Journal of Theoretical and Applied Information Technology
15th March 2024. Vol.102. No 5

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

AN EFFICIENT ANOMALY DETECTION BASED HEMODIALYSIS MORTALITY PREDICTION FOR MIXED HEMODIALYSIS DATABASES

T HEMALATHA[1], K.V.D KIRAN[2]

[1]T Hemalatha, Research scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[2]Professor Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

ABSTRACT

As the size of mixed databases increases, challenges arise in improving the true positive rate. Several critical issues such as missing values, attribute noise, and imbalanced classes substantially impact data accuracy. High-quality input data becomes imperative to optimize classification algorithms, especially in the context of imbalanced mixed data types. Thus, the optimization of classic machine learning models becomes essential to ensure accurate predictions on imbalanced datasets. This study addresses these challenges inherent in hemodialysis mixed datasets by proposing an enhanced ensemble classification model incorporating optimal filtering and classification algorithms. A novel framework is proposed to handle missing data with classes, feature ranking, and ensemble classification approaches to improve the true positive rate and error rate on imbalance hemodialysis databases. Experimental results have demonstrated that the proposed approach outperforms conventional techniques in terms of statistical metrics.

Keywords: *Imbalance Hemodialysis Dataset, Probabilistic Classification , Support Vector Machine, Ensemble Learning Model.*

.

1. INTRODUCTION

Data mining and machine learning heavily rely on classification, the process of assigning items to predetermined groups according to their characteristics. This essential activity is foundational to many real-world applications, including medical diagnosis, picture and text classification, among others. Unfortunately, unbalanced data makes classification difficult, and existing models aren't always accurate. One major reason for this is that imbalanced datasets often have under-represented and undervalued minority classes, which can negatively impact the accuracy of the model [1]. When the distribution of classes in a dataset is not uniform, a typical problem in real-world applications is class imbalance. In medical diagnosis data, for example, there can be far fewer cases of a particular condition compared to cases without the disease. This disparity in data distribution can cause biased findings, and current models may fail to account for the minority class, leading to poor performance [2]. Several approaches, such as cost-sensitive learning, oversampling, and undersampling, have been suggested to deal with imbalanced datasets. However, these approaches aren't without their flaws, and they frequently lead to overfitting or data loss. They are also computationally expensive and require a lot of resources. When one category has a disproportionately high number of instances compared to another, we say that there is a class imbalance. This kind of skewed hemodialysis dataset is commonly used in real-world applications including medical diagnostics, credit scoring, and fraud detection [3-5]. However, in cases of class distribution imbalance, the numbers of instances in each class are not drastically different, but there is an uneven distribution of the classes. Overfitting, skewed results, and inaccurate results can result from using machine learning algorithms on imbalanced datasets. Subpar performance and skewed results are common outcomes when current algorithms fail to account for minority classes in imbalanced datasets [6]. One way to get the most out of machine learning models is to employ ensemble learning, which involves combining many models into one prediction. This method has

demonstrated promising results in actual applications, despite being computationally demanding and requiring large resources. To tackle imbalanced datasets, various strategies have been suggested, including deep learning and transfer learning, in addition to traditional machine learning techniques. These methods provide many approaches to deal with the problem of biased datasets in machine learning, and they've been successful in the real world. There are several oversampling methods to choose from, such as adaptive synthetic oversampling, random oversampling, and synthetic oversampling [7].

Both binary and multi-class imbalanced classification types can be effectively handled by resampling techniques, including oversampling and undersampling. The main benefit of these methods is their independence from the underlying classifier. According to the data, pre-processing is an effective strategy for achieving a more uniform distribution of classes in variables. The goal of random oversampling is to achieve a more equal distribution of classes by randomly replicating instances from the minority class. While this method is simple and easy to grasp, it may increase calculation time and lead to overfitting. Creating artificial members of the minority group using methods like bagging and bootstrapping is known as synthetic oversampling [8]. The task at hand and the properties of the data will dictate which of the two oversampling methods is most suited for analysis. However, it has been demonstrated that these resampling strategies effectively reduce class imbalance in ML classification models.Predictive model bias, overfitting, and metric misrepresentation are some of the challenges that can arise from hemodialysis dataset imbalances in real-time machine learning systems. Oversampling and undersampling are two sampling approaches that can be used to even out an unbalanced dataset's class distribution. However, these techniques also carry the risk of information loss and an increase in data noise [9]. Creating data samples of the minority class using synthetic methods, like SMOTE (Synthetic Minority Over-sampling Technique), is possible; nevertheless, there is a risk of overfitting and worse model performance. Careful analysis and resolution of these difficulties are critical for building and deploying machine learning models in real-time applications. These are among the most important problems with imbalanced datasets in machine learning systems that run in real-time.Important steps in preparing a dataset for ML include cleaning it up, eliminating noise, filling in missing values, and handling attribute and class noise. Converting data into a numerical representation, dealing with missing values, and scaling features to the same range are all possible steps in this process [10]. The term "noise filtering" refers to the steps used to eliminate or fix data instances that are inaccurate or inconsistent, such as outliers or duplicates. This step can be useful in improving the performance of machine learning systems and preventing overfitting. Incomplete or missing feature values in dataset instances might lead to missing values, while mistakes or discrepancies in dataset features are referred to as attribute noise.

This could be the result of features that are either inaccurately measured or reported or that include unnecessary details. Eliminating superfluous features, fixing data mistakes, or rescaling all features to the same value can all help reduce attribute noise. Traditional classifiers commonly misclassify minority class instances when faced with imbalanced learning, a phenomenon observed in many areas of computer science and finance [11]. This includes areas such as fraud detection, software defect prediction, credit scoring, and cancer malignancy grading.In binary class imbalance, there is a clear dominance of either the majority or the minority class. However, if the minority group is crucial to the prediction system, incorrect results or unreliable system performance may occur. Class imbalance is a typical issue in data mining, which can arise due to the ratio of the majority and minority classes. Incorrect categorization of uncommon occurrences, distinct behaviors, or anomalous trends can lead to negative outcomes. Attempts to artificially balance the data may exacerbate the problem rather than resolving it.New methods in healthcare, cybersecurity, IMS, and software defect prediction can learn from datasets and historical instances to forecast future outcomes, thereby enhancing system performance. Predictive systems have numerous potential applications in fields including healthcare, cybersecurity, and finance. Predictive analytics in healthcare, for instance, can utilize regular patient data such as blood pressure, heart rate, and blood sugar levels. Since producing reliable predictions is the primary objective of these systems, high-quality datasets are of paramount importance. Balanced and imbalanced datasets are terms used to describe dataset quality. Accurate forecasts and enhanced system performance result from a balanced dataset.Computer systems can now autonomously examine data using machine learning (ML) techniques. To maximize the efficiency of ML-based systems, enhancing the quality of datasets is

crucial [12]. Methods at the data level modify the distribution of classes by updating training datasets. The data level technique employs two methods for updating datasets: over-sampling and under-sampling. Under-sampling removes outliers from samples taken from the majority class to achieve statistical parity, while over-sampling evenly distributes outliers from the minority class by randomly duplicating them. Traditional clustering approaches cannot handle the massive amount of data due to their complexity and processing expense. The main goal is to increase the throughput of existing clustering techniques with minimal degradation in clustering quality.Because it is possible to achieve high accuracy numbers in these cases even if no instance of the minority class is correctly predicted, leading to the misclassification of important class occurrences, the accuracy measures used to evaluate classifiers fail in these situations. Conventional classifiers struggle to perform well on unbalanced data. To tackle this, a number of approaches and performance measures have been suggested. While there is a lack of data for model learning, ensemble approaches like boosting and bagging can improve the performance of poor classifiers. Data-level solutions, which often include resampling approaches, are frequently utilized to improve learning by balancing training datasets [13]. Such methods are flexible and not dependent on the classifier in use since they do not necessitate alterations to the algorithms. The one-class learning method relies on learning from a single class and can be more easily implemented with resampling techniques than the cost-sensitive learning method, which requires a cost matrix for different types of samples used in classification.This paper is organized as follows: A thorough analysis of the pertinent literature is presented in Section 2. The suggested method is described in depth in Section 3. Section 4 offers the results of the empirical evaluation of the proposed method. Finally, in Section 5, the report wraps up by summarizing the study's key findings and contributions.

Research Gaps:

1 One research gap in the field of software reliability estimation is the need to incorporate the dynamic nature of user experience and the learning effect over time. While existing models consider the fault detection rate and assume reliability growth with the fixing of underlying faults, they often overlook the impact of user familiarity and their ability to adapt to the software. As users gain experience, they tend to develop workarounds to handle situations that previously caused failures, resulting in an increase in reliability over time. However, current models do not explicitly capture this phenomenon.

2. Moreover, software reliability models make various assumptions related to fault detection rates, the location of faults in the software and data space, and the testing environment. These assumptions may not always hold true in practical scenarios, leading to limitations in the accuracy and applicability of the models.

3. Another research gap lies in the classification and characterization of software reliability models. While some models treat the software as a black box without considering its internal structure, others, known as white box models, explicitly incorporate the software's architecture and module interactions. There is a need to explore and develop more comprehensive and flexible models that can effectively handle diverse software systems and testing scenarios.

4. Additionally, there is a distinction between parametric and non-parametric models in terms of the interpretation and limitations of model parameters. Parametric models assume that the parameters have physical meanings and explicit ranges, while non-parametric models lack such restrictions. Exploring and comparing the strengths and weaknesses of these different modeling approaches can provide valuable insights for improving software reliability estimation.

2. RELATED WORK

To address the issue of class imbalance, ensembles prove to be effective. An ensemble is defined as a collection of classifiers that employ an aggregation approach to combine their judgments for classifying new input data [14–16]. The method involves running the basic learner multiple times with the training dataset to alter its distribution. The resulting base classifiers are then concatenated to create the final classifier for categorizing the testing dataset. Decision trees, neural networks, and naive Bayes are some of the alternatives for base learners.Differentiating small classes from the majority class becomes more challenging when there are highly distinct patterns within each class or when patterns overlap between classes, especially in cases where the classes are very small. In classes with a lot of overlap, there may be a significant decrease in the proportion of minority class samples correctly identified. A person's sensitivity to class imbalance increases in direct correlation with the degree of conceptual complexity.The K-Nearest Neighbor (KNN) method is an example of a lazy learner; it predicts output based on training cases but discards the

abstraction it learned from. However, if there is bias in the training data, instances from minority classes may be over- or under-classified due to their rarity.Vapnik's Support Vector Machines (SVMs) are among the most popular binary classifiers, based on the concept of maximum margin. Originally, they aimed to find the optimal separation hyperplane with the largest margin in linearly separable two-class issues involving margin. By applying nonlinear functions to nonlinearly separable issues, we can transform them into high-dimensional feature spaces, allowing linearly separable support vector machines to be used in nonlinear scenarios [17].

For the purpose of classifying instances into a small number of classes, decision trees utilize a simple knowledge representation. The tree consists of nodes representing features, edges connecting those features to their values, and leaves representing class labels. It becomes easy to categorize test instances after constructing a decision tree. Modeling a decision tree classifier involves two steps: constructing the tree and pruning it. In the first phase, the training set is partitioned repeatedly using locally optimal criteria until every partition contains tuples with the same class label. After building the decision trees, the next step involves pruning them to prevent overfitting of the training dataset. Pruning is performed based on the forecast of a lower predicted error rate for hemodialysis when a subtree is replaced with a tree leaf or its most commonly utilized branch [18].

Decision trees require a battery of tests to differentiate between majority and minority classes when the data is skewed toward one or the other. After analyzing the training samples covered by each leaf's class label, a decision tree is constructed, and the class with the highest weight is selected. Undersampling was found to be more effective than bagging in resolving binary class imbalance. Accurate predictions are achieved by using the Data Level approach and improving the balance of datasets by class. Prediction accuracy is enhanced through feature extraction.

The EUSBoost ensemble classifier has been shown to be the most successful approach in rectifying class imbalance in medical diagnostics, including malignant breast cancer diagnosis. The severity of class imbalance, dataset complexity, training data amount, and hemodialysis classifiers all contribute to the binary class imbalance problem.

The concept behind ensemble learning is to construct numerous classifiers using the training dataset, which are then combined to produce a final judgment. It is understood that all classifiers will make mistakes because they are trained on different sets of data. However, it's not necessarily the same samples of patterns that different classifiers get wrong. Classifiers can be broken down into their intrinsic error, bias error, and variance using the bias-variance decomposition. Furthermore, the issue of merging redundant ensembles is studied in connection with the notions of bias and variance.

In practical settings, managing several classes with unequal distributions poses difficulties. In scenarios with more than two classes, the procedures used to address two-class imbalance won't work. Data-level solutions that alter the class size ratio of two classes by under- or oversampling the majority class and then iteratively running the classification algorithm to discover the optimal distribution don't work when there are more than two classes. Algorithmic-level solutions are employed to make algorithms more biased in favor of the minority class. However, adapting the algorithm becomes more complicated when dealing with many minority classes [19].

While undersampling could enhance classifier sensitivity, it could lead to data loss. Therefore, the study's goals and the dataset's characteristics should be carefully considered when choosing the optimal method. Classification modeling, by distinguishing between common and rare samples, is a practical way to handle the issue of uneven class distribution. A critical factor in determining the model's ability to deal with class imbalance is the sample size. Expanding the training set to a larger size can greatly enhance the model's accuracy when looking for patterns in the tiny class of hemodialysis [20].

Classifiers are already complicated, and within-class imbalance adds further challenges, as a single class may have subclasses or subconcepts with different numbers of examples. Decision trees typically examine the class label associated with a leaf by analyzing the training instances it covered and selecting the most common class. Modifying decision trees may be necessary to increase the accuracy of the categorization model. Ensuring that smaller classes aren't excluded from the learning process should be a top priority: developing assessments that can distinguish between large and small classes. Refining the model by removing a branch that predicts a minor

class is akin to carefully trimming a tree; it produces a healthy leaf node with a distinct label.

One useful tool for dealing with classification problems is the multi-layer perceptron (MLP) trained using the backpropagation (BP) approach. The exceptional classification and categorization capabilities of the MLP result from its complex neural network, composed of interconnected layers of neurons. Each neuron in this network is an integral component, linked to every other neuron in the layers above and below it. While some neurons in the input layer may not activate, others are ready to take action [21].

Feature selection approaches are necessary to handle imbalanced datasets [22], as imbalance problems often involve high-dimensional data. Instead of using independent hemodialysis features, highly associated features should be explored when adding features to imbalanced datasets [23]. To handle small samples of imbalanced datasets, a novel feature selection metric called Feature Assessment by Sliding Threshold (FAST) was devised [24]. An imbalance in the relationship between feature distributions based on the probability density function can be addressed by unsupervised feature selection based on the filtering technique [25]. Iterative feature selection using different sample datasets enables the finding and ranking of effective feature lists [26]. Optimal Feature Weighting (OFW) and one-versus-one support vector machines (SVMs) are two stochastic techniques that can be employed to extract optimal features from a high-dimensional, imbalanced feature space [27].

Cost sensitivity serves as a feature weight, determining which characteristics are included in the Chaos Algorithm technique's feature selection [28]. The characteristics of the experimental datasets dictate the best feature selection method [4]. Under the data level method, training datasets can be updated using sampling approaches. To achieve statistical parity, under-sampling removes outliers from samples taken from the majority class, while over-sampling randomly duplicates the minority class. If better results are desired compared to utilizing only one sampling method, combining algorithms and sampling strategies is recommended [29]. However, because this method uses original datasets, the training dataset could be biased, and the subset of features that emerges could be inappropriate for different samples [30].Adaptive semi-unsupervised weighted oversampling is one method for sampling imbalanced datasets with overlapping classes [31]. Another approach to handling overlaps in unbalanced datasets using evolutionary fuzzy systems was suggested in [32] as feature weighting. To avoid overfitting problems, another study suggested using an oversampling method with rejection to create synthetic data for the minority class [33]. Addressing the issue of class imbalance, one solution was proposed in [34] by combining various ensemble approaches. A new strategy for unbalanced biological dataset classification was suggested in [35] with an ensemble of sampling and two classification techniques (SMOTE, Rotation Forest, and AdaBoost) [36].

Problem Statement:

Imbalanced mixed datasets, such as those encountered in hemodialysis databases, pose significant challenges for accurate predictive modeling due to missing values, attribute noise, and class imbalance. Conventional techniques often struggle to effectively handle these issues, leading to suboptimal performance in classification tasks. There is a pressing need for robust frameworks that can address these challenges and improve the accuracy of predictive models in the context of imbalanced mixed datasets.

Research Objectives and contributions

To address the challenges inherent in handling imbalanced mixed datasets, particularly in the context of hemodialysis datasets.

To propose an enhanced ensemble classification model incorporating optimal filtering and classification algorithms.

To evaluate the effectiveness of the proposed framework in improving the true positive rate and error rate on imbalanced hemodialysis databases.

To compare the performance of the proposed framework with conventional techniques using statistical metrics such as accuracy, precision, recall, and F1 score.

The proposed framework offers a systematic approach to handle missing data, perform feature ranking, and apply ensemble classification techniques in the context of imbalanced mixed datasets.

By incorporating optimal filtering and classification algorithms, the framework aims to optimize accuracy and reduce error rates in predictive modeling, particularly for hemodialysis datasets.

The experimental evaluation provides empirical evidence of the effectiveness of the proposed approach in outperforming conventional techniques, highlighting its potential for improving classification accuracy in medical diagnostics and other fields with imbalanced datasets.

3. PROPOSED MODEL

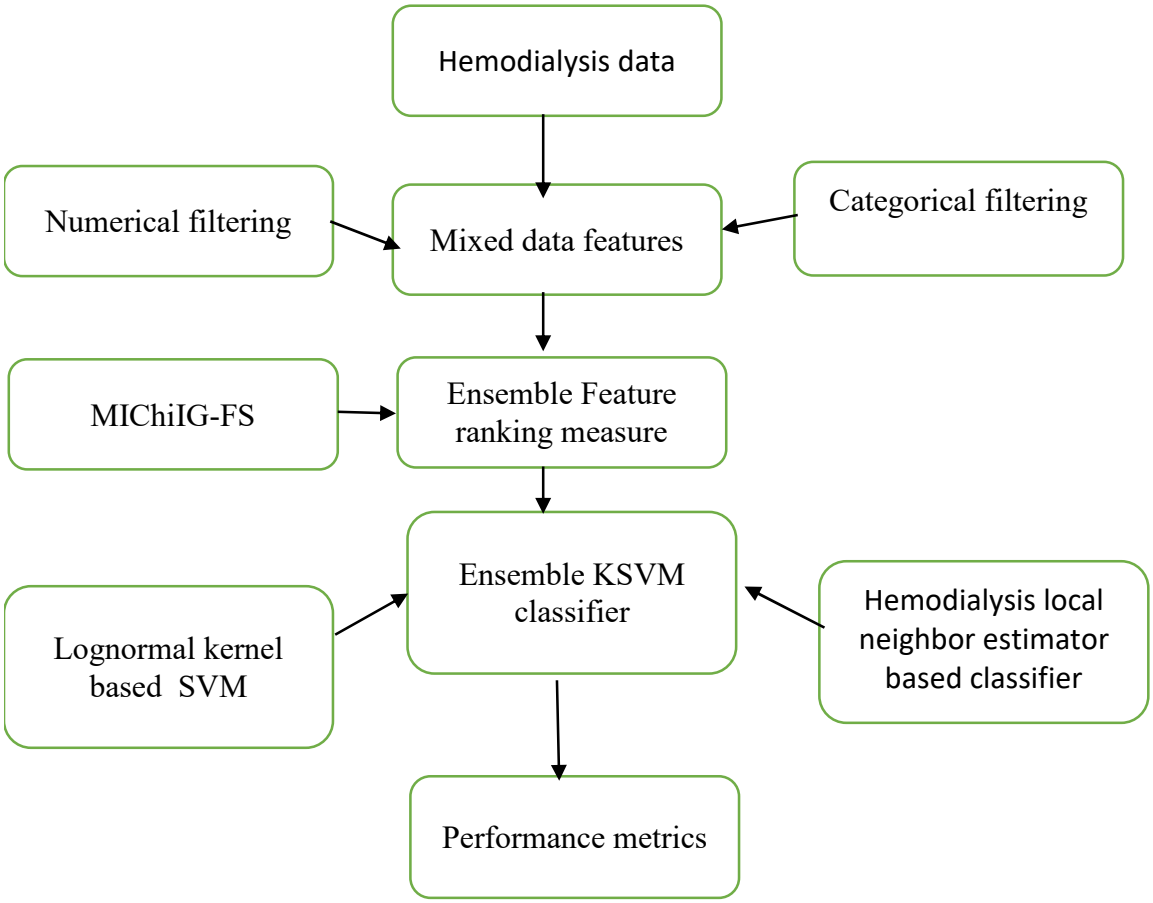

Figure 1: Proposed Model

The framework is designed to handle hemodialysis datasets containing both nominal and numeric attributes. It commences with pre-processing steps, including data filtering to address missing values and outliers. Attribute importance is then evaluated using ensemble feature ranking measures, which combine various ranking techniques for comprehensive analysis. The subsequent phase applies ensemble learning strategies that utilize the results of the ranking process. For classification tasks, the framework employs optimal support vector machines (SVM) and density probability function-based k-nearest neighbors (KNN) algorithms, leveraging SVM's capability to find optimal hyperplanes and KNN's density-based predictions. The models' performance is evaluated using statistical classification metrics, considering factors such as accuracy, precision, recall, F1 score, and other metrics. Overall, this framework offers an organized and robust approach

to handling mixed datasets, encompassing effective preprocessing, feature ranking, ensemble learning, and accurate classification utilizing SVM, density probability-based KNN, and statistical metrics for evaluation as shown in Figure 1.

Filtering approach :

Input: Training dataset with missing values
Separate the dataset into two subsets:
- Training data: The subset of data without any missing values.
- Missing data: The subset of data with missing values.
For each nominal attribute in the dataset:
Calculate the mode (most frequent value) of the attribute from the training data.
Replace missing values in the nominal attribute of the missing data subset with the calculated mode.
For each numeric attribute in the dataset:
Calculate the mean of the attribute from the training data.
Replace missing values in the numeric attribute of the missing data subset with the calculated mean.
Class missing value prediction using eq (1)

$$P\left(z_i = k \mid x_i, \theta^{(t)}\right) = \frac{P\left(x_i \mid z_i = k, \theta^{(t)}\right) P\left(z_i =\right.}{P\left(x_i \mid \theta^{(t)}\right)}$$

Output: The dataset with missing values replaced using modes for nominal attributes and means for numeric attributes.

The missing values in the numeric attribute of the missing data subset are replaced with the calculated mean. This approach uses the average value of the attribute to fill in the missing values, providing a reasonable estimate.Finally, the paragraph mentions "Class missing value prediction using eq (1)," which suggests that a specific equation or algorithm is used to predict missing values in the class variable (the target variable to be predicted).

2.Feature ranking measures

Feature ranking is an essential step in developing effective machine learning models for imbalanced datasets. To enhance model performance and avoid overfitting, selecting the most relevant features for imbalanced datasets is crucial, especially when the majority and minority classes are strongly imbalanced.
A variety of feature ranking measures can be applied to imbalance datasets, including:
ReliefF: This feature selection algorithm assesses the differences between the closest instances of the same and other classes when determining the significance of each feature. It is useful in identifying the features that most effectively differentiate the majority class from the minority class.
Gini Index: The Gini index measures the probability of an instance being incorrectly classified based on the distribution of classes in a node. It can help identify the features that provide the most information for classification and is commonly used in decision trees.
Chi-squared test: This analysis determines whether two categorical variables are independent of each other. It can be employed to identify the features that significantly influence the class variable.
Mutual Information (MI): MI measures the amount of information a feature contributes to the class variable. It is frequently used in feature selection algorithms to identify the most informative features for classification.

Ensemble Feature ranking algorithm

Function EnsembleFeatureRanking(Dataset, Class, Corr_threshold)
Inputs:
Dataset: the imbalanced dataset
Class: the target variable
Corr_threshold: the correlation threshold for identifying redundant features
Outputs:
Important_features: the subset of important features
Step 1: Calculate the Chi-Squared statistic for each feature:
For each feature X in Dataset:
Calculate the observed and expected frequencies
Calculate the Chi-Squared statistic for feature X
Identify the set of important features using a threshold or statistical test
Step 2: Calculate the Mutual Information for each feature:
For each feature X in Dataset:
Calculate the Mutual Information between feature X and the target variable Class
Identify the set of important features using a threshold or statistical test
Step 3: Calculate the Gain Ratio for each feature:
For each feature X in Dataset:
Calculate the entropy of the target variable Class
Calculate the entropy of feature X given Class
Calculate the Gain Ratio for feature X
Identify the set of important features using a threshold or statistical test

Step 4: Identify the set of highly correlated features:
Calculate the correlation matrix between all pairs of features
Identify pairs of features with a correlation greater than Corr_threshold
Create a set of highly correlated features
Step 5: Combine the sets of important features from the Chi-Squared Test, Mutual Information, and Gain Ratio:
Create a set of important features from the intersection of the three feature ranking measures
Step 6: Select the set of highly correlated features from the set of important features:
Return the subset of important features.
End Function

The EnsembleFeatureRanking function is designed to identify and select the most important features from an imbalanced dataset for predictive modeling. It takes three inputs: the imbalanced dataset itself, denoted as Dataset, the target variable, denoted as Class, and a correlation threshold, denoted as Corr_threshold, for identifying redundant features. The function aims to output a subset of important features, referred to as Important_features.Firstly, it calculates the Chi-Squared statistic for each feature in the dataset. This involves computing observed and expected frequencies for each feature and deriving the Chi-Squared statistic. The set of important features is then identified based on a specified threshold or statistical test.Next, the function calculates the Mutual Information for each feature, measuring the dependency between each feature and the target variable Class. Similarly, it identifies the set of important features using a threshold or statistical test.In the third step, the function computes the Gain Ratio for each feature. This involves determining the entropy of the target variable Class and the entropy of each feature given Class, followed by calculating the Gain Ratio for each feature. Again, the set of important features is identified using a threshold or statistical test.In the fourth step, the function identifies highly correlated features by calculating the correlation matrix between all pairs of features. Features with a correlation greater than the specified Corr_threshold are considered highly correlated, and a set of highly correlated features is created.Subsequently, the function combines the sets of important features obtained from the Chi-Squared Test, Mutual Information, and Gain Ratio. It creates a unified set of important features from the intersection of these three feature ranking measures.Finally, the function selects the set of highly correlated features from the set of important features and returns this subset of important features as the output. Overall, the EnsembleFeatureRanking function systematically analyzes the dataset and selects the most relevant features for predictive modeling, considering both their individual statistical properties and their interrelationships.

3. Ensemble classification framework

SVM optimization with kernel function

1. Ensemble classification framework:
a) SVM optimization with kernel function:
Function SVM_Optimization_with_Kernel(D, sigma, n, c):
Inputs:
D: Number of dimensions
sigma: Parameter controlling the width of the radial basis function (RBF)
n: Degree vector specifying the degree of each variable in the polynomial
c: Coefficients vector for the polynomial function
Output:
Kernel function for SVM optimization
Step 1: Compute the radial basis function (RBF) term:
Compute P(x) = exp^(-(sigma^2)*(d1^2 + ... + dD^2)) where d1, ..., dD are input vectors
(Note: d1^2, ..., dD^2 represent the squared distances from the origin along each dimension)
Step 2: Compute the normalization constant term:
Compute B(d, n) = sqrt((2*sigma^2)^(c1+...+cD) / (c1! * ... * cD!) * d1^n1 * ... * dD^nD)
(Note: d1^n1, ..., dD^nD represent the degrees specified in the degree vector n)

Step 3: Compute the polynomial function term:
Compute A(c) = P(x) * B(d, n)
Step 4: Return the kernel function: ker_n(d) = A(c)
End Function

Radial Basis Function (RBF): The first term, represented by P(x), is a radial basis function (RBF) that measures the similarity between pairs of input vectors based on their distance from the origin. It decreases exponentially as the distance between vectors increases, and its width is controlled by the

parameter sigma, which balances the tradeoff between bias and variance in the SVM.
Normalization Constant: The second term, denoted as B(d, n), serves as a normalization constant ensuring that the kernel function is positive and finite. It is computed using the multinomial coefficient formula and is proportional to the number of possible monomials of degree n in D dimensions.
Polynomial Function: The third term, represented by A(c), is a polynomial function of the input vector x and the degree vector n. This term captures the non-linear interactions between the input features and enables the SVM to model complex decision boundaries.

b) Local Probability estimator based KNN

Function Improved_KNN(data_set, test_instance, k_neighbors):
Inputs:
data_set: The dataset containing training instances
test_instance: The sample for which the class is to be predicted
k_neighbors: The number of nearest neighbors to consider
Output:
Predicted_class: Predicted class label for the test_instance
Step 1: Compute the Euclidean distance between test_instance and each instance in the dataset:
For each instance p in data_set:
Calculate the Euclidean distance using the formula:
Euclidean_distance(p, test_instance) = sqrt(Σ((pi - pj)²))

Step 2: Calculate the log normal likelihood for the computed distances:
For each computed Euclidean distance:
Calculate the log normal likelihood using the formula:
Log_Normal_Likelihood(distance) = ln(1 / (sqrt(2π) * σ)) * e^(-0.5 * ((distance / σ)²))
where σ is the standard deviation of the distances

Step 3: Select the k-nearest neighbors of the test_instance based on the maximum Euclidean distance and log normal likelihood:
Sort the dataset based on the computed Euclidean distances and log normal likelihoods in descending order
Select the top k instances as the nearest neighbors

Step 4: Compute distance probabilities for the selected k-nearest neighbors:
For each neighbor in the k-neighbors:
Compute the distance probability using the formula:
Distance_Probability(neighbor) = (1 / √(2π)) * ∫ e^(-0.5 * ((distance / σ)²)) d(distance) / |N|,
where N is the total number of attributes

Step 5: Compute class membership probabilities for each class:
For each neighbor in the k-neighbors:
Compute the membership probabilities of each class using the classifier

Step 6: Assign the class label to the test_instance based on the highest probability:
Select the class with the highest probability as the predicted class label for the test_instance

Return the predicted class label

End Function

4.Experimental Analysis

This section presents experimental results to evaluate our proposed framework in terms of accuracy, precision, recall, and F-measure, specifically on classification tasks on different imbalance datasets. The chronic disease(hemodialysis) dataset serves as an invaluable tool for conducting comprehensive analysis, identifying risk factors, building predictive models, and gaining deeper insights into the intricacies of chronic kidney disease. With a total of 25 columns, the dataset comprises both nominal and numerical attributes. The nominal attributes encompass critical indicators such as hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia, and the overall classification of CKD. Additionally, they include the presence of abnormalities in red blood cells, pus cells, pus cell clusters, and bacteria. On the other hand, the numerical attributes provide quantitative measurements for various blood components, including urea, creatinine, sodium, potassium, hemoglobin, and cell counts. They also encompass essential factors such as age, blood pressure, and blood glucose levels. By harnessing this extensive repository of medical data, researchers and medical practitioners gain a

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

valuable resource to explore, comprehend, and manage chronic kidney disease effectively.

Table 1: Hemodialysis dataset with mixed longitudinal datatypes and sparse null values.

No.	1: id Nominal	2: age Nominal	3: bp Nominal	4: sg Nominal	5: al Nominal	6: su Nominal	7: rbc Nominal	8: pc Nominal	9: pcc Nominal	10: ba Nominal	11: bgr Nominal	12: bu Nominal	13: sc Nominal	14: sod Nominal	15: pot Nominal	16: hemo Nominal	17: pcv Nominal	18: wc Nominal	19: rc Nominal	20: htn Nomina
1	0	48	80	1.02	1	0		normal	notpre...	notpre...	121	36	1.2			15.4	44	7800	5.2	yes
2	1	7	50	1.02	4	0		normal	notpre...	notpre...		18	0.8			11.3	38	6000		no
3	2	62	80	1.01	2	3	normal	normal	notpre...	notpre...	423	53	1.8			9.6	31	7500		no
4	3	48	70	1.005	4	0	normal	abnormal	present	notpre...	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes
5	4	51	80	1.01	2	0	normal	normal	notpre...	notpre...	106	26	1.4			11.6	35	7300	4.6	no
6	5	60	90	1.015	3	0			notpre...	notpre...	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes
7	6	68	70	1.01	0	0		normal	notpre...	notpre...	100	54	24	104	4	12.4	36			no
8	7	24		1.015	2	4	normal	abnormal	notpre...	notpre...	410	31	1.1			12.4	44	6900	5	no
9	8	52	100	1.015	3	0	normal	abnormal	present	notpre...	138	60	1.9			10.8	33	9600	4	yes
10	9	53	90	1.02	2	0	abnormal	abnormal	present	notpre...	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes
11	10	50	60	1.01	2	4		abnormal	present	notpre...	490	55	4			9.4	28			yes
12	11	63	70	1.01	3	0	abnormal	abnormal	present	notpre...	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes
13	12	68	70	1.015	3	1		normal	present	notpre...	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes
14	13	68	70						notpre...	notpre...	98	86	4.6	135	3.4	9.8				yes
15	14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes
16	15	40	80	1.015	3	0		normal	notpre...	notpre...	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes
17	16	47	70	1.015	2	0		normal	notpre...	notpre...	99	46	2.2	138	4.1	12.6				no
18	17	47	80						notpre...	notpre...	114	87	5.2	139	3.7	12.1				yes
19	18	60	100	1.025	0	3		normal	notpre...	notpre...	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes
20	19	62	60	1.015	1	0		abnormal	present	notpre...	100	31	1.6			10.3	30	5300	3.7	yes
21	20	61	80	1.015	2	0	abnormal	abnormal	notpre...	notpre...	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes
22	21	60	90						notpre...	notpre...		180	76	4.5		10.9	32	6200	3.6	yes
23	22	48	80	1.025	4	0	normal	abnormal	notpre...	notpre...	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes
24	23	21	70	1.01	0	0		normal	notpre...	notpre...										no
25	24	42	100	1.015	4	0	normal	abnormal	notpre...	present		50	1.4	129	4	11.1	39	8300	4.6	yes
26	25	61	60	1.025	0	0		normal	notpre...	notpre...	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes
27	26	75	80	1.015	0	0		normal	notpre...	notpre...	156	45	2.4	140	3.4	11.6	35	10300	4	yes
28	27	69	70	1.01	3	4	normal	abnormal	notpre...	notpre...	264	87	2.7	130	4	12.5	37	9600	4.1	yes
29	28	75	70		1	3			notpre...	notpre...	123	31	1.4							no

Table 1 presents the hemodialysis dataset, which consists of mixed data types and missing values. Upon examination of the table, it becomes evident that many attributes exhibit sparse values. Both numerical and categorical attributes within the dataset have sparse null values, indicating the presence of missing data points in these variables.

Journal of Theoretical and Applied Information Technology
15th March 2024. Vol.102. No 5

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

Table 2: Hemodialysis Dataset With Mixed Longitudinal Datatypes And Filled Null Values.

No.	1: id Nominal	2: age Nominal	3: bp Nominal	4: sg Nominal	5: al Nominal	6: su Nominal	7: rbc Nominal	8: pcc Nominal	9: ba Nominal	10: bgr Nominal	11: bu Nominal	12: sc Nominal	13: sod Nominal	14: pot Nominal	15: hemo Nominal	16: pcv Nominal	17: wc Nominal	18: rc Nominal	19: htn Nominal	20: dm Nominal
1	0	48	80	1.02	1	0	normal	notpre...	notpre...	121	36	1.2	135	3.5	15.4	44	7800	5.2	yes	yes
2	1	7	50	1.02	4	0	normal	notpre...	notpre...	99	18	0.8	135	3.5	11.3	38	6000	5.2	no	no
3	2	62	80	1.01	2	3	normal	notpre...	notpre...	423	53	1.8	135	3.5	9.6	31	7500	5.2	no	yes
4	3	48	70	1.005	4	0	normal	present	notpre...	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no
5	4	51	80	1.01	2	0	normal	notpre...	notpre...	106	26	1.4	135	3.5	11.6	35	7300	4.6	no	no
6	5	60	90	1.015	3	0	normal	notpre...	notpre...	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes
7	6	68	70	1.01	0	0	normal	notpre...	notpre...	100	54	24	104	4	12.4	36	9800	5.2	no	no
8	7	24	80	1.015	2	4	normal	notpre...	notpre...	410	31	1.1	135	3.5	12.4	44	6900	5	no	yes
9	8	52	100	1.015	3	0	normal	present	notpre...	138	60	1.9	135	3.5	10.8	33	9600	4	yes	yes
10	9	53	90	1.02	2	0	abnormal	present	notpre...	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes
11	10	50	60	1.01	2	4	normal	present	notpre...	490	55	4	135	3.5	9.4	28	9800	5.2	yes	yes
12	11	63	70	1.01	3	0	abnormal	present	notpre...	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes
13	12	68	70	1.015	3	1	normal	present	notpre...	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes
14	13	68	70	1.02	0	0	normal	notpre...	notpre...	98	86	4.6	135	3.4	9.8	41	9800	5.2	yes	yes
15	14	68	80	1.01	3	2	normal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes
16	15	40	80	1.015	3	0	normal	notpre...	notpre...	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no
17	16	47	70	1.015	2	0	normal	notpre...	notpre...	99	46	2.2	138	4.1	12.6	41	9800	5.2	no	no
18	17	47	80	1.02	0	0	normal	notpre...	notpre...	114	87	5.2	139	3.7	12.1	41	9800	5.2	yes	no
19	18	60	100	1.025	0	3	normal	notpre...	notpre...	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes
20	19	62	60	1.015	1	0	normal	present	notpre...	100	31	1.6	135	3.5	10.3	30	5300	3.7	yes	no
21	20	61	80	1.015	2	0	abnormal	notpre...	notpre...	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes
22	21	60	90	1.02	0	0	normal	notpre...	notpre...	99	180	76	4.5	3.5	10.9	32	6200	3.6	yes	yes
23	22	48	80	1.025	4	0	normal	notpre...	notpre...	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no
24	23	21	70	1.01	0	0	normal	notpre...	notpre...	99	46	1.2	135	3.5	15	41	9800	5.2	no	no
25	24	42	100	1.015	4	0	normal	notpre...	present	99	50	1.4	129	4	11.1	39	8300	4.6	yes	no
26	25	61	60	1.025	0	0	normal	notpre...	notpre...	108	75	1.9	141	5.2	9.9	29	8400	3.7	yes	yes
27	26	75	80	1.015	0	0	normal	notpre...	notpre...	156	45	2.4	140	3.4	11.6	35	10300	4	yes	yes
28	27	69	70	1.01	3	4	normal	notpre...	notpre...	264	87	2.7	130	4	12.5	37	9600	4.1	yes	yes
29	28	75	70	1.02	1	3	normal	notpre...	notpre...	123	31	1.4	135	3.5	15	41	9800	5.2	no	yes

Table 2 provides an overview of the hemodialysis dataset after addressing sparse values through a data preprocessing approach. The approach involved filling the sparse values in both numerical and nominal attributes with the mean or mode of the respective attributes. Additionally, missing class values were addressed by utilizing a probabilistic measure to fill in the gaps. As a result of this preprocessing step, the dataset is now devoid of sparse values, ensuring a more complete and reliable representation of the hemodialysis data.

```
Correctly Classified Instances         394               98.5     %
Incorrectly Classified Instances         6                1.5     %
Kappa statistic                          0.9683
Mean absolute error                      0.017
Root mean squared error                  0.1141
Relative absolute error                  3.6247 %
Root relative squared error             23.5649 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.976     0.000     1.000       0.976   0.988      0.969    1.000     1.000     ckd
                 1.000     0.024     0.962       1.000   0.980      0.969    1.000     1.000     notckd
Weighted Avg.    0.985     0.009     0.986       0.985   0.985      0.969    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 244   6 |   a = ckd
   0 150 |   b = notckd
```

Figure 2: Proposed Ensemble Classification Model

The classification results for the given dataset are highly accurate, with 98.5% of instances correctly classified and only 1.5% misclassified. Figure 2, demonstrates the effectiveness of the classification model in accurately predicting the classes of the instances. Overall, these results highlight the effectiveness of the classification model in accurately predicting the classes of instances in the *Existing Ensemble Learning Model*

```
Correctly Classified Instances         383               95.75    %
Incorrectly Classified Instances        17                4.25    %
Kappa statistic                          0.9113
Mean absolute error                      0.042
Root mean squared error                  0.1759
Relative absolute error                  8.9607 %
Root relative squared error             36.3268 %
Total Number of Instances              400

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.932    0.000    1.000      0.932    0.965      0.915    1.000     1.000     ckd
                 1.000    0.068    0.898      1.000    0.946      0.915    1.000     1.000     notckd
Weighted Avg.    0.958    0.026    0.962      0.958    0.958      0.915    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 233  17 |   a = ckd
   0 150 |   b = notckd
```

Figure 3: Existing Ensemble Classification Model

The existing algorithm used for classification in this scenario achieved an overall accuracy of 95.75%, correctly classifying 383 instances out of 400. The algorithm's performance is evaluated using various metrics.

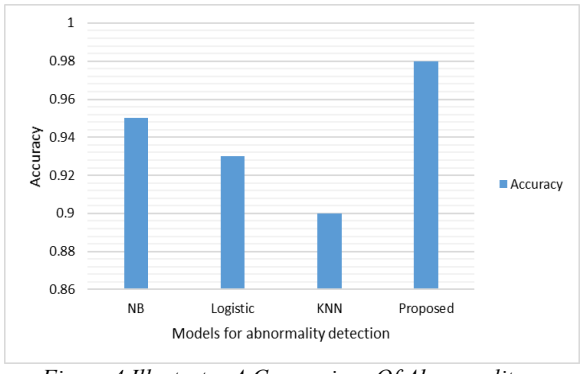

Figure 4 Illustrates A Comparison Of Abnormality Between The Proposed Model And Conventional Models(Accuracy).

The accuracy values represent the proportion of correctly classified instances by each model, reflecting their performance in predicting class labels accurately. A higher accuracy signifies superior performance in classification. According to the findings, the Proposed model exhibited the highest accuracy at 98%, indicating its superior performance in accurately classifying instances. Naive Bayes also demonstrated a commendable accuracy of 95%, closely followed by Logistic Regression with 93% accuracy. In comparison, K-Nearest Neighbors achieved the lowest accuracy among the evaluated models, reaching 90%.

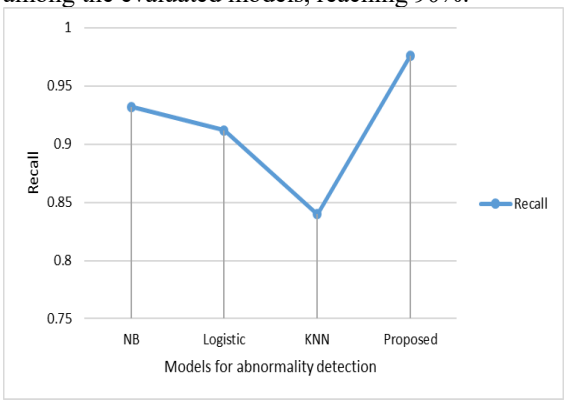

Figure 5: Illustrates A Comparison Of Abnormality Between The Proposed Model And Conventional Models(Recall)

hemodialysis dataset. The high accuracy, along with the detailed accuracy metrics and confusion matrix, demonstrate the model's ability to discriminate between the "ckd" and "notckd" classes with a high level of precision and recall. These findings provide valuable insights for understanding and managing chronic kidney disease based on the dataset.

Recall, also recognized as sensitivity or the true positive rate, quantifies the proportion of actual positive instances correctly identified by the model. Figure 5 highlights the model's capability to accurately detect positive instances. According to the provided recall values, the Proposed model attained the highest recall of 0.976, indicating its exceptional performance in correctly identifying positive instances. Naive Bayes also exhibited a high recall of 0.932, followed by Logistic Regression with a recall of 0.912. In contrast, K-Nearest Neighbors showed a relatively lower recall of 0.84 among the evaluated models.

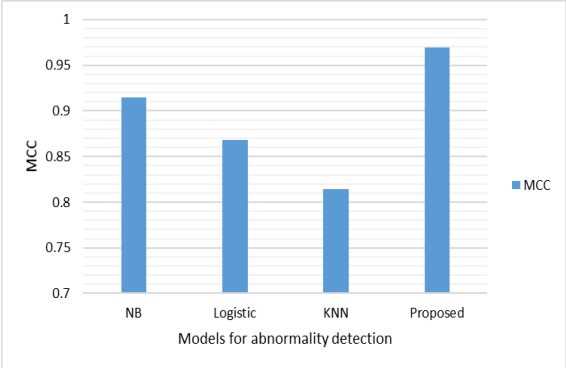

Figure 6: Illustrates A Comparison Of Abnormality Between The Proposed Model And Conventional Models(MCC)

The Matthews Correlation Coefficient (MCC) is a metric that synthesizes true positive, true negative, false positive, and false negative rates to offer a comprehensive evaluation of a model's performance. It considers the balance among these rates and yields a score ranging from -1 to +1, where +1 signifies perfect classification, 0 indicates random classification, and -1 implies inverse classification. Analyzing the provided MCC values, the Proposed model attained the highest MCC of 0.969, signifying outstanding performance in terms of overall classification quality. Naive Bayes also demonstrated strong performance with an MCC of 0.915, followed by Logistic Regression with an MCC of 0.868. In contrast, K-Nearest Neighbors displayed a lower MCC of 0.814 compared to the other models.

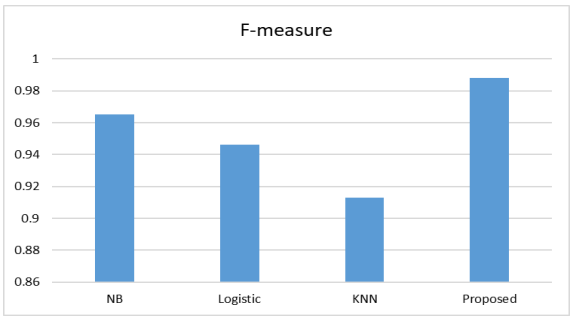

Figure 7: Illustrates A Comparison Of Abnormality Between The Proposed Model And Conventional Models(F-Measure)

The F-measure serves as a metric that harmonizes precision and recall to offer a well-rounded evaluation of a model's performance. It represents the harmonic mean of these two measures and provides an overall assessment of the model's capability to balance between accurately identifying positive instances (precision) and capturing all positive instances (recall). Assessing the provided F-measure values, the Proposed model attained the highest F-measure of 0.988, underscoring its exceptional ability to balance precision and recall in classifying instances effectively. Naive Bayes also showcased a commendable F-measure of 0.965, followed closely by Logistic Regression with an F-measure of 0.946. However, K-Nearest Neighbors displayed a comparatively lower F-measure of 0.913 among the evaluated models.

5. CONCLUSION

In this paper, an ensemble cluster-based classification model to address challenges associated with imbalanced datasets. In conclusion, the proposed framework presents a comprehensive solution for handling imbalanced mixed datasets, with a specific focus on hemodialysis databases. By incorporating optimal filtering, feature ranking, and ensemble classification techniques, the framework offers a systematic approach to address missing data, optimize feature selection, and improve classification accuracy. Experimental results demonstrate the superiority of the proposed approach over conventional techniques, underscoring its potential to enhance predictive modeling in medical diagnostics and other domains with imbalanced datasets. Moving forward, further research can explore extensions and refinements of the proposed framework to tackle additional challenges in predictive modeling and data analysis. The experiments conducted on imbalanced hemodialysis databases demonstrated that the proposed model outperforms conventional techniques in terms of statistical metrics such as

accuracy, recall, and the Matthews Correlation Coefficient. The results indicate that the proposed model achieved higher accuracy compared to other models, correctly classifying instances with an accuracy of 98%. It also demonstrated a high recall, correctly identifying positive instances with a recall of 0.976. The Matthews Correlation Coefficient further substantiates the excellent performance of the proposed model, with a value of 0.969.

REFERENCES

[1] T. K. J. Kam and C. T. Chan, "14 - Home Preparation and Installation for Home Hemodialysis," in Handbook of Dialysis Therapy (Sixth Edition), A. R. Nissenson, R. N. Fine, R. Mehrotra, and J. Zaritsky, Eds., New Delhi: Elsevier, 2023, pp. 149–153. doi: 10.1016/B978-0-323-79135-9.00014-8.

[2] "33 Distinct Subtypes of Hepatorenal Syndrome and Associated Outcomes as Identified by Machine Learning Consensus Clustering," American Journal of Kidney Diseases, vol. 81, no. 4, Supplement 1, p. S10, Apr. 2023, doi: 10.1053/j.ajkd.2023.01.035.

[3] "380 Animal-assisted Intervention for Hemodialysis Patients' Treatment Adherence, Pain, and Depression," American Journal of Kidney Diseases, vol. 81, no. 4, Supplement 1, p. S111, Apr. 2023, doi: 10.1053/j.ajkd.2023.01.382.

[4] "381 Experience of Starting a Rural Home Hemodialysis Program," American Journal of Kidney Diseases, vol. 81, no. 4, Supplement 1, p. S111, Apr. 2023, doi: 10.1053/j.ajkd.2023.01.383.

[5] "383 Associations of Skeletal Muscle Mass Index and Protein Markers in Patients With Stage 5 Chronic Kidney Disease Receiving Maintenance Hemodialysis," American Journal of Kidney Diseases, vol. 81, no. 4, Supplement 1, pp. S111–S112, Apr. 2023, doi: 10.1053/j.ajkd.2023.01.385.

[6] M. Asghari and S. M. J. Mirzapour Al-e-hashem, "A green delivery-pickup problem for home hemodialysis machines; sharing economy in distributing scarce resources," Transportation Research Part E: Logistics and Transportation Review, vol. 134, p. 101815, Feb. 2020, doi: 10.1016/j.tre.2019.11.009.

[7] T. Matsuzaki, Y. Kato, H. Mizoguchi, and K. Yamada, "A machine learning model that emulates experts' decision making in vancomycin initial dose planning," Journal of Pharmacological Sciences, vol. 148, no. 4, pp. 358–363, Apr. 2022, doi: 10.1016/j.jphs.2022.02.005.

[8] Y. Wang, Y. Zhu, G. Lou, P. Zhang, J. Chen, and J. Li, "A maintenance hemodialysis mortality prediction model based on anomaly detection using longitudinal hemodialysis data," Journal of Biomedical Informatics, vol. 123, p. 103930, Nov. 2021, doi: 10.1016/j.jbi.2021.103930.

[9] P. Kotanko and G. N. Nadkarni, "Advances in Chronic Kidney Disease Lead Editorial Outlining the Future of Artificial Intelligence/Machine Learning in Nephrology," Advances in Kidney Disease and Health, vol. 30, no. 1, pp. 2–3, Jan. 2023, doi: 10.1053/j.akdh.2022.11.008.

[10] J. Hu et al., "An effective model for predicting serum albumin level in hemodialysis patients," Computers in Biology and Medicine, vol. 140, p. 105054, Jan. 2022, doi: 10.1016/j.compbiomed.2021.105054.

[11] X. Yang et al., "An optimized machine learning framework for predicting intradialytic hypotension using indexes of chronic kidney disease-mineral and bone disorders," Computers in Biology and Medicine, vol. 145, p. 105510, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105510.

[12] Y. Liu, J. Xue, and J. Jiang, "Application of machine learning algorithms in electronic medical records to predict amputation-free survival after first revascularization in patients with peripheral artery disease," International Journal of Cardiology, Apr. 2023, doi: 10.1016/j.ijcard.2023.04.040.

[13] K. Hegerty et al., "Australian Workshops on Patients' Perspectives on Hemodialysis and Incremental Start," Kidney International Reports, vol. 8, no. 3, pp. 478–488, Mar. 2023, doi: 10.1016/j.ekir.2022.11.012.

[14] X. Yang et al., "Boosted machine learning model for predicting intradialytic hypotension using serum biomarkers of nutrition," Computers in Biology and Medicine, vol. 147, p. 105752, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105752.

[15] S. K. Dey, K. M. M. Uddin, H. Md. H. Babu, Md. M. Rahman, A. Howlader, and K. M. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," Intelligent Systems with Applications, vol. 16, p. 200144, Nov. 2022, doi: 10.1016/j.iswa.2022.200144.

[16] Z. Zheng, Q. H. Soomro, and D. M. Charytan, "Deep Learning Using Electrocardiograms in Patients on Maintenance Dialysis," Advances in Kidney Disease and Health, vol. 30, no. 1, pp. 61–68, Jan. 2023, doi: 10.1053/j.akdh.2022.11.009.

[17] K. Misumi et al., "Derivation and validation of a machine learning-based risk prediction model in patients with acute heart failure," Journal of Cardiology, vol. 81, no. 6, pp. 531–536, Jun. 2023, doi: 10.1016/j.jjcc.2023.02.006.

[18] T. Kalelioglu et al., "Detecting biomarkers associated with antipsychotic-induced extrapyramidal syndromes by using machine learning techniques," Journal of Psychiatric Research, vol. 158, pp. 300–304, Feb. 2023, doi: 10.1016/j.jpsychires.2023.01.003.

[19] T. Ferguson et al., "Development and External Validation of a Machine Learning Model for Progression of CKD," Kidney International Reports, vol. 7, no. 8, pp. 1772–1781, Aug. 2022, doi: 10.1016/j.ekir.2022.05.004.

[20] S. Sulaiman et al., "Development and Validation of a Machine Learning Score for Readmissions After Transcatheter Aortic Valve Implantation," JACC: Advances, vol. 1, no. 3, p. 100060, Aug. 2022, doi: 10.1016/j.jacadv.2022.100060.

[21] J. Yang et al., "Development of a machine learning model for the prediction of the short-term mortality in patients in the intensive care unit," Journal of Critical Care, vol. 71, p. 154106, Oct. 2022, doi: 10.1016/j.jcrc.2022.154106.

[22] A. V. Karhade et al., "Development of machine learning algorithms for prediction of mortality in spinal epidural abscess," The Spine Journal, vol. 19, no. 12, pp. 1950–1959, Dec. 2019, doi: 10.1016/j.spinee.2019.06.024.

[23] F. Miller et al., "Evaluation of a wearable biosensor to monitor potassium imbalance in patients receiving hemodialysis," Sensing and Bio-Sensing Research, vol. 40, p. 100561, Jun. 2023, doi: 10.1016/j.sbsr.2023.100561.

[24] M. Hecking, M. Madero, F. K. Port, D. Schneditz, P. Wabel, and C. Chazot, "Fluid volume management in hemodialysis: never give up!," Kidney International, vol. 103, no. 1, pp. 2–5, Jan. 2023, doi: 10.1016/j.kint.2022.09.021.

[25] Y. Komaru, T. Yoshida, Y. Hamasaki, M. Nangaku, and K. Doi, "Hierarchical Clustering Analysis for Predicting 1-Year Mortality After Starting Hemodialysis," Kidney International Reports, vol. 5, no. 8, pp. 1188–1195, Aug. 2020, doi: 10.1016/j.ekir.2020.05.007.

[26] S. D. Bieber and B. A. Young, "Home Hemodialysis: Core Curriculum 2021," American Journal of Kidney Diseases, vol. 78, no. 6, pp. 876–885, Dec. 2021, doi: 10.1053/j.ajkd.2021.01.025.

[27] W. F. Hussein, P. N. Bennett, and B. Schiller, "Innovations to Increase Home Hemodialysis Utilization: The Transitional Care Unit," Advances in Chronic Kidney Disease, vol. 28, no. 2, pp. 178–183, Mar. 2021, doi: 10.1053/j.ackd.2021.02.009.

[28] J. J. Squiers et al., "Machine learning analysis of multispectral imaging and clinical risk factors to predict amputation wound healing," Journal of Vascular Surgery, vol. 75, no. 1, pp. 279–285, Jan. 2022, doi: 10.1016/j.jvs.2021.06.478.

[29] S. Chaudhuri et al., "Machine learning directed interventions associate with decreased hospitalization rates in hemodialysis patients," International Journal of Medical Informatics, vol. 153, p. 104541, Sep. 2021, doi: 10.1016/j.ijmedinf.2021.104541.

[30] C. T. Ryan et al., "Machine learning for dynamic and early prediction of acute kidney injury after cardiac surgery," The Journal of Thoracic and Cardiovascular Surgery, Oct. 2022, doi: 10.1016/j.jtcvs.2022.09.045.

[31] A. J. Weiss et al., "Machine learning using institution-specific multi-modal electronic health records improves mortality risk prediction for cardiac surgery patients," JTCVS Open, Apr. 2023, doi: 10.1016/j.xjon.2023.03.010.

[32] A. M. McKnite, K. M. Job, R. Nelson, C. M. T. Sherwin, K. M. Watt, and S. C. Brewer, "Medication based machine learning to identify subpopulations of pediatric hemodialysis patients in an electronic health record database," Informatics in Medicine Unlocked, vol. 34, p. 101104, Jan. 2022, doi: 10.1016/j.imu.2022.101104.

[33] P. Escandell-Montero et al., "Optimization of anemia treatment in hemodialysis patients via reinforcement learning," Artificial Intelligence in Medicine, vol. 62, no. 1, pp. 47–60, Sep. 2014, doi: 10.1016/j.artmed.2014.07.004.

[34] K.-Y. Lin et al., "Optoelectronic online monitoring system for hemodialysis and its data analysis," Sensors and Actuators B:

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

Chemical, vol. 364, p. 131859, Aug. 2022, doi: 10.1016/j.snb.2022.131859.

[35] R. C. Walker, D. Tipene-Leach, A. Graham, and S. C. Palmer, "Patients' Experiences of Community House Hemodialysis: A Qualitative Study," Kidney Medicine, vol. 1, no. 6, pp. 338–346, Nov. 2019, doi: 10.1016/j.xkme.2019.07.010.

[36] M. Lavoie-Cardinal and A.-C. Nadeau-Fredette, "Physical Infrastructure and Integrated Governance Structure for Home Hemodialysis," Advances in Chronic Kidney Disease, vol. 28, no. 2, pp. 149–156, Mar. 2021, doi: 10.1053/j.ackd.2021.02.008.