Explainable Dynamic Ensemble Framework for Classification Based on the Late Fusion of Heterogeneous Multimodal Data

Firuz Juraev[1], Shaker El-Sappagh[1,2], and Tamer Abuhmed[1(✉)]

[1] College of Computing and Informatics, Sungkyunkwan University, Seoul, South Korea
fjuraev@g.skku.edu, shaker@skku.edu, tamer@skku.edu

[2] Faculty of Computer Science and Engineering, Galala University, Suez, Egypt

Abstract. Ensuring precise and reliable classification effectiveness holds paramount importance in essential sectors such as medicine, industry, and healthcare. Machine learning (ML) techniques have evolved in recent years to address the performance, efficiency, and robustness of the applied models. Recent advancements in Machine Learning (ML) techniques have aimed to enhance the performance, efficiency, and robustness of applied models. Ensemble learning has been shown to exhibit superior accuracy, robustness, and generalization capability over classical single ML models in most classification problems. Although dynamic ensembles have extended the performance of static ensembles such as random forest and boosting, current literature on dynamic ensembles primarily focuses on the early fusion of multimodal data. In this study, we present a novel framework that combines dynamic ensemble selection (DES) with a late fusion of heterogeneous multimodal data and model explainability. We evaluated our approach on a classification task of hospital mortality prediction, and our approach achieves a testing accuracy of 90.16%, surpassing existing techniques and providing physicians with case-based reasoning and deep-based classifiers contributions explanations to support their decision-making. We compare our proposed framework against nine widely used ML techniques, including static and dynamic ensemble models with early fusion and static ensemble models with late fusion on a dataset of 6,600 patients from MIT's GOSSIS dataset. The dynamic ensemble model with early fusion achieves a testing accuracy of 86.89%, the LightGBM model achieves a test accuracy of 87.72%, and the soft voting model reaches 87.97% and 89.45% using early and late fusion, respectively. Our proposed framework not only improves the accuracy and robustness of in-hospital mortality prediction models but also offers explainability and potential for further optimization to achieve even higher performance.

Keywords: Dynamic ensemble classifier · Multi-modality · Early fusion · Late fusion · Static ensemble classifier · Explainable ai · In-hospital mortality prediction

K. Arai (Ed.): IntelliSys 2023, LNNS 824, pp. 555–570, 2024.
https://doi.org/10.1007/978-3-031-47715-7_38

1 Introduction

Reliable prediction in a classification task is critical in ensuring precise decision-making, performance assessment, efficient resource allocation, and enhanced confidence in the entire system. This is particularly crucial in the realm of medical decision-making, where reliable prediction can potentially lower healthcare expenses and ease the burden on medical resources [14]. The literature has extensively investigated the use of various machine-learning approaches to improve prediction performance [6]. For example, ensemble learning techniques, such as static and dynamic ensembles, have been widely explored in the field of in-hospital mortality prediction [8,14], Sepsis prediction in intensive care [7], Alzheimer's disease progression [10] to improve classification accuracy. While static ensemble methods combine the predictions of all available base classifiers, dynamic ensemble selection chooses a subset of the most relevant classifiers for a given task, resulting in improved predictive performance and reduced computational cost [6,11].

The use of multimodal data in healthcare is prevalent due to the representation of numerous intricate medical problems through datasets comprising heterogeneous modalities. Different types of data such as medical images, clinical notes, and laboratory results are combined to improve diagnosis and treatment outcomes and to provide customized and individualized decisions. The most popular two approaches for combining multimodal data are early fusion and late fusion. Early fusion combines the raw data of different modalities at the feature level, while late fusion combines the outputs of classifiers trained on individual modalities at the decision level [9,26]. Early fusion has been demonstrated to be efficient when the modalities are highly correlated since the combining of raw data allows for more complete feature extraction and integration. On the other hand, late fusion is more suitable when the modalities have low correlation or requires different feature representations, as it allows for a more flexible combination of the outputs of individual classifiers [27]. Late fusion has also been shown to be effective when the modalities have varying amounts of missing data since it can handle missing data in particular modalities while improving overall performance [12].

In this work, we propose a novel framework that merges the advantages of dynamic ensemble modeling and the late fusion of data. We proposed a new architecture following a similar design of the existing dynamic ensemble model, i.e., the k-Nearest Oracle Union (KNORA-U) [15], with a novel late fusion capability. In addition, to the best of our knowledge, there is no work that provides explainability to dynamic ensemble models. In this work, we proposed two different explainability techniques for our proposed dynamic ensemble model with late fusion. The contributions of the study can be summarized as follows.

1. We propose the first dynamic ensemble selection classifier with a late fusion of multiple modalities. The proposed model is capable of learning complex tasks based on heterogeneous multi-modalities.

2. We propose a novel explainability for dynamic ensemble selection classifiers based on Case-Based Reasoning and deep-based classifiers contributions techniques.

3. We compare our proposed models' performance with existing approaches including the static ensemble model with early and late fusion, and the dynamic ensemble model with early fusion. The comparison is based on a real-world dataset of 6,600 patients from MIT's GOSSIS (Global Open Source Severity of Illness Score) dataset.

The study is organized as follows. Section 2 highlights the related work. Section 3 presents the proposed model, Sect. 4 discusses the results, and Sect. 5 introduces explainability. Section 6 provides discussion for our work and then Sect. 7 concludes the paper.

2 Related Work

Ensemble learning has emerged as a highly effective technique in the field of machine learning and pattern recognition. It involves generating and combining multiple models to improve the accuracy, robustness, and generalizability of predictions. By leveraging the strengths of multiple models, ensemble learning techniques have consistently demonstrated their ability to improve performance, reduce overfitting, and provide more reliable and robust predictions than single models alone. This led to successfully applying ensemble learning in various domains, including computer vision, natural language processing, and medical diagnosis, among others [3]. The ensemble learning approach can be broadly categorized into two main types: static and dynamic ensemble learning.

2.1 Static Ensemble

Static ensemble learning combines a fixed set of models to generate diverse models that work collectively to produce a better prediction (see Fig. 1). This approach leverages methods such as bagging [2], boosting [25], and stacking [23]. Bagging (Bootstrap Aggregating) creates multiple training datasets through bootstrapping, training independent models on these datasets, and then combining their predictions by averaging or voting to reduce overfitting and improve generalizability [20]. Boosting, on the other hand, entails training weak models consecutively on the misclassified data points, with each new model focused on the regions where the prior models failed. The final prediction is obtained by combining these weak models through a weighted majority vote or a weighted sum, ultimately resulting in a strong model with improved accuracy and reduced bias. Popular boosting algorithms include AdaBoost, which adjusts the weights of misclassified instances at each iteration, and Gradient Boosting, which utilizes the gradients of the loss function to guide the addition of new weak models [22,25]. Stacking involves training several diverse base models, which could include a mix of machine learning algorithms such as linear regression, decision trees, support vector machines, or even deep learning models like neural networks. Once base

models are trained to generate predictions on the given dataset, instead of combining the predictions through simple methods like averaging or majority voting, stacking introduces a meta-model, which is trained on the predictions generated by the base models, effectively learning how to optimally combine their outputs to achieve better performance [22].

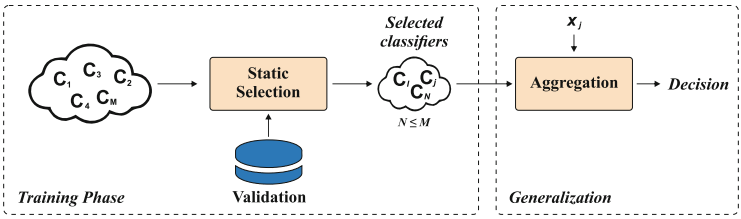


Fig. 1. Architecture of the static ensemble

2.2 Dynamic Ensemble

Dynamic ensemble learning constitutes an approach in which models are generated and combined in real-time based on specific query samples. This technique can be divided into two primary categories, namely dynamic classifier selection (DCS) and dynamic ensemble selection (DES), as illustrated in Fig. 2. DCS [21] is a technique that selects the most competent model from a pool of models, based on the characteristics of the query sample. This involves calculating the competences of each model in the pool and selecting the one that is best suited to make a final prediction for the given query. On the other hand, DES [15] is a technique that selects a subset of the most competent models from a pool of models and combines their predictions to make a final decision for the given query. This involves calculating the competences of each model in the pool and selecting the N most competent models to form the ensemble. The predictions of the selected models are then aggregated to make a final decision.

The literature has shown that dynamic ensemble learning outperforms static ensemble learning for several reasons. Dynamic ensemble learning is more robust and reliable because it can adapt to changes in data characteristics and distributions. It also reduces the risk of overfitting by selecting the most competent models for each query, and it produces better interpretable results by identifying the most relevant models and their contributions to the final prediction. These advantages make dynamic ensemble learning a powerful and effective approach for improving the accuracy and robustness of machine learning models [6].

2.3 Early Fusion-Based Ensemble

Early fusion-based ensemble modeling is a frequent data fusion technique where multiple modalities are integrated at an initial stage of the learning process

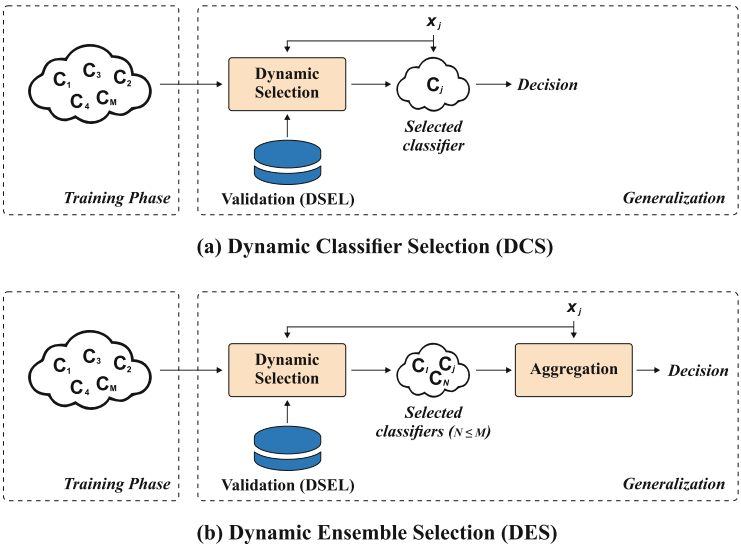


Fig. 2. The architecture of the dynamic classifier and dynamic ensemble selection

before training individual models within the pool. By combining various modalities during the early stages, this approach aims to capture the potential interaction and interdependencies among the modalities to enhance the model's overall performance [3]. However, a critical limitation of early fusion-based ensemble modeling is the training of base models within the pool using the entire dataset. This technique not only slows the training time for each model but also limits the diversity in the generated predictions. The diversity is reduced by the fact that all the models are trained on identical datasets, which may result in models with similar biases and limitations, thereby affecting the ensemble's overall robustness and generalization. Despite this drawback, early fusion-based ensemble modeling has been thoroughly investigated and employed in both static and dynamic ensemble modeling contexts within the literature. In the case of static ensemble modeling, early fusion has been used to create diverse models by combining multiple modalities, assuming that the data distribution remains consistent throughout the process. In contrast, dynamic ensemble modeling has incorporated early fusion to adaptively select and combine models in response to specific query samples, thereby leveraging the advantages of fusing multiple modalities while accounting for changing data characteristics and distributions [22].

Furthermore, researchers have also attempted to address the limitations of early fusion-based ensemble modeling by exploring alternative techniques such as late fusion and intermediate fusion, which integrate the modalities at different stages of the learning process. These alternative fusion strategies aim to balance the benefits of early integration with the need for maintaining diversity

and adaptability among the models within the ensemble, and offering potential alternatives for further enhancement in ensemble learning methods [5,17].

2.4 Late Fusion-Based Ensemble

Late fusion-based ensemble modeling is a widely adopted strategy that includes training separate models in various modalities and combinations of modalities and then integrating their output predictions at a later stage of the learning process. A key advantage of late fusion lies in its ability to permit models to learn tasks across multiple modalities, thereby fostering greater diversity in their decision boundaries. This diversity within the model pool is crucial for constructing a robust ensemble that can effectively generalize to previously unseen data. While late fusion has been extensively researched and applied for static ensemble learning [19], its potential for dynamic ensemble selection remains comparatively under-explored in the literature. We want to solve this research gap by examining the applicability of late fusion to dynamic ensemble selection models in this paper. By combining the output predictions of multiple dynamically selected models, we aim to improve the accuracy and robustness of the ensemble model, especially in scenarios where the data distribution may change over time. Another significant contribution of our work is the development of an innovative explainability technique for dynamic ensemble selection models, an aspect that has not been previously investigated. This novel approach aims to provide insights into the decision-making process of the ensemble, clarifying the contributions of individual models and the rationale behind the dynamic selection of models. Our proposed explainability technique can potentially facilitate the adoption of dynamic ensemble selection models in various domains, particularly those where interpretability is significantly important, such as healthcare and finance.

3 Methods

This section describes the medical dataset utilized in this study, the architecture of the suggested framework for predicting in-hospital mortality, and the procedures utilized for data preparation.

3.1 Dataset

In this study, we used MIT's GOSSIS (Global Open Source Severity of Illness Score) dataset [18]. The study includes 6600 subjects (57.3% male). The average age of subjects was between 16 and 89 years. There are two classes in the dataset whether the patient is dead or survived. There are 5264 (79.7%) subjects who survived and 1336 (20.3%) as dead patients.

3.2 Proposed Model

Figure 3 illustrates our proposed framework. This framework learns in-hospital mortality based on the multimodal late fusion paradigm. The proposed algorithm extends the KNORA-U (KNORA-UNION) and works as follows:

1. Split dataset D into training *Train*, validation (for DES) *Dsel*, and testing *Test*.

2. Define the models pool M with corresponding feature set F (modality or combination of modalities): $M = \{m_1, m_2, m_3, ..., m_n\}$, $F = \{f_1, f_2, f_3, ..., f_n\}$

3. Train models' pool M with their corresponding subsets of F masked on the *Train* set.

4. Define a number of neighbors K for selection.

5. Select K nearest samples from *Dsel* for the given x_i (test sample) using the k-Nearest Neighbors algorithm with Minkowski distance (Eq. 1). These K nearest samples are called regions of competence (RoC) for x_i.

$$d_M(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{1}$$

where p is a positive constant that determines the order of the distance and $y_i \in Dsel$.

6. Make predictions P of the samples in RoC using the trained models' pool M: $P = \{p_1, p_2, p_3, ..., p_n\}$ and p_i is predictions of m_i (where $m_i \in M$) for K samples in RoC.

7. Calculate the competence scores C of each model in the models' pool M using their prediction P in RoC. The competence score is the accuracy of the model in RoC: $C(m_i) = accuracy(p_i)$ where $m_i \in M, p_i \in P$.

8. Select models that have a competence score higher than zero. This means that the model predicted at least one sample in RoC correctly.

9. Predict the given test sample x_i with selected models.

10. Aggregate the predictions of selected models for x_i using soft voting and use competence scores as weights.

Our model also provides two types of explainability: case-based reasoning (CBR) and deep-based classifiers contribution (DBCC) to a final decision. For CBR, the model provides a map for sample x_i with the given test sample x_i, selected K nearest samples, and other samples on Dsel set. In addition, our model provides a detailed table of selected K nearest samples that help the physicians to get more insight into similar cases to the given test sample. For DBCC, our model provides an explanation of how each selected model contributes to the final decision.

3.3 Data Preprocessing

3.3.1 Missing Data Handling A feature that is absent more than 25% of the time gets eliminated. k-Nearest Neighbors (KNNImputer) [28] with 5 nearest neighbors is used to impute features with less than 25% missing. The mean value

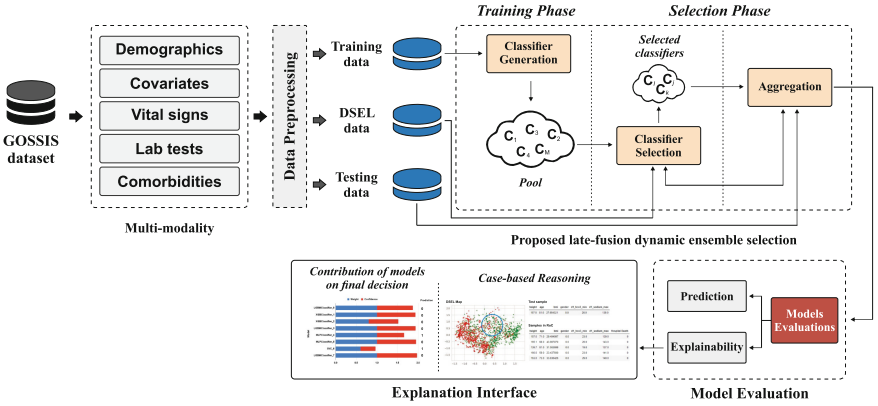


Fig. 3. The architecture of the proposed framework

from the five nearest neighbors discovered in the training set is used to impute the missing values in each sample.

3.3.2 Data Standardization We used the MinMax normalization approach to transform the data into a homogenous normal distribution, as described in Eq. 2.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

3.3.3 Feature Selection We employed a recursive feature elimination (RFE) approach [13] to select a subset of important features from a large set of 185 statistical features. To carry out the selection process, we utilized the Random Forest classifier in conjunction with RFE (RFE-RF). This method was used to rank the features based on their relevance and contribution to the classification task. After applying RFE-RF, we were able to identify 83 important features that were deemed essential for accurate classification.

3.3.4 Data Balancing Biased results are always the result of unbalanced datasets. In classification modeling, unbalanced datasets may be addressed using a variety of techniques, such as either oversampling the minority stage or undersampling the majority stage. In this work, although our dataset is not highly imbalanced, we applied the synthetic minority over-sampling technique (SMOTE) [4] for balancing our dataset.

4 Results

In this section, we discuss and compare the results of different machine-learning approaches with our proposed late-fusion dynamic ensemble selection model.

We collect and present the testing results for every model. In order to get more consistent results, we applied the 10-holdout testing method [24]. The results are shown as (mean ± standard deviation).

4.1 Early Fusion-Based Ensemble Modeling

In early fusion, we combine all modalities first and train machine learning (ML) models with them. For ensemble techniques, we need a pool of ML models. In this study, we used nine well-known and diverse ML algorithms [1,16]: XGboost (XGB), LightGBM (LGBM), Random Forest (RF), Support Vector Classifier (SVC), Logistic Regression (LR), Multi-Layer Perceptron (MLP), Decision Tree (DT), Naive Bayes (NB), and k-Nearest Neighbors (KNN).

First, Table 1 shows the performance of the nine selected models. Among the models, the LightGBM ensemble model achieved the highest accuracy of 87.72%. Following the LGBM, XGB, and RF models reached 86.97% and 81.35%, respectively. Please note that the ensemble models achieved better results than classical ML models.

Table 1. The performance of the base models with the early fusion of modalities

Base classifier	Accuracy	Precision	Recall	F1-score
LGBM	87.72±0.47	91.00±0.61	83.73±0.90	87.21±0.52
XGB	86.97±0.56	90.80±1.10	82.30±1.07	86.33±0.60
RF	81.35±1.11	86.21±0.73	74.63±2.29	79.99±1.43
SVC	80.72±1.14	81.32±0.85	79.76±2.51	80.52±1.38
LR	78.20±1.19	77.48±0.63	79.51±2.87	78.46±1.52
MLP	76.83±1.01	83.03±1.41	67.47±2.22	74.42±1.37
DT	73.11±1.27	77.39±1.42	65.31±1.78	70.83±1.49
NB	72.82±1.76	73.25±1.49	71.86±2.76	72.54±2.02
KNN	72.44±1.51	71.24±1.32	75.25±2.65	73.18±1.70

In Table 2, we can see the results of the static ensemble models with a different set of the model pool. The best accuracy of 87.97% is reached with the pool of the three most accurate models: LGBM, XGB, and RF. We also tried to add more diverse models, but the results do not improve. The training data of all models are the same and for that reason, there is no high diversity in predictions.

For testing the performance of the dynamic ensemble selection model with early fusion, we selected KNORA-U because we extended this technique for late fusion. The performance of the KNORA-U is not high as compared to the static ensemble model. The results of KNORA-U are shown in Table 3, and the highest accuracy that KNORA-U reached is 86.89% with the pool of the three most accurate models.

Table 2. The performance of the static ensemble model (soft voting) with the early fusion ofmModalities

Selection strategy	Model pool	Accuracy	Precision	Recall	F1-score
All models	XGB, RF, LGBM, SVC, MLP, LR, NB, KNN, DT	83.96±0.82	85.70±0.77	81.50±1.73	83.53±0.96
3 most accurate	**LGBM, XGB, RF**	87.97±0.54	90.95±0.64	84.35±0.93	87.50±0.60
4 most accurate	LGBM, XGB, RF, SVC	87.28±0.57	89.69±0.65	84.27±1.00	86.90±0.61
3 most accurate + 1 diverse	LGBM, XGB, RF, MLP	87.10±0.52	90.54±0.83	82.91±0.97	86.55±0.57
3 most accurate + 2 diverse	LGBM, XGB, RF, MLP, KNN	86.10±0.68	88.71±0.69	82.74±1.33	85.61±0.76
2 most accurate + 3 diverse	LGBM, XGB, MLP, KNN, DT	85.65±0.63	88.56±0.88	81.95±0.94	85.12±0.68

Table 3. The performance of the dynamic ensemble model (KNORAU) with the early fusion of modalities

Selection strategy	Model pool	Accuracy	Precision	Recall	F1-score
All models	XGB, RF, LGBM, SVC, MLP, LR, NB, KNN, DT	83.78±1.03	84.07±0.91	83.78±1.03	83.72±1.04
3 most accurate	LGBM, XGB, RF	86.89±0.96	87.16±0.88	**86.89±0.96**	**86.85±0.97**
4 most accurate	LGBM, XGB, RF, SVC	85.80±0.94	86.20±0.85	85.80±0.94	85.74±0.98
3 most accurate + 1 diverse	LGBM, XGB, RF, MLP	85.95±1.03	86.48±0.91	85.95±1.03	85.92±1.05
3 most accurate + 2 diverse	LGBM, XGB, RF, MLP, KNN	85.33±1.18	85.72±1.04	85.33±1.18	85.30±1.20
2 most accurate + 3 diverse	LGBM, XGB, MLP, KNN, DT	85.57±0.81	85.89±0.71	85.57±0.81	85.53±0.83

4.2 Late Fusion-Based Ensemble Modeling

In late fusion, instead of combining all modalities, we can train each model in the pool with different modalities and different combinations of modalities. This approach increases the diversity of models' prediction and gives better performance by ensembling than early fusion scenarios.

Table 4 shows the results of the static ensemble, and as expected the performance is increased as compared to the static ensemble with early fusion. We tested different combinations of models' pools and modalities. We got the best result with the following combinations: LGBM with lab tests (L), XGB with demographics (D), XGB with covariates (COV), LGBM with vital signs (V), MLP with commodities (COM), MLP with a combination of lab tests and demographics (L, D), SVC with a combination of covariates and demographics (COV, D), and RF with a combination of all modalities. With this setting, we achieved the best accuracy of 89.45%.

In Table 5, we report the results of our proposed model: dynamic ensemble selection model with late fusion. We tested different combinations with our model as well and to get the highest results, we utilized a model pool consisting of four highly accurate models (2 LightGBM and 2 XGBoost) and 4 diverse models (1 RF, 1 SVC, and 2 MLP). The average accuracy of these models was 77.75 ± 8.3 and we achieved an accuracy of 90.16% that outperforms all previous techniques. By leveraging the diversity of the model pool, we were able to achieve a high level of accuracy in the DES approach, which supports our hypothesis that diversity in the model pool is important in the dynamic ensemble techniques.

Table 4. The performance of the static ensemble model (soft voting) with the late fusion of modalities

Modalities	Strategy	Models	Accuracy	Precision	Recall	F1-score
[L] + [D] + [COV] + [V] + [ACOM]	5 accurate	LGBM, XGB, XGB, LGBM, LR	89.08±0.43	89.96±0.46	89.08±0.43	89.03±0.43
[L] + [D] + [COV] + [V]	4 accurate	LGBM, XGB, XGB, LGBM	88.98±0.46	89.88±0.49	88.98±0.46	88.95±0.45
[L] + [D] + [ACOV] + [V] + [ACOM]	4 accurate + 1 diverse	LGBM, XGB, XGB, LGBM, MLP	89.05±0.40	89.97±0.43	89.05±0.40	89.01±0.41
[L] + [D] + [COV] + [V] + [ACOM] + [COV]	4 accurate + 2 diverse	LGBM, XGB, XGB, LGBM, MLP, KNN	88.82±0.79	89.40±0.78	88.82±0.79	88.75±0.79
[L] + [D] + [COV] + [V] + [ACOM] + [L, D]	4 accurate + 2 diverse	LGBM, XGB, XGB, LGBM, MLP, MLP	89.22±0.61	89.76±0.58	89.22±0.61	89.17±0.60
[L] + [D] + [COV] + [V] + [COM] + [L, D] + [COV, D]	4 accurate + 3 diverse	LGBM, XGB, XGB, LGBM, MLP, MLP, SVC	89.15±0.81	89.45±0.79	89.15±0.81	89.12±0.81
[L] + [D] + [COV] + [V] + [COM] + [L, D] + [COV, D] + [ALL]	**4 accurate + 4 diverse**	**LGBM, XGB, XGB, LGBM, MLP, MLP, SVC, RF**	**89.45±0.81**	**89.76±0.82**	**89.45±0.81**	**89.43±0.80**

Table 5. The performance of the dynamic ensemble model (our Proposed) with the late fusion of modalities

Modalities	Strategy	Models	Accuracy	Precision	Recall	F1-score
[L] + [D] + [COV] + [V] + [COM]	5 accurate	LGBM, XGB, XGB, LGBM, LR	89.30±0.35	90.29±0.33	89.30±0.35	89.26±0.34
[L] + [D] + [COV] + [V]	4 accurate	LGBM, XGB, XGB, LGBM	89.22±0.31	90.21±0.28	89.22±0.31	89.16±0.32
[L] + [D] + [COV] + [V] + [COM]	4 accurate + 1 diverse	LGBM, XGB, XGB, LGBM, MLP	89.25±0.36	90.28±0.29	89.25±0.36	89.21±0.34
[L] + [D] + [COV] + [V] + [COM] + [COV]	4 accurate + 2 diverse	LGBM, XGB, XGB, LGBM, MLP, KNN	89.24±0.43	89.85±0.38	89.24±0.43	89.18±0.43
[L] + [D] + [COV] + [V] + [COM] + [L, D]	4 accurate + 2 diverse	LGBM, XGB, XGB, LGBM, MLP, MLP	89.66±0.34	90.18±0.33	89.66±0.34	89.63±0.33
[L] + [D] + [COV] + [V] + [COM] + [L, D] + [COV, D]	4 accurate + 3 diverse	LGBM, XGB, XGB, LGBM, MLP, MLP, SVC	89.72±0.43	90.01±0.40	89.72±0.43	89.68±0.43
[L] + [D] + [COV] + [V] + [COM] + [L, D] + [COV, D] + [ALL]	**4 accurate + 4 diverse**	**LGBM, XGB, XGB, LGBM, MLP, MLP, SVC, RF**	**90.16±0.57**	**90.42±0.53**	**90.16±0.57**	**90.14±0.58**

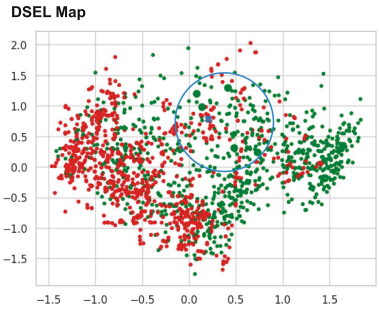
5 Model Explainability

Accurate models are not sufficient to get the trust of medical experts. Understanding of how the models are working and why they take specific decisions are important for trustworthy AI models. In previous experiments, we achieved the best results by integrating the dynamic ensemble selection and late fusion of heterogeneous data modalities. We extend our proposed late fusion dynamic ensemble selection model to provide explainability. Our model provides two types of explanation: case-based reasoning (CBR) and deep-based classifiers contributions (DBCC). In this section, these explanabilities are discussed with examples.

5.1 Case-Based Reasoning

In dynamic ensemble selection, we choose base classifiers for the final decision based on the performance of classifiers in the region of competence (RoC). We select the most similar samples as RoC for the given test sample as discussed in Sect. 3.2. If we provide those samples in RoC to physicians as similar cases to the given new test sample, the physicians can get more insight. It is also called case-based reasoning (CBR). The physicians can select the features that they want to see. In Fig. 4, you can see an example of case-based reasoning. On the left of Fig. 4, there is a DSEL map where we can see the given test sample (blue X), selected samples for RoC (bigger points), and other samples in DSEL data.

On the left of Fig. 4, there are tables that show the detailed data about the given test sample and each nearest sample in the RoC. In the table (on the right side of Fig. 4), the physicians can select which features they want to see or as default features, we show the most important features. Statistics can be collected from the collected cases, and domain experts can test if the neighbor cases are medically similar to the test case or not. Also, visualization techniques can be used to show the important features that affect the decisions of the classifier.



Test sample

height	age	bmi	gender	d1_hco3_min	d1_sodium_max
157.5	81.0	27.654321	0.0	26.0	139.0

Samples in RoC

height	age	bmi	gender	d1_hco3_min	d1_sodium_max	Hospital Death
157.0	71.0	29.494097	0.0	23.0	139.0	0
165.1	68.0	43.987070	0.0	20.0	143.0	0
139.7	81.0	31.563699	0.0	19.0	137.0	0
160.0	59.0	23.437500	0.0	23.0	141.0	0
163.0	73.0	33.836426	0.0	29.0	140.0	0

Fig. 4. Case-based reasoning. Samples that are selected in the DSEL map (left) and their detailed table with selected features (right)

5.2 Deep-Based Classifiers Contributions

In the previous explainability, we discussed the explanations suitable for physicians. Furthermore, we can also get explanations on a deeper level: how each classifier in the selected ensemble contributes to the final decision, what are their individual predictions, and what their confidence is. Our model provides the table as shown in Fig. 5 with the features of weight (competence score), individual prediction for the given test sample, and a confidence score for that prediction. With this table, we can generate figures as illustrated on the right side of Fig. 5. With this explanation, we can see how each model affects our result and it helps us to add or remove other models for better performance.

Fig. 5. Selected ensemble contribution. Our model has produced a table (on the left), which may be used to generate a corresponding figure (on the right)

6 Discussion

The primary objective of our paper is to present an innovative approach to dynamic ensemble selection (DES) that employs a late fusion strategy for the fusion of multi-modal data. Moreover, our approach introduces a novel method to enhance the explainability of dynamic ensemble selection and address the need for better transparency and interpretability in the decision-making process.

Our experimental findings indicate that the proposed DES method outperforms existing techniques, including static ensembles such as voting and stacking that employ both early and late fusion, as well as the dynamic ensemble selection approaches that utilize early fusion. The performance improvement of our proposed DES is related to the advantages of dynamic selection in comparison to static selection, in addition to the benefits of the late fusion of data over early fusion.

By effectively combining the strengths of dynamic ensemble selection with the advantages of late fusion, our proposed approach demonstrates its potential to improve the accuracy, robustness, and generalizability of machine learning models. Furthermore, the introduction of a novel explainability technique contributes to a better understanding of the ensemble's decision-making process, ultimately fostering increased confidence and trust in the model's predictions, particularly in domains where interpretability is of critical importance.

Limitations. Our study involved the evaluation of our proposed late fusion-based DES approach on a single well-balanced dataset, which was used for binary classification. However, real-world datasets often exhibit varying degrees of complexity, class imbalances, and distribution shifts. Consequently, further research is warranted to assess the performance of our proposed late fusion-based DES approach on more complex, diverse, and imbalanced datasets, as well as in multi-class classification scenarios. This would provide a more comprehensive understanding of the generalizability and robustness of our approach when faced with the intricacies and challenges commonly encountered in real-world applications.

7 Conclusion

This paper proposed an interpretable dynamic ensemble selection model with a late fusion of different modalities for detecting in-hospital mortality. Our proposed model outperformed the existing approaches, and the results suggest that the proposed DES approach for late fusion can be an effective technique for improving the accuracy of machine learning models, particularly in situations where diverse modalities and model pools are involved. These findings may have important implications for the development of more robust and accurate machine learning systems in various applications. Our future work will aim to address the limitations of our current study. Specifically, we intend to expand the evaluation of our proposed DES approach on diverse and unbalanced multi-class datasets, which will provide a more comprehensive assessment of its performance. To further enhance the methodology, we also plan to incorporate additional distance metrics, including Euclidean, Manhattan, Cosine, and customized metrics, as well as various selection methods. Furthermore, we intend to explore different types of dynamic ensemble selection techniques with late fusion setting and provide novel approaches to explainability.

Acknowledgments. This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience Program (IITP-2021-2020-0-01821) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1011198).

References

1. Bonaccorso, G.: Machine Learning Algorithms. Packt Publishing Ltd (2017)
2. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
3. Cao, Y., Geddes, T.A., Yang, J.Y.H., Yang, P.: Ensemble deep learning in bioinformatics. Nat. Mach. Intell. **2**(9), 500–508 (2020)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: synthetic minority over-sampling technique: smote. J. Artif. Intell. Res. **16**, 321–357 (2002)
5. Cruz, R.M.O., Hafemann, L.G., Sabourin, R., Cavalcanti., G.D.C.: Deslib: a dynamic ensemble selection library in python. J. Mach. Learn. Res. **21**(1), 283–287 (2020)
6. Cruz, R.M.O., Sabourin, R., Cavalcanti, G.D.C.: Recent advances and perspectives: dynamic classifier selection. Inf. Fusion **41**, 195–216 (2018)
7. El-Rashidy, N., Abuhmed, T., Alarabi, L., El-Bakry, H.M., Abdelrazek, S., Ali, F., El-Sappagh, S.: Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. Neural Computing and Applications, pp. 1–30 (2022)
8. El-Rashidy, N., El-Sappagh, S., Abuhmed, T., Abdelrazek, S., El-Bakry, H.M.: An improved patient-specific stacking ensemble model: intensive care unit mortality prediction. IEEE Access **8**, 133541–133564 (2020)
9. El-Sappagh, S., Abuhmed, T., Riazul Islam, S.M., Kwak, K.S.: Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. Neurocomputing, **412**, 197–215 (2020)

10. El-Sappagh, S., Ali, F., Abuhmed, F., Singh, J., Alonso, J.M.: Automatic detection of alzheimer's disease progression: an efficient information fusion approach with heterogeneous ensemble classifiers. Neurocomputing **512**, 203–224 (2022)
11. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**(1), 3133–3181 (2014)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
13. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**, 389–422 (2002)
14. Juraev, F., El-Sappagh, S., Abdukhamidov, E., Ali, F., Abuhmed, T.: Multilayer dynamic ensemble model for intensive care unit mortality prediction of neonate patients. J. Biomed. Inf. **135**, 104216 (2022)
15. Ko, A.H.R., Sabourin, R., Britto Jr, A.S.: From dynamic classifier selection to dynamic ensemble selection. Pattern Recognit. **41**(5), 1718–1731 (2008)
16. Mahesh, B.: Machine learning algorithms-a review. Int. J. Sci. Res. (IJSR).[Internet] **9**, 381–386 (2020)
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
18. Raffa, J.D., Johnson, A.E.W., O'Brien, Z., Pollard, T.J., Mark, R.G., Celi, L.A., Pilcher, D., Badawi, O.: The global open source severity of illness score (gossis). Crit. Care Med. **50**(7), 1040–1050 (2022)
19. Raschka, S.: Mlxtend: providing machine learning and data science utilities and extensions to python's scientific computing stack. J. Open Source Softw. **3**(24) (2018)
20. Ruta, D., Gabrys, B.: Classifier selection for majority voting. Inf. Fusion **6**(1), 63–81 (2005)
21. Sabourin, M., Mitiche, A., Thomas, D., Nagy, G.: Classifier combination for handprinted digit recognition. In: Proceedings of 2nd international conference on document analysis and recognition (ICDAR'93), pp. 163–166. IEEE (1993)
22. Sagi, O., Rokach, L.: Ensemble learning: a survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Disc. **8**(4), e1249 (2018)
23. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., Stamatopoulos, P.: Stacking classifiers for anti-spam filtering of e-mail (2001). cs/arXiv:0106040
24. Sammut, C., Webb, G.I. (eds.) Holdout Evaluation, pp. 506–507. Springer, US, Boston, MA (2010)
25. Schapire, R.E.: A brief introduction to boosting. In: Ijcai, vol. 99, pp. 1401–1406. Citeseer (1999)
26. Stahlschmidt, S.R., Ulfenborg, B., Synnergren, J.: Multimodal deep learning for biomedical data fusion: a review. Briefings Bioinform. **23**(2), bbab569 (2022)
27. Tan, W., Tiwari, P., Pandey, H.M., Moreira, C., Jaiswal, A.M.: Multimodal medical image fusion algorithm in the era of big data. Neural Computing and Applications, pp. 1–21 (2020)
28. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. Bioinformatics **17**(6), 520–525 (2001)