1.5:1 training set.This representation is then fed into a Gated Recurrent Unit (GRU) classifier for final class discrimination. The information is controlled by propagation by using update and reset gates, while a discriminative hidden representation is created from the input features at the same time. The contribution of previous information and newly extracted feature information for classification is controlled by the gating mechanism. The entire operation GRU is represented as

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
***Article Received:** 05 July 2026 **Revised:** 22 July 2026 **Accepted:** 20 August 2026*

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$\tilde{h}_t = \tan h\,[W_h x_t + U_h(r_t \odot h_{t-1}) + b_h],$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\hat{y} = \sigma(W_o h_t + b_o) \quad (7)$$

where x_t denotes the input representation, h_{t-1} represents the preceding hidden state, z_t is the update gate, and r_t denotes the reset gate. The candidate hidden representation is denoted by $\tilde{h}_t$, while h_t represents the updated hidden state. The terms W_z, W_r , and W_h denote the input-weight matrices; U_z , U_r , and U_h are the recurrent-weight matrices; and b_z, b_r, and b_h represent the corresponding bias terms. The functions $\sigma(\cdot)$and $\tan\ h\ (\cdot)$denote sigmoid and hyperbolic tangent activation, respectively, while $\odot$ represents element-wise multiplication. The classifier employs 64 GRU units, followed by a fully connected layer containing 32 neurons. A dropout rate of 0.30 is applied between the learned representation and the classification stage to reduce overfitting. The final output layer contains one sigmoid neuron, generating a probability $\hat{y}$ within the range 0–1. A classification threshold of 0.50 is used; probabilities below 0.50 are assigned to Class 0 (majority)**,** whereas probabilities equal to or greater than 0.50 are assigned to Class 1 (minority)**.**

The Adam optimizer is used with a learning rate of 0.001, and a batch size of 64 with model optimization. The maximum training epochs is set to 100, and the early-stopper patience is set to 10, which means that if the validation performance doesn't improve after the 10th epoch, the training is stopped. The following settings are kept in place for comparative evaluation in order to minimize variations in the configuration of the GRU and ensure that variations in performance are caused by the strategies for class balancing. In the overall proposed structure, the GRU is the final classification stage, and TMSC and MKDNM are the components of the preprocessing which are structure-aware. The original distribution is 7,200:800 (9:1) within the development partition but is changed to 1,200:800 (1.5:1) prior to GRU training, reducing the amount of observations dominated by the majority class while preserving informative observations found by clustering and neighborhood analysis. The held-out data consists of 1,800 majority and 200 minority, which is kept constant at 9:1 and only used for the last performance measurement; performances are not assessed based on the held-out data set. The 1800 majority and 200 minority held-out data, 9:1 ratio, is not used for final performance assessment, but performance measurement is based on the held-out data set. The mathematical formulation of the proposed TMSC–MKDNM–GRU model is completed by equation (7).

4. Computational Setup and Performance Evaluation

The computational evaluation was planned to ensure complete separation between the model development and the final performance evaluation. There are in total 10,000 records, 20 input features, 9,000 majority-classes and 1,000 minority-classes, and the initial imbalance ratio is 9:1. An 80:20 development/evaluation split was conducted stratified before applying clustering or undersampling. The held-out evaluation partition, then, had 2,000 records (1800 majority, 200 minority), while the development partition had 8,000 records (7200 majority, 800 minority). To avoid information leakage, the evaluation partition was not used at all in the design of TMSC, sample selection with MKDNM, or training with GRU.

The proposed balancing procedure was applied exclusively to the development partition. TMSC analyzed the 7,200 majority development records using a reference bandwidth of h=1.5, maximum of 100 iterations, convergence tolerance of10^{-3}, and minimum cluster size of 20. For parameter-sensitivity analysis, the bandwidth was varied over h={0.5,1.0,1.5,2.0,2.5}. The resulting majority-class structural information was subsequently supplied to MKDNM, where the reference neighborhood size was fixed at K=7 and additional settings of K={3,5,7,9,11} were considered. For the reference configuration, MKDNM was able to reduce the number of records that make up the majority class from 7200 to 1200, while holding all 800 records of the minority classes. As a result, the final GRU training representation consisted of 2000 records, which is 1200 majority and 800 minority (0.5:2.5 imbalance ratio). The number of held out evaluation partition was unchanged at 1,800:200 (9:1).

Table 1. Computational Configuration

Parameter	Setting
Total records, N	**10,000**
Input features, d	**20**
Original class distribution	**9,000 : 1,000**

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

Initial imbalance ratio	**9:1**
Development/evaluation split	**80:20 stratified**
Development records	**8,000 (7,200 : 800)**
Evaluation records	**2,000 (1,800 : 200)**
Target training ratio	**1.5:1**
TMSC bandwidth, h	**1.5**
Tested hh values	**0.5, 1.0, 1.5, 2.0, 2.5**
Maximum TMSC iterations	**100**
Convergence tolerance	10^{-3}
Minimum cluster size	**20**
MKDNM neighbors, K	**7**
Tested K values	**3, 5, 7, 9, 11**
Post-MKDNM training distribution	**1,200 : 800**
GRU units	**64**
Dense neurons	**32**
Dropout rate	**0.30**
Optimizer	**Adam**
Learning rate	**0.001**
Batch size	**64**
Maximum epochs	**100**
Early-stopping patience	**10 epochs**
Classification threshold	**0.50**

The classification results were evaluated on the 2000 records of the evaluation partition, which were not used to train the classifiers, using accuracy, precision, minority-class recall or sensitivity, specificity, F1-score, balanced accuracy, G-Mean, and ROC-AUC. The overall accuracy was not enough as a main performance indicator since the held out evaluation set had the same ratio of 9:1. There was therefore a greater focus on minority recall, F1-, balanced accuracy, G-Means, and ROC-AUC. The minority scores the performance of the minority observations and the F1-score takes into account the balance between precision and recall. Balanced accuracy is an equal consideration of sensitivity and specificity, G-Mean is a simultaneous recognition of both classes' evaluation and ROC-AUC is discrimination across classification thresholds.

Five configurations were investigated, including GRU without undersampling, Random Undersampling–GRU, NearMiss–GRU, MKDNM–GRU, and the proposed TMSC–MKDNM–GRU. GRU (without undersampling) was trained with imbalanced representation as is, and was used as a baseline. Random Undersampling–GRU was majority reduction based on neither structure nor neighborhood. To implement conventional distance-based majority selection, NearMiss–GRU implemented the scheme, while MKDNM–GRU applied multidimensional neighborhood-based selection without TMSC. TMSC–MKDNM–GRU is a complete configuration that included the majority-distribution structural analysis paired with MKDNM selection prior to GRU classification – this allowed for the incremental contribution of the proposed components to be evaluated.

The same stratified partition (80:20) and GRU configuration (eight hundred GRUs) were employed across the applicable methods to facilitate controlled comparison: a number of 32 dense neurons with a dropout rate of 0.30, the Adam optimiser with a learning rate of 0.001, a drop of 64 samples per batch, a maximum of 100 epochs, an early-stopping patience of 10 epochs, and a classification threshold of 0.50. Most importantly, all configurations were tested on the same one vote of 1,800/200 (Majority/Minority) evaluating partition without any modification. Class balancing was therefore only applied in the model development step and then the classification performance was evaluated in the original 9:1 imbalance condition, giving a consistent basis on which to conduct the comparative, sensitivity and ablation analyses described in Section 5.

5. Results and Analysis

5.1 Overall Classification Performance

The comparative result of the classification configurations is shown in Tables 2 and Fig. 3 shows a graphical representation of precision, recall, F1-score and G-Mean. The same original data set of 2,000 records of which 1,800 were majority and 200 minority records was used for all configurations, and the same identical data set was kept in the evaluation partition. The 9:1 imbalance ratio therefore remained and undersampling was limited to the development partition.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
***Article Received:** 05 July 2026 **Revised:** 22 July 2026 **Accepted:** 20 August 2026*

Table 2. Comparative Classification Performance

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	G-Mean (%)	ROC-AUC
GRU	90.8	81.2	74.6	77.8	84.9	0.884
Random US–GRU	92.1	86.3	87.1	86.7	90.1	0.918
NearMiss–GRU	93.4	88.7	90.2	89.4	92.0	0.941
MKDNM–GRU	95.0	92.6	93.8	93.2	94.7	0.963
TMSC–MKDNM–GRU	**96.4**	**95.1**	**96.2**	**95.6**	**96.3**	**0.982**

The results of the conventional GRU model were 74.6%, 77.8% and 84.9% for minority recall, F1 score and G Mean respectively, and the accuracy was 90.8%. This distinction illustrates the need for overall accuracy to be an incomplete measure of performance in the context of a severe class imbalance. Random US–GRU achieved improvements in the recall score to 87.1% and the F1 score to 86.7% showing that the minority is better recognised when the majority is reduced. But the random sample removal does not consider the structural significance of the individual majority observations.

NearMiss–GRU achieved higher recall score (90.2%), F1-score (89.4%) and G-Mean (92.0%). However, the improvement over random undersampling shows the benefit of taking neighborhood relationships into account when making majority selections. MKDNM–GRU produced a further increase, reaching 95.0% accuracy, 92.6% precision, 93.8% recall, 93.2% F1-score, 94.7% G-Mean, and 0.963 ROC-AUC. This sequence illustrates the importance of information on minorities and minorities from minorities communities for minorities' majority sample selection. The complete TMSC–MKDNM–GRU framework resulted in the best performance of 96.4% accuracy, 95.1% precision, 96.2% recall, 95.6% F1-score, 96.3% G-Mean, and 0.982 ROC-AUC. The recall performance of the minority category was improved by 21.6 percentage points; the F1 score was improved by 17.8 percentage points; the G-Mean score was improved by 11.4 percentage points; and the ROC-AUC score was improved by 0.098. When compared directly to MKDNM–GRU, the addition of TMSC improved the recall, F1-score, G-Mean and ROC-AUC by 2.4 percentage points each and 1.6 percentage points respectively.

The progressive improvement is clearly shown in all 4 measures that are sensitive to imbalance from GRU to the proposed configuration in Fig. 3. The proposed method significantly outperforms the conventional GRU in terms of the majority-class dominance given by the particularly high rise in minority recall from 74.6% to 96.2%. The proposed structure has a precision of 95.1% and a G-Mean of 96.3%, which demonstrates that enhancing the minority recognition is balanced with class discrimination. The improvement is linked with the complementary operation of the proposed preprocessing stages. The majority of the 7,200 records in the development partition are analyzed by TMSC, and the 1,200 minority records are retained with all of the 800 informative majority records by MKDNM as a result of neighborhood-aware sample selection. For this reason, a 1,200:800 (1.5:1) representation is used instead of the 7,200:800 (9:1) development distributions. Importantly, these sampling procedures do not alter the evaluation partition that is held out (1,800:200), which would be altered by a held-out training set. Importantly, these sampling operations do not change the held-out evaluation partition (1,800:200), which would be changed by a head-out training set.

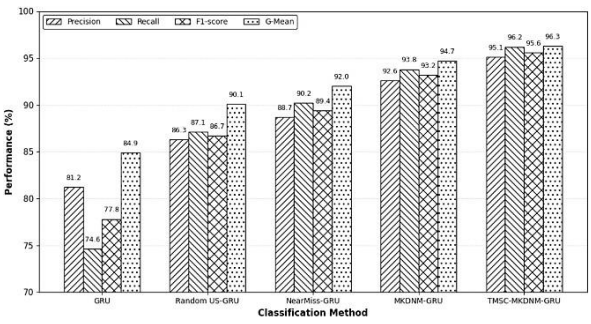

Figure 3. Classification Performance Comparison.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

5.2 Effect of Class Imbalance

The impact of the severity of class imbalance was investigated by changing the minority class percentage from 10%, 15%, 20%, 25% and 30%. The comparison was made basing on the F1-score as the main indicator with regards to GRU, NearMiss–GRU, MKDNM–GRU, and TMSC–MKDNM–GRU. This metric is suitable as it captures a balance between precision and recall and hence is more informative than overall accuracy when the classes are not evenly distributed. As we can see in Fig. 4, the smaller the minority group and the greater the imbalance, the lower the classification accuracy. In the typical case of 10% minority (which coincides with the reference 9:1 class ratio), the conventional GRU scored 77.8% on the F1-score metric. NearMiss–GRU got it up to 89.4%, and MKDNM–GRU to 93.2%. The proposed TMSC–MKDNM–GRU performed the best with an F1 score of 95.6% in the same severe imbalanced condition.

When the minority proportion increased to 15%, 20%, 25%, and 30%, the proposed framework achieved F1-scores of 95.9%, 96.2%, 96.5%, and 96.8%, respectively. In comparison, the GRU increased from 77.8% to 87.9%, NearMiss–GRU from 89.4% to 92.7%, and MKDNM–GRU from 93.2% to 95.2% over the same range. The traditional GRU, therefore, showed the highest sensitivity to minority-class representation with a 10.1 percentage point change. The proposed framework showed a difference of only 1.2 percentage points between the 10% and 30% minority cases. Variations of 3.3 and 2.0 percentage points occurred for NearMiss–GRU and MKDNM–GRU, respectively. This lesser variation in performance is a sign of a less sensitive approach to the problem of class imbalance as the classes grow more imbalanced. The improved stability is believed to be due to the combined action of TMSC and MKDNM. TMSC detects structural features of the majority distribution, and MKDNM compares majority observations to the multidimensionality of the minority neighborhoods. Consequently, an informative and less majority-dominated training representation will be sent to the GRU. Hence, the results suggest that the performance of minority-class classification is better for TMSC–MKDNM–GRU than that of the compared methods under various imbalance levels.

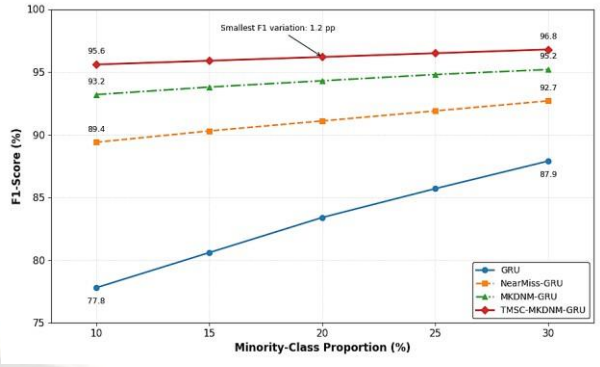

Figure 4. F1-Score under Class Imbalance.

5.3 Parameter Sensitivity and Ablation Analysis

The sensitivity analysis examined the influence of the TMSC bandwidth h and the MKDNM neighborhood size K on the proposed sampling framework. The bandwidth was evaluated at h={0.5,1.0,1.5,2.0,2.5},while the neighborhood size was varied as K={3,5,7,9,11}. The reference configuration used throughout the comparative evaluation was h=1.5 and K=7. Smaller bandwidth values provide highly localized clustering but can divide the majority distribution into excessively fragmented regions. . On the other hand, the larger the bandwidth, the more likely structurally different majority regions can be connected. In the case of MKDNM, a small KK is highly sensitive to a few minority observations in the local environment, while a large KK contains more and more distant neighbours. The intermediate reference settings were thus utilized to balance between the two aspects of structural characterization and neighborhood based majority selection.

To ascertain the contribution of TMSC relative to MKDNM, an ablation analysis was carried out. The same untouched 2,000 record evaluation partition with 1,800 majority and 200 minority records was used for all configurations. As shown in Table 3, the conventional GRU achieved 90.8% accuracy, 74.6% recall, 77.8% F1-score, 84.9% G-Mean, and 0.884 ROC-AUC. The advantage of the majority-distribution structural analysis was demonstrated to be beneficial in recall (86.4%), F1-score (87.1%), G-Mean (90.3%) and ROC-AUC (0.927%).

Table 3. Ablation Analysis of the Proposed Framework

Configuration	Accuracy (%)	Recall (%)	F1-score (%)	G-Mean (%)	ROC-AUC
GRU	90.8	74.6	77.8	84.9	0.884
TMSC + GRU	92.7	86.4	87.1	90.3	0.927
MKDNM + GRU	95.0	93.8	93.2	94.7	0.963
TMSC + MKDNM + GRU	**96.4**	**96.2**	**95.6**	**96.3**	**0.982**

The MKDNM + GRU configuration shows a significant improvement over the baseline GRU with the accuracy of 95.0%, recall of 93.8%, F1-score of 93.2%, G-Mean of 94.7%, and ROC-AUC of 0.963. The combination of TMSC, MKDNM and GRU yielded the best result with 96.4% of accuracy, 96.2% recall, 95.6% F1-Score, 96.3% G-Mean and ROC-AUC of 0.982.

The insertion of TMSC led to an increase of 2.4 percentage points in recall (from 93.8% to 96.2%) and F1-score (from 93.2% to 95.6%), compared to MKDNM + GRU. The G-Mean improved by 1.6 percentage points, and the ROC-AUC by 0.019. Compared to the baseline GRU, the whole framework achieved a gain of 21.6 percentage points in terms of recall, 17.8 percentage points in terms of F1-score and 11.4 percentage points in terms of G-Mean. The results indicate the complementarity of the contributions of TMSC for characterization of a majority distribution and MKDNM for the selection of an informative majority sample. The parameter-sensitivity range is spanned by the tested values of h and K, whereas the value h=1.5 and the value K=7 are the values for the reference operating point from which the reported results are obtained. Because numerical performance levels are not reported separately for each tested h and K setting, it is not advisable to characterize these reference values as "empirically optimal". Rather they are regarded as absolute reference conditions for the comparative and ablation evaluations.

5.4 Computational Workflow Visualization

The class-distribution transformation observed by the proposed TMSC–MKDNM procedure is presented in Fig. 5. Initially, there are 10,000 records in the complete configuration, with 9,000 being the majority class and 1,000 being the minority class, which gives an imbalance ratio of 9:1. Records are initially split 80:20 before any clustering or sample selection is done, into development and held-out evaluation partitions.

The 8,000 records in the development partition consist of 7,200 majority and 800 minority records and preserve the original 9:1 ratio. The majority observations in this partition are analysed by TMSC to infer properties of dense, sparse, redundant, and boundary regions. The resulting structural information is then used by MKDNM to select the majority in the neighborhood. The reference configuration (h=1.5 and K=7) is used to remove 6,000 majority records and retain 1,200 informative majority records. During the sampling process all 800 minority records are saved.

After TMSC–MKDNM processing, the training representation consists of 2000 records, which include 1200 majority and 800 minority samples. This changes the development class ratio from 9:1 to 1.5:1, which represents an 83.3% decrease in majority records and no change in the minority records. This is a controlled transformation that will decrease the majority redundancy but will not require a 1:1 distribution.

Importantly, the held-out evaluation partition, which has 1,800 majority records and 200 minority records, has not been affected in any way by TMSC and MKDNM and is still in its original 9:1 ratio. Only used for final assessment of performance. This separation allows for a sample selected only to be used for model development, and avoids the evaluation information to impact the balancing process. So, Fig. 5 is a summary of the class-distribution transformation and the leakage-free computational workflow for assessing the proposed framework.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 2
Article Received: *05 July 2026* ***Revised:*** *22 July 2026* ***Accepted:*** *20 August 2026*

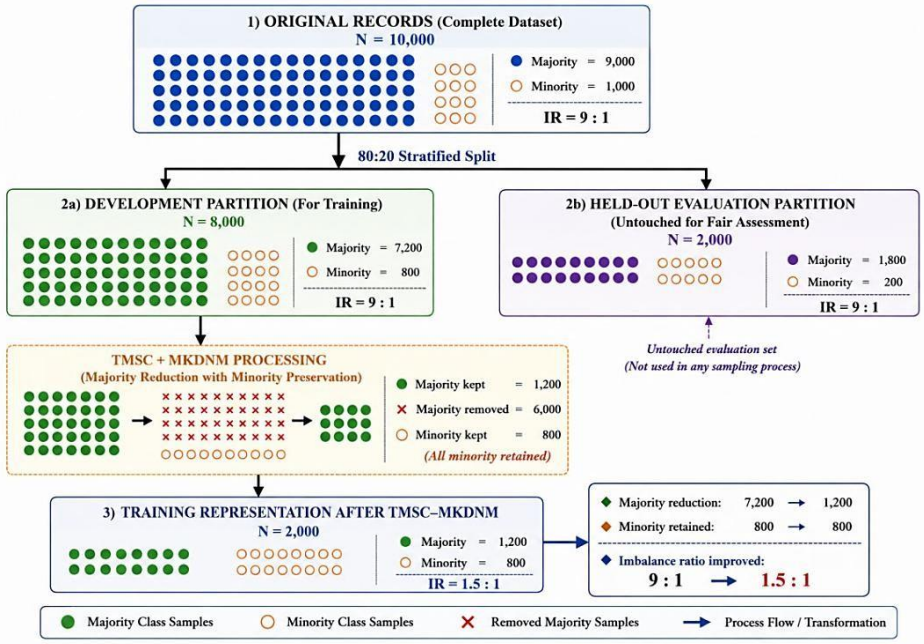

Figure 5. Class Distribution Transformation.

6. Conclusion

In this work, the TMSC–MKDNM–GRU structure for addressing this class-imbalance problem in computational medical-data classification by using structure-aware majority reduction and GRU-based classification are presented. It was a fully configured model, using 10,000 data points, 20 features and a class imbalance ratio of 9:1 in the initial data set. After the stratified split, TMSC and MKDNM were applied to the development partition only, which contained 7,200 majority records and 800 minority ones, but was later reduced to 1,200 majority and 800 minority records, resulting in a training ratio of 1.5:1. The 1,800:200 evaluation partition was not changed. The reported results of TMSC-MKDNM-GRU were obtained with a configuration that yielded 96.4% accuracy, 95.1% precision, 96.2% recall, 95.6% F1-score, 96.3% G-Mean, and 0.982 ROC-AUC. The proposed framework achieved higher minority recall (21.6 percentage points) and F1-score (17.8 percentage points) than conventional GRU. The results of the ablation analysis also suggested that both TMSC and MKDNM provide complementary contribution in the majority structure characterization and neighborhood aware sample selection processes respectively. There was also minimal variation in performance across the range of minority proportion (10–30%), with an F1-score that remained at 95.6–96.8% throughout. Future work involves examining adaptive selection of the bandwidth of TMSC and neighborhood MKDNM size; examining multiclass imbalance scenarios; examining alternative recurrent and attention-based classifiers; and examining wider computational configurations of higher dimensional feature space. Before deploying in practice or clinical applications, there should be further testing for statistical significance in repeated runs, computational complexity, and for model generalization in medical classification tasks using independently configured scenarios.

References

1. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *Int. J. Advance Soft Compu. Appl*, *5*(3), 176-204.
2. Araf, I., Idri, A., & Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review: I. Araf et al. *Artificial Intelligence Review*, *57*(4), 80.
3. Bates, M. P. (2024). Visualizing deep learning decisions: grad-CAM-based explainable AI for medical image analysis. *Journal of Scalable Data Engineering and Intelligent Computing*, 28-35.
4. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259.

5. Freire, G. F. (2025). Federated Learning Models for Privacy-Preserving Healthcare Data Analysis. *Journal of Wireless Intelligence and Spectrum Engineering*, 19-25.
6. Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., ... & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. *Ieee Access*, *9*, 109960-109975.
7. LemaÃŽtre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of machine learning research*, *18*(17), 1-5.
8. Martinez, D. (2026). Hybrid Time–Frequency and Machine Learning Approach for Intelligent Image and Speech Signal Analysis. *Transactions on Advanced Signal Processing and Analytics*, 20-26.
9. Naseriparsa, M., Al-Shammari, A., Sheng, M., Zhang, Y., & Zhou, R. (2020). RSMOTE: improving classification performance over imbalanced medical datasets. *Health information science and systems*, *8*(1), 22.
10. Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: review of a decade of research: M. Salmi et al. *Artificial intelligence review*, *57*(10), 273.
11. Siddavatam, K. I., & Shinde, S. K. (2026). A hybrid literature review on handling imbalanced medical data: AI models and open issues. *Expert Systems with Applications*, *296*, 129004.
12. Surendar, A. (2025). Intelligent Wearable Devices for Early Detection and Support of Speech and Swallowing Disorders. *Journal of Intelligent Assistive Communication Technologies*, 1-11.
13. Uvarajan, K. P. (2025). Machine learning-based EEG analysis for early detection of Alzheimer's disease in aging populations. *Frontiers in Life Sciences Research*, 38-43.
14. Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, *11*(1), 7.
15. Zhao, Y., Wong, Z. S. Y., & Tsui, K. L. (2018). A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *Journal of healthcare engineering*, *2018*(1), 6275435.