DOI: https://doi.org/10.15276/hait.08.2025.18
UDC: 004.1

Method for building ensemble classifiers of structured and unstructured data based on a unified approach

Olena O. Arsirii[1)]
ORCID: https://orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Oleksandr K. Andronati[1)]
ORCID: https://orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID: 58677655800
[1)] Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

Effectively classifying heterogeneous data, including structured and unstructured data, is essential in diverse fields such as healthcare, finance, information security, and audio content analysis. This study aims to develop a unified approach for constructing ensemble classifiers capable of handling diverse data formats within a single framework, enhancing classification accuracy and robustness. The methodology integrates feature extraction and data preprocessing techniques, transforming heterogeneous datasets to a standardized numerical format suitable for ensemble learning. Eight base classifiers including K-nearest neighbors, support vector machines, random forest, extreme gradient boosting, logistic regression, multilayer perception, convolution neural networks and long short-term memory networks–were trained with optimized hyperparameters. The ensemble classification uses stacking with various aggregation types such as hard voting, soft voting, and soft voting using Gompertz fuzzy ranking to effectively combine model predictions while accounting for uncertainty and noise. Experimental evaluation across five datasets, covering medical diagnosis, credit risk, emotion recognition, music genres and deepfake detection–demonstrates consistent improvement in accuracy and F1-score metrics, with gains up to 8 percent compared to the best individual classifiers. The approach proves particularly effective for unstructured audio data, where temporal and spectral dependencies pose significant challenges. The results underscore the versatility the proposed unified ensemble methodology in addressing class imbalance and noise offering a scalable solution adaptable to various domains. This work contributes a comprehensive framework facilitating the development of robust classifiers for complex real-world data and paves the way for future research integrating heterogeneous data sources within cohesive predictive models.

Keywords: Ensemble classifiers; machine learning; hard voting; soft voting; Gompertz function

***For citation*: Arsirii O. O., Andronati O. K. "Method for building ensemble classifiers of structured and unstructured data based on a unified approach" . *Herald of Advanced Information Technolog*. 2025; Vol.8 No.3: 288–300. DOI: https://doi.org/10.15276/hait.08.2025.18**

INTRODUCTION

Solving data classification problems that arise in various subject areas, such as healthcare, finance, information security and audio content analysis, involves developing effective machine learning models. For example, in healthcare, early diagnosis of autism based on behavioral and clinical signs allows for timely treatment, significantly improving patients' quality of life [1]. In finance, accurate credit risk assessment based on transaction data, credit history, and customer profiles minimizes financial losses and optimizes decision-making processes [2]. In the audio data analysis field, emotion recognition, music genre recognition, and fake speech detection have applications in security systems, recommendation systems, and healthcare [3], [4]. At the same time, the challenges in developing machine learning models to solve such tasks is associated with the need to process both structured tabular data (e.g., medical and financial datasets) and unstructured audio data (e.g., for emotion recognition, music genre recognition, or speech authenticity analysis). The data format heterogeneity, large volumes, and structural complexity create additional challenges for ML developers, including problems of overfitting, class imbalance, and insufficient classification accuracy [5]. Traditional classification methods, such as neural networks, decision trees, or support vector machine, are known to be successfully applied to data analysis, but their effectiveness is limited by the data complexity and heterogeneity, as well as problems of overfitting and insufficient generalization ability [5]. To overcome these limitations, a promising solution is to use ensemble classifiers, which combine several models to achieve higher accuracy compared to individual classifiers. Building ensembles allows combining different structured and unstructured data processing methods using a unified approach. This minimizes the shortcomings of individual models and increases resistance to noise and data uncertainty.

In previous works, the authors have demonstrated the effectiveness of ensemble

classifiers in processing audio data, such as recognizing emotions in human speech and classifying music genres [3]. Similarly, ensemble methods have shown high performance in structured data classification tasks, including medical diagnosis and financial analysis [6]. This study generalizes these approaches, extending them to new subject areas, including information security, where audio data authenticity analysis is becoming critically important [4]. The proposed method for constructing ensemble classifiers is also based on a unified approach, but, taking into account the authors' experience, it allows heterogeneous structured and unstructured data to be reduced to a single classification task, using a single end-to-end technology for data preprocessing, feature extraction and selection, developing base classifier models (weak learners), building ensembles for a meta-classifier, and assessing classification quality. The implementation of classifiers based on the proposed method improves the accuracy and versatility of data analysis, regardless of its nature and subject area.

LITERATURE REVIEW AND PROBLEM STATEMENT

Despite significant progress in the developing classification algorithms in fields like healthcare, finance, information security, and audio analysis, existing methods face a number of limitations related to the heterogeneous data processing, class imbalance, and the complexity of feature extraction [3], [4], [5]. In this section, based on an analysis of the literature, the need to develop ensemble classifiers based on a unified approach is justified, and the features of its stages are considered.

It is known that, according to the unified approach, the development of ensemble classifiers is carried out in three stages: feature preparation and extraction, development of weak learners (basic classifiers), and ensemble formation (combining weak learners into a meta-classifier) [3], [6], [7].

According to the authors [8], the challenges in converting structured and unstructured data to a standardized numerical format during feature preparation and extraction lead to global classification problems in subject areas. For example, in healthcare, classifying medical data to diagnose diseases like autism is complicated by the heterogeneity of data, including clinical records, behavioral indicators, and biomarkers. Class imbalance due to the rarity of positive cases and limited data volume reduces the accuracy of models [1]. In addition, missing values and noise in medical data require complex preprocessing methods [9]. In finance, the classification of banking data for credit risk assessment faces problems of missing values, nonlinear dependencies, and noise, which makes it difficult to build reliable models [2], [6]. For example, banking data often contains incomplete records of transactions or credit history, which reduces the generalizability of algorithms [10].

In the field of information security, detecting deepfakes in audio data poses significant challenges due to subtle distortions that are difficult to detect. The high audio data dimensionality and the need to extract features like MFCC or STFT increase sensitivity to noise [4]. Similarly, in audio content analysis, for example, in emotion or music genre recognition, problems are associated with processing large amounts of data and the need to extract stable features in the presence of noise and variations in recordings [3], [5]. These challenges highlight the need to develop approaches that can effectively process heterogeneous data and take into account their specific characteristics.

Literature sources [7], [8] note that various machine learning methods are used to develop basic classifiers for further assembly, including K-nearest neighbors (KNN), support vector machines (SVM), random forest, gradient boosting (XGBoost), logistic regression, multilayer perceptrons (MLP), convolutional neural networks (CNN), and recurrent neural networks (LSTM). Each of these methods has its advantages, but also has significant limitations.

The authors of the study [9] note that the KNN algorithm has several limitations, including sensitivity to noisy data and outliers, high computational cost with large datasets, and the need to select an optimal k value. In high-dimensional datasets, the effectiveness of distance metrics degrades due to the curse of dimensionality, which adversely affects KNN classification accuracy.

According to the study [12], the SVM classifier is effective for medium-sized datasets, but when working with large volumes of banking data, such as mortgage loans with a large number of categorical and quantitative features, it faces serious computational limitations and deterioration in classification quality. The authors note that traditional SVM has difficulty processing large and unbalanced samples, which reduces its applicability in credit scoring tasks.

The authors of the study [13] note that the Random Forest classifier has a built-in mechanism for dealing with imbalance, but may show less accurate predictions when there is a strong class skew. Also, in Random Forest, small changes in

hyperparameters affect all trees at once, which can lead to poor predictions.

In study [14], the authors describe the XGBoost algorithm, noting that the effectiveness of XGBoost largely depends on the correct selection of hyperparameters and the application of calibration methods, which are necessary to reduce overfitting and improve accuracy on unbalanced data. The authors note that, despite its high performance, XGBoost requires careful tuning and model verification, especially in the financial field with its diverse data structure and noise.

Regarding logistic regression, study [15] showed that when applying logistic regression to banking data, the accuracy of the model is usually lower compared to more complex methods, especially on data with nonlinear or complex dependencies.

Study [4] describes the effectiveness of deep models like CNN and LSTM, for audio data classification due to their ability to capture temporal and spectral dependencies, but they require large amounts of data. In addition, these models are difficult to tune and sensitive to noise.

Thus, the literature analysis shows that the limitations of basic models, such as sensitivity to noise, overfitting, and inability to effectively handle nonlinear dependencies or class imbalance, make them insufficiently versatile for working with heterogeneous data [4], [5], [16].

To overcome these limitations when building classifiers, it is proposed to use ensemble methods such as bagging, boosting, and stacking [13], [16]. The authors believe that it is the use of meta-classifiers that allows these limitations to be overcome by combining the predictions of several models, which increases accuracy and robustness.

In [16], it is shown that bootstrap aggregating creates several subsamples of data and trains independent models, reducing the variance of predictions. However, bootstrap aggregating may be less effective on unbalanced data. Boosting, on the other hand, sequentially trains models, correcting the errors of previous ones, which improves accuracy but increases the risk of overfitting and the complexity of configuration.

Classic stacking [3] uses a meta-classifier to aggregate the predictions of base models such as KNN, SVM, or MLP. This method is easy to implement and flexible and allows to combine heterogeneous models. At the same time, stacking is considered particularly effective for accounting for uncertainty and increasing noise resistance, which makes it preferable for tasks with heterogeneous data.

Thus, having considered the current unified approach to the development of ensemble classifiers and analyzed the advantages and disadvantages of its stages, we can formulate the goal and objectives of this study.

THE AIM AND OBJECTIVES OF THE RESEARCH

The aim of the research is to improve the accuracy of autism prediction, credit risk assessment, deepfake detection, emotion recognition, and music genre classification by developing a method for creating ensemble classifiers of structured and unstructured data based on a unified approach.

Research objectives:

1) justify the datasets selection from four subject areas (medicine, finance, information security, audio content analysis), bring data into a unified numerical format that ensures compatibility with the ensemble classifier and perform preprocessing of heterogeneous data (tabular and audio data).

2) develop models of weak learners–basic classifiers (KNN, SVM, Random Forest, XGBoost, MLP, CNN, LSTM) by optimizing hyperparameters when training on datasets from four subject areas.

3) develop a stacking-based meta-classifier that includes hard and soft voting, as well as soft voting with fuzzy prediction aggregation methods to account for uncertainty and improve model robustness.

4) experimentally evaluate the meta-classifier's accuracy in autism prediction, credit risk assessment, emotion recognition, deepfake detection, and music genre classification using metrics such as accuracy, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC), and compare the results with the evaluations of the base classifiers.

5) analyze the results and justify the possibility of using a unified approach to classify data of different nature.

THE RESEARCH MATERIALS AND METHODS

Acquisition and Preprocessing of Data

To solve the problem of converting data to a single numerical format compatible with the ensemble classifier of structured and unstructured data under development, a number of datasets from

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

open sources [14], [4], [17], [18], [19], [20] were analyzed according to the following parameters: data volume, quality and markup, class diversity, structured or unstructured nature, representation formats, and source data. For structured data, the completeness and correctness of tabular information is important, while for audio data, the readability of the sound signal, sampling frequency, noise level, and type of recorded sound events are important. These parameters were taken into account when selecting datasets for the development of an ensemble classifier that ensures versatility and reliability.

The following selection was made. Structured data is presented in tabular formats: Autistic Spectrum Disorder Screening Data for Adults (for predicting autism in healthcare) and Home Equity Line of Credit (HELOC, for assessing credit risk in finance). Unstructured data – audio signals: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS, for speech emotion recognition), GTZAN (for music genre classification), and Fake or Real (FoR for synthetic speech detection in information security).

1. Autistic Spectrum Disorder Screening Data for Adults [17]: contains 800 records with 13 attributes, including binary behavioral characteristics (AQ-10) and demographic data (age, gender). The target variable is the presence/absence of autism. The dataset was split into 600 records in training set and 200 records in testing set.

2. HELOC [18]: Contains 10,459 records with 23 attributes, including financial indicators (time since last delinquency, number of transactions). The target variable is credit risk (binary). The dataset was split into 7,844 records in the training set and 2,615 in the testing set.

3. RAVDESS [19]: 1,440 audio recordings with 7 emotions (anger, happiness, etc.). The dataset was split into 1080 records in the training set, 360 in the testing set.

4. GTZAN [20]: 1000 audio files, 30 seconds each, 10 genres (100 files per genre). The dataset was split into 750 records in the training set and 250 in testing set.

5. FoR [21]: 17870 audio files for synthetic speech detection (real and fake audio). The dataset was split into 13,956 in the training set and 3,914 in the testing set.

For the Autistic Spectrum Disorder Screening Data for Adults dataset, preprocessing included encoding categorical features (gender) and normalizing numerical features (age) to the range [0, 1].

For the HELOC dataset, preprocessing consisted of filling in missing values with mean/median values and normalizing numerical features to the range [0, 1].

For the unstructured RAVDESS, GTZAN, and FoR datasets, spectral characteristics were obtained from raw audio files. According to the researches [22], [23], [24] the most promising spectral characteristics for audio data classification are spectral centroid, spectral flatness, spectral contrast, spectral roll-off, zero crossing rate.

Spectral centroid – the location of the center of mass of the spectrum.

The calculation of the spectral centroid is given in formula (1).

$$SpectralCentroid = \frac{\sum_{k=1}^{N} kF[k]}{\sum_{k=1}^{N} F[k]}, \quad (1)$$

where $F[k]$ is the amplitude corresponding to the k-th bin in the discrete Fourier transform (DFT) spectrum.

Spectral centroid is shown in Fig. 1.

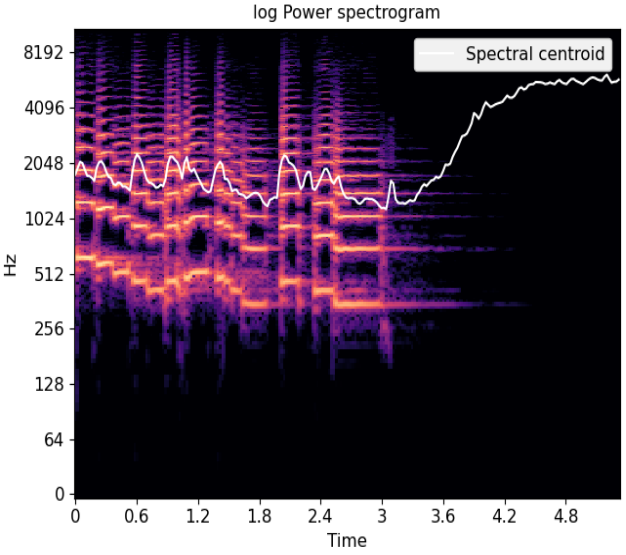


Fig. 1. **Spectral centroid**
***Source*: compiled by the [25]**

Spectral flatness – the measure to quantify how much noise-like a sound is, as opposed to being tone-like. The calculation of the spectral flatness is given in formula (2).

$$SpectralFlatness = \frac{\sqrt[N]{\prod_{k=0}^{N-1} F[k]}}{\frac{\sum_{k=0}^{N-1} F[k]}{N}} \quad (2)$$

where $F[k]$ is the amplitude corresponding to the k-th bin in the discrete Fourier transform spectrum.

Spectral contrast – the measure of the energy of frequency at each timestamp.

To obtain spectral contrast, it is necessary to calculate spectral peaks and spectral declines for each bin.

$$Peak_k = log\left(\frac{1}{aN}\sum_{i=1}^{aN} F[k]_i\right), \quad (3)$$

$$Valley_k = log\left(\frac{1}{aN}\sum_{i=1}^{aN} F[k]_{N-i+1}\right), \quad (4)$$

where $F[k]$ is the amplitude corresponding to the k-th bin in the DFT spectrum; k is the bin number; N is the number of subbands of each bin; a is an additional coefficient determined experimentally, most often its value is close to 0.02.

The spectral contrast is calculated as their difference.

$$Spectral\ Conrast = Peak_k - Valley_k. \quad (5)$$

Spectral contrast is shown in Fig. 2.

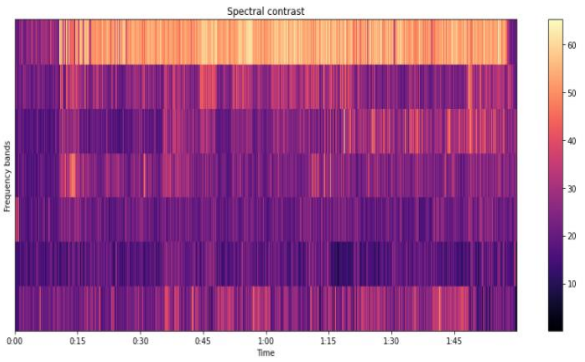


Fig. 2. **Spectral** contrast
***Source*: compiled by the [25]**

Spectral roll-off – the action of a specific type of filter which is designed to roll off the frequencies outside to a specific range.

Spectral roll-off is shown in Fig. 3.

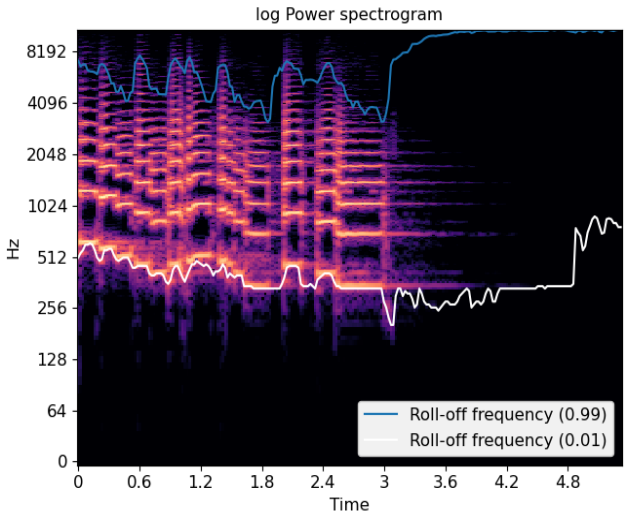


Fig. 3. **Spectral roll-off**
***Source*: compiled by the [25]**

Zero crossing rate – the measure of the rate at which the signal is changing positive to negative or vice versa.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} I_{R<0}(s_t s_{t-1}). \quad (6)$$

where s is the signal; T is the signal length; $I_{R<0}$ $(s_t$ $s_{t-1})$ is the indicator of a sign change in the signal during the time interval s_t s_{t-1}

RMS (root mean square) – measures the average loudness of an audio.

MFCC – the coefficients that are derived from a type of inverse Fourier transform (cepstral) representation. MFCC allow a better representation of sound because the frequency bands are equally distributed on the Mel scale which approximates the human auditory system's response more closely.

Mel is a unit of sound pitch based on the perception of this sound by our hearing organs. Since the frequency response of the human ear is not linear, amplitude is not an entirely accurate measure of sound loudness. Similarly, the pitch of a sound perceived by human hearing does not depend linearly on its frequency [16].

This dependence is described by the simple formula (7):

$$m = 1125\ ln\left(1 + \frac{f}{700}\right). \quad (7)$$

where f is the frequency in hertz.

The graph showing the dependence of sound pitch in mels on the frequency of vibrations is shown in Fig. 4.

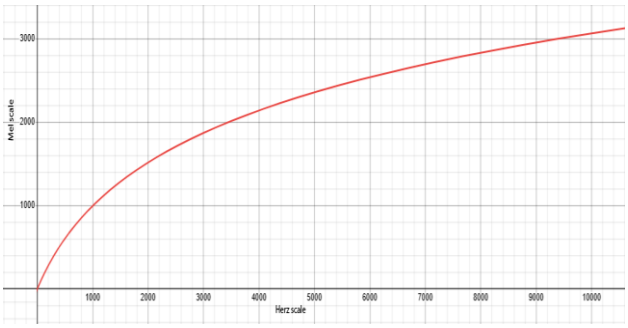


Fig. 4. **Dependence of sound pitch in mels on the frequency of vibrations**
***Source*: compiled by authors**

For the analysis of the audio data of the GTZAN dataset, additional spectral characteristics such as chromagram, Constant-Q chromagram, and Chroma Energy Normalized were selected.

Chromagram is defined as the whole spectral audio information mapped into one octave. Each octave is divided into 12 bins representing each one semitone [3].

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Also, for RAVDESS, GTZAN, and FoR, spectra were used as images for CNN, and MFCC were used as a time series input for LSTM.

Development of classifier models

The next step in the unified approach to classifying structured and unstructured data is to select a set of eight basic classifiers covering both classical machine learning methods and deep neural networks. The set includes: K-nearest neighbors (KNN), support vector machines (SVM), random forest, extreme gradient boosting (XGBoost), multilayer perceptron (MLP), convolutional neural network (CNN), long short-term memory (LSTM) and logistic regression (LR).

This selection was driven by the need to combine models with different strengths. This set allows the models to be adapted to heterogeneous data, including tabular structured sets (Autistic Spectrum Disorder Screening Data and HELOC) and audio signals (RAVDESS, GTZAN, FoR), where all data is converted to a single numerical format after preprocessing.

Model development began with the definition of their basic architecture and parameters.

Table 1 lists the main hyperparameters for each algorithm.

In cases where it was difficult to manually select the optimal hyperparameters for the models, Grid Search Cross Validation (GSCV) technology was used with 5-fold cross-validation on training samples to avoid overfitting and ensure generalization. For each model, a grid range of parameters was set (for XGBoost, for example, the hyperparameter grid included variations in learning rate (0.01, 0.05, 0.1, 0.2, 0.3), maximum tree depth (3, 5, 7, 10), and number of estimators (50, 100, 200, 300), and combinations were selected that maximized the F1-score or accuracy on the validation subsample.

Overfitting control was also implemented by monitoring the difference in metrics between the training and test samples, and early stopping was used in the case of neural network algorithms.

The development stages included:

1) Initialization of models with basic parameters and pre-processed data;

2) Launch of GSCV to search for optimal hyperparameters, fixing the best ones according to cross-validation metrics;

3) Final training on the full training sample and validation on the test sample. This approach ensures uniformity: all models generate probabilistic predictions (soft outputs) for subsequent ensemble training, increasing overall stability and accuracy in diverse subject areas.

Table 1. **Algorithms and their hyperparameters**

Algorithm	Hyperparameters
KNN	1. Number of neighbors to use 2. Metric for distance computation
Logistic Regression	1. Maximum number of iterations
SVM	1. Kernel type 2. Regularization parameter C 3. Degree of the kernel function (if kernel type is polynomial)
Random Forest	1. The number of trees in the forest. 2. The maximum depth of the tree. 3. The minimum number of samples required to be at a leaf node.
XGBoost	1. Number of trees 2. The maximum depth of the tree. 3. Minimum sum of instance weight needed in a child. 4. Boosting learning rate 5. Gamma – minimum loss reduction required to make a further partition on a leaf node of the tree
MLP	1. The number of hidden layers 2. The number of neurons in each layer 3. Activation function of hidden layers 4. Percent in dropout layers 5. The optimizer
CNN	1. The number of convolutional layers 2. Number of Pooling layers 3. The number of fully connected layers 4. The number of neurons in them 5. Layer activation functions 6. The optimizer
LSTM	1. The number of LSTM layers 2. The number of neurons in LSTM layers 3. The number of fully connected layers 4. The number of neurons in fully connected layers 5. Layer activation functions 6. Recurrent activation function 7. The optimizer

***Source*: compiled by the authors**

Development of Ensemble Classifiers

The following approaches are used to aggregate predictions in stacking: hard voting, soft voting, and soft voting with fuzzy Gompertz ranking.

In hard voting, the final decision of the ensemble is based on the majority of model votes; each model makes a prediction, and the option with the highest number of votes is selected. This method is suitable for balanced classifiers with reliable predictions.

The calculation of hard voting:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_n), \quad (8)$$

where $\hat{y}$ is the final predicted class determined by the ensemble; y_i is the predicted class by the i-th individual classifier, where i ranges from 1 to n; n is the total number of individual classifiers in the ensemble.

In soft voting, each classifier assigns probabilities to classes, and the final decision is determined by the weighted average of these probabilities, taking into account the confidence of the models.

The calculation of Soft Voting:

$$p_k = (1/n) \Sigma\, p_{ik}$$
$$\hat{y} = \text{argmax}\, p_k, \quad (9)$$

where $\hat{y}$ is the final predicted class determined by the ensemble; p_{ik} is the probability of class k predicted by the i-th classifier, where i ranges from 1 to n and k represents the class index; n is the total number of individual classifiers in the ensemble.

Soft voting using fuzzy ranks (Gompertz) allows taking into account uncertainty and heterogeneity in the data, which improves the accuracy and stability of the ensemble classifier in the classification of audio data [3], [11].

The calculation of Soft Voting using fuzzy ranks (Gompertz):

$$p'_{i,k} = a \cdot exp(-b \cdot exp(-c \cdot p_{i,k})),$$
$$p_k = (1/n) \Sigma\, p'_{i,k},$$
$$\hat{y} = \text{argmax}_k\, p_k. \quad (10)$$

where $\hat{y}$ is the final predicted class determined by the ensemble; $p'_{i,k}$ is the adjusted probability of class k for the i-th classifier after applying the Gompertz function; p_k is the average probability of class k across all classifiers in the ensemble; n is the total number of individual classifiers in the ensemble; a, b, c are the parameters of the Gompertz function used for adjusting probabilities, where a controls the upper asymptote, and b and c control the shape of the curve.

As part of the proposed methodology, ensemble classifiers consisting of different combinations of weak learners were created for each of the five classification tasks, each of which represented its own subject area. For the Autistic Spectrum Disorder Screening Data for Adults and HELOC datasets, six individual classifiers were used (all except CNN and LSTM, given the nature of the data), while all eight were used for the remaining three datasets. The chosen ensemble approach is based on a stacking method in which several baseline models are trained on the same data and their predictions are passed as new features to a meta-model, the ensemble classifier, which is trained to combine these predictions [16].

Given the fact that we have three types of prediction aggregations we have following number of ensemble classifiers.

For the Autistic Spectrum Disorder Screening Data for Adults and HELOC datasets:

$$C_n^k = 3\,(C_6^3 + C_6^4 + C_6^5 + C_6^6) = 126.$$

For RAVDESS, GTZAN and FOR datasets:

$$C_n^k = 3\,(C_8^3 + C_8^4 + C_8^5 + C_8^6 + C_8^7 + C_8^8) = 657,$$

where C_n^k is the number of combinations from n to k, multiplication with a factor of 3 is explained by the fact that we have three types of ensemble classifiers aggregation – hard voting, soft voting, soft voting with Gompertz aggregation.

Despite the construction of a large number of ensembles with various weak classifier combinations, the computational complexity of the process remains moderate due to the simplicity of the voting stage. The main computational cost is the training weak classifiers, which is done once. This allows generating many ensembles by varying the composition of classifiers without significantly increasing the overall computational load, since the additional costs of voting are minimal compared to model training.

The performance of all ensembles was evaluated on test samples using key classification quality metrics: accuracy, F1-score, and ROC-AUC, which provided a comprehensive picture of model effectiveness and allowed to select the best configurations based on maximum accuracy and F1-score values. The use of these metrics is critical for objective evaluation, as they cover different aspects of model performance and help avoid biases associated with data characteristics. Accuracy measures the proportion of correct predictions out of the total number of instances, providing an intuitive overall measure of success, but it can be misleading in unbalanced datasets where the prevalence of one class (e.g., negative cases in Autism prediction) leads to artificially high values by ignoring errors in minor classes. The F1-score, as the harmonic mean of precision and recall, is especially valuable for tasks with class imbalance, as it balances between minimizing false positives (precision) and false negatives (recall). In multi-class classification (such as RAVDESS with 7 emotions or GTZAN with 10 genres), the use of the F1-score made it possible to take into account performance across all classes. As a result, the best ensembles were selected based on a combined criterion of maximum accuracy and F1-score, which ensured the selection of configurations that are optimal for real-world applications in healthcare, finance, and information security, where

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

minimizing errors is critical. If there is no ensemble that has the best performance on both metrics (for example, one ensemble has the maximum accuracy value, while another has the maximum f-score value), preference is given to the ensemble with the maximum f-score value.

RESULTS OF THE STUDY

The key unified approach feature is building many ensembles with different classifier combinations of classifiers. The diversity and variability are the main reasons for the improvement in classification metrics, as they allow the strengths of different models to be taken into account and their weaknesses to be compensated for in heterogeneous data conditions. In the experiments, a significant number of ensembles were created for each dataset (126 for the structured Autism and HELOC datasets, 657 for the unstructured RAVDESS, GTZAN and FoR audio datasets), which ensured a comprehensive search for optimal configurations. Analysis of the results showed that ensembles with soft voting and Gompertz fuzzy ranking often outperform hard voting aggregation, especially in tasks with unbalanced classes or high-dimensional data, where accounting for prediction uncertainty plays a key role. In addition, there is a tendency for the inclusion of neural network models (MLP, CNN, LSTM) in ensembles to significantly improve metrics for audio data, while combinations with XGBoost and SVM dominate for tabular data.

For the Autistic Screening Data dataset, out of 126 designed ensemble classifiers, the best results were shown by ensemble consisting of NN, SVM, XGB with soft voting with Gompertz fuzzy ranking. The following metrics were obtained:

• Accuracy = 0.875 (+1.5% relative to the best classifier in the compound);

• F1 Score = 0.873 (+1.3% relative to the best classifier in the compound);

• AUC = 0.874 (the ROC-AUC curve is shown in Fig. 5).

Complete information comparing the best ensemble with individual classifiers is given in Table 2.

For the HELOC dataset, out of 126 designed ensemble classifiers, the best results were shown by ensemble consisting of NN, SVM, Random Forest, XGB, and Logistic Regression with soft voting.

The following metrics were obtained:

• Accuracy = 0.736 (+1 % relative to the best classifier in the compound);

• F1 Score = 0.734 (+1 % relative to the best classifier in the compound);

• AUC = 0.733 (the ROC-AUC curve is shown in Fig. 6).

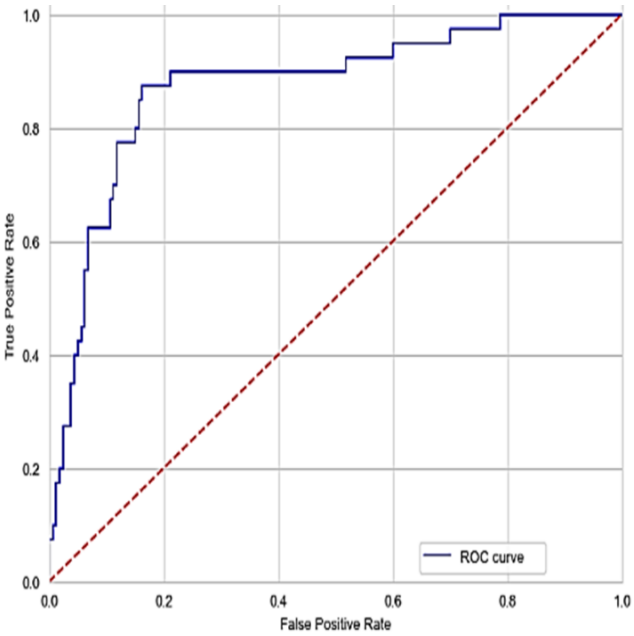


Fig. 5. **ROC-AUC curve for the best ensemble for autistic screening data**

***Source*: compiled by the authors**

Table 2. **Best ensemble for autism prediction**

Algorithm	Metrics for Autistic Screening Data			
	Accuracy	***F1-score***	***Accuracy difference***	***F1-score difference***
Ensemble	***0.875***	***0.873***		
SVM	0.845	0.846	3 %	2.7 %
MLP	0.85	0.853	2.5 %	2 %
XGB	0.86	0.86	1.5 %	1.3 %
Random Forest	0.85	0.845	2.5 %	2.8 %
KNN	0.835	0.834	4 %	3.9 %
Logistic Regression	0.855	0.856	2 %	1.7 %

***Source*: compiled by the authors**

Complete information comparing the best ensemble with individual classifiers is given in Table 3.

Table 3. **Best ensemble for HELOC**

Algorithm	Metrics for HELOC			
	Accuracy	***F1-score***	***Accuracy difference***	***F1-score difference***
Ensemble	***0.736***	***0.734***		
SVM	0.715	0.714	2.1 %	2 %
MLP	0.720	0.720	1.6 %	1.4 %
XGB	0.725	0.725	1.1 %	1 %
Random Forest	0.722	0.721	1.4 %	1.3 %
KNN	0.685	0.685	5.1 %	4.8 %
Logistic Regression	0.726	0.72	1 %	1 %

***Source*: compiled by the authors**

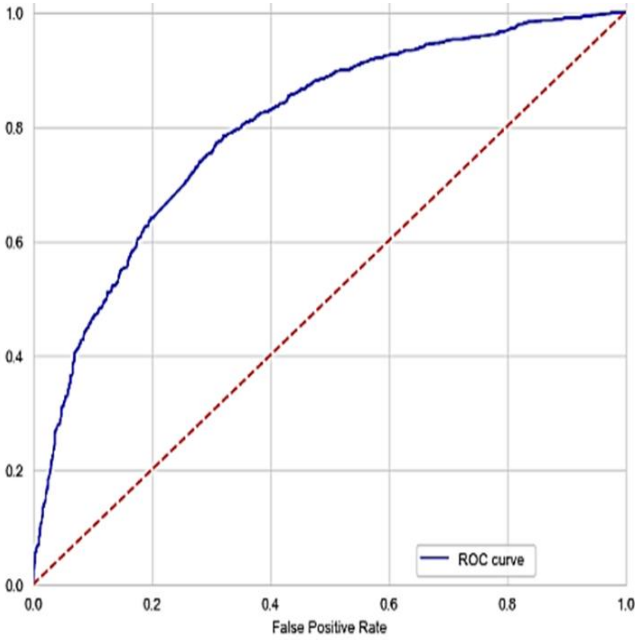


Fig. 6. **ROC-AUC curve for best ensemble for HELOC**

***Source*: compiled by the authors**

For the GTZAN dataset, out of 657 designed ensemble classifiers, the best results were shown by ensemble consisting of MLP, KNN, SVM, Random Forest, CNN, LSTM, XGB with soft voting using fuzzy ranks (Gompertz) aggregation. The following metrics were obtained:

• Accuracy = 0.812 (+5.6 % relative to the best classifier in the compound);

• F1 Score = 0.808 (+5.1 % relative to the best classifier in the compound).

Complete information comparing the best ensemble with individual classifiers is given in Table 4.

Table 4. **Best ensemble for GTZAN**

Algorithm	Metrics for music classification			
	Accuracy	***F1-score***	***Accuracy difference***	***F1-score difference***
Ensemble	0.812	0.808		
SVM	0.74	0.738	7.6%	7%
MLP	0.76	0.757	5.6%	5.1%
XGB	0.688	0.685	12.8%	12.3%
Random Forest	0.668	0.659	14.8%	14.9%
KNN	0.676	0.671	14%	13.7%
CNN	0.64	0.623	17.6%	18.5%
LSTM	0.448	0.438	36.8 %	36 %
Logistic Regression	0.696	0.689	11.6 %	11.9 %

***Source*: compiled by the authors**

For the RAVDESS dataset, out of 657 designed ensemble classifiers, the best results were shown by ensemble consisting of MLP, KNN, SVM, Random Forest, CNN, and LSTM with soft voting using fuzzy ranks (Gompertz) aggregation. The following metrics were obtained:

• Accuracy = 0.808 (+8 % relative to the best classifier in the compound);

• F1 Score = 0.806 (+8 % relative to the best classifier in the compound).

Complete information comparing the best ensemble with individual classifiers is given in Table 5.

Table 5. **Best ensemble for RAVDESS**

Algorithm	Metrics for emotion classification			
	Accuracy	***F1-score***	***Accuracy difference***	***F1-score difference***
Ensemble	***0.808***	***0.806***		
SVM	0.703	0.699	10.5 %	10.7 %
MLP	0.728	0.726	8 %	8 %
XGB	0.656	0.648	15.2 %	15.8 %
Random Forest	0.653	0.643	15.5 %	16.3 %
KNN	0.62	0.608	18.8 %	19.8 %
CNN	0.617	0.597	19.1 %	20.9 %
LSTM	0.597	0.592	21.1 %	21.4 %
Logistic Regression	0.558	0.533	25 %	27.3 %

***Source*: compiled by the authors**

For the Deepfake FOR dataset, out of 657 designed ensemble classifiers, the best results were shown by Ensemble consisting of KNN, RF, CNN

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

with soft voting. The following metrics were obtained:

• Accuracy = 0.935 (+3.9 % relative to the best classifier in the compound);

• F1 Score = 0.935 (+3.9 % relative to the best classifier in the compound).

Complete information comparing the best ensemble with individual classifiers is given in Table 6.

To summarize the results for all datasets, Table 7 was compiled, showing the average ensemble metric improvements over the best individual models.

This confirms the effectiveness of the approach: the average improvement in accuracy is 4 %, F1-score – 3.8 %, with the largest increase for audio data (up to 8 %), where data's unstructured nature requires model combinations to capture temporal and spectral dependencies.

Table 6. **Best ensemble for FoR**

Algorithm	Metrics for FoR			
	Accuracy	***F1-score***	***Accuracy difference***	***F1-score difference***
Ensemble	***0.935***	***0.935***		
SVM	0.816	0.813	11.9 %	12.2 %
MLP	0.852	0.827	8.3 %	10.8 %
XGB	0.792	0.789	14.3 %	14.6 %
Random Forest	0.828	0.827	10.7 %	10.8 %
KNN	0.896	0.896	3.9 %	3.9 %
CNN	0.894	0.903	4.1 %	3.2 %
LSTM	0.881	0.876	5.4 %	5.9 %
Logistic Regression	0.816	0.813	11.9 %	12.2 %

***Source*: compiled by the authors**

Table 7. **Summarizing the results**

Dataset	Increase in Accuracy (%)	Increase in F1-score (%)	Best aggregation type
Autistic Screening Data	1.5	1.3	Soft voting with Gompertz
HELOC	1	1	Soft voting
GTZAN	5.6	5.1	Soft voting with Gompertz
RAVDESS	8	8	Soft voting with Gompertz
FoR	3.9	3.9	Soft voting
Average	**4**	**3.8**	

***Source*: compiled by the authors**

The results also indicate that improvements are smaller for binary tasks (Autism, HELOC, FoR) than for multi-class tasks (GTZAN, RAVDESS), which is due to the greater complexity of the latter. Overall conclusion of the section: the unified approach demonstrates the consistent superiority of ensembles over individual models, with the greatest effect in tasks with unstructured data.

CONCLUSIONS

The developed method for constructing ensemble classifiers based on a unified approach has demonstrated high efficiency in classifying structured and unstructured data across various fields, including healthcare, finance, information security, and audio content analysis. The research objective has been achieved: the proposed methodology, which includes data preprocessing, optimization of individual classifiers (KNN, SVM, Random Forest, XGBoost, MLP, CNN, LSTM, Logistic Regression) and stacking with various types of aggregation (hard voting, soft voting, soft voting with Gompertz fuzzy ranking), ensures that heterogeneous data is converted to a single format and improves classification accuracy. Experiments on five datasets (Autistic Spectrum Disorder Screening Data for Adults, HELOC, RAVDESS, GTZAN, FoR) confirmed the universality of the approach: for structured tabular data (healthcare and finance), the metric improvements were 1-1.5 % in accuracy and F1-score, and for unstructured audio data (emotion and genre analysis, and synthetic speech detection), the improvement was up to 8%, with an overall average gains of 4 % in accuracy and 3.8 % in F1-score over the best individual models. This highlights the method's ability to adapt to different types of data and tasks, minimizing class imbalance, noise, and overfitting issues by combining models. Using this method allows to automatically determine the composition of an ensemble classifier for a specific dataset of a specific subject area within the method.

The statistics obtained confirm the practical value: the best ensembles, often using Gompertz to account for uncertainty, outperform individual classifiers by 1-8 % on key metrics, making the approach applicable in real-world systems where accuracy is critical – from early diagnosis of autism and credit risk assessment to deepfake detection in security systems and music recommendations. The method's unified nature allows the use of a single methodology for preprocessing (extracting MFCC

and spectrograms for audio, normalization for tabular data), tuning, and evaluation, regardless of the nature of the data, which simplifies implementation in interdisciplinary projects.

Hence, using this unified approach to build ensemble classifiers for structured and unstructured data is a promising research direction.

REFERENCES

1. Thabtah, F., Abdelhamid, N. & Peebles, D. "A machine learning autism classification based on logistic regression analysis". *Health Information Science and Systems*. 2019; 7 (1): 12, https://www.scopus.com/authid/detail.uri?authorId=8349253300. DOI: https://doi.org/10.1007/s13755-019-0073-5.

2. Dastile, X., Celik, T. & Potsane, M. "Statistical and machine learning models in credit scoring: A systematic literature survey". *Applied Soft Computing*. 2020; 91: 106263. DOI: https://doi.org/10.1016/j.asoc.2020.106263.

3. Andronati, O., Antoshchuk, S., Babilunha, O., Arsirii, O., Nikolenko, A. & Mikhalev, K. "A method of constructing ensemble classifiers for recognizing audio data of various ,atures". *14th International Conference on Advanced Computer Information Technologies (ACIT)*. Ceske Budejovice, Czech Republic. 2024. p. 758–761, https://www.scopus.com/authid/detail.uri?authorId=58677655800.
DOI: https://doi.org/10.1109/ACIT62333.2024.10712469.

4. Khanjani, Z., Watson, G. & Janeja, V. P. "Audio deepfakes: A survey". *Frontiers in Big Data*. 2023; 5: 1001063. DOI: https://doi.org/10.3389/fdata.2022.1001063.

5. Gourisaria, M. K., Agrawal, R., Sahni, M., et al. "Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques". *Discov Internet Things*. 2024; 4: 1. DOI: https://doi.org/10.1007/s43926-023-00049-y.

6. Zhang, D., Yin, C., Zeng, J., et al. "Combining structured and unstructured data for predictive models: a deep learning approach". *BMC Medical Informatics and Decision Making*. 2020; 20: 280. DOI: https://doi.org/10.1186/s12911-020-01297-6.

7. Large, J., Lines, J. & Bagnall, A. "The heterogeneous ensembles of standard classification algorithms (HESCA): The Whole is Greater than the Sum of its Parts". *arXiv*. 2017. DOI: https://doi.org/10.48550/arXiv.1710.09220.

8. Vichare, S. S. "Probabilistic ensemble machine learning approaches for unstructured textual data classification". *Purdue University Graduate School. Thesis*. 2024. DOI: https://doi.org/10.25394/PGS.25669425.v1.

9. An, Q., Rahman, S., Zhou, J. & Kang, J. "A comprehensive review on machine learning in healthcare industry". Classification, Restrictions, Opportunities and Challenges. *Sensors*. 2023; 23 (9): 4178. DOI: https://doi.org/10.3390/s23094178.

10. Lan, Q., Xu, X., Ma, H. & Li, G. "Multivariable data imputation for the analysis of incomplete credit data". *Expert Systems with Applications*. 2020; 41: 112926. DOI: https://doi.org/10.1016/j.eswa.2019.112926.

11. Uddin, S., Haque, I., Lu, H., et al. "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction". *Sci Rep*. 2022; 12: 6256. DOI: https://doi.org/10.1038/s41598-022-10358-x.

12. Stecking, R. & Schebesch K. D. "Clustering large credit client data sets for classification with SVM". 2009. – Available from: https://www.crc.business-school.ed.ac.uk/sites/crc/files/2023-10/Clustering-large-credit-client-data-sets-for-classification-with-SVM.pdf. – [Accessed: Jun 2024].

13. Fatima, S., Hussain, A., Amir, S. B., Ahmed, S. H. & Aslam, S. M. H. "XGBoost and random forest algorithms: An In-Depth analysis". *Pakistan Journal of Scientific and Industrial Research*. 2023; 3 (1): 26–31. DOI: https://doi.org/ 10.57041/pjosr.v3i1.946.

14. Koshti, R. M., Molia, S. J. & Varia D. J. "Improving credit score classification using predictive analysis and machine learning techniques". *International Journal on Science and Technology*. 2025; 16 (2): 1–12. DOI: https://doi.org/10.71097/IJSAT.v16.i2.4759.

15. Altunöz, U. “Prediction of banking credit risk using logistic regression and the artificial neural network models: a case study of english banks”. *Journal of Social Research and Behavioral Sciences*. 2024; 10 (21): 862–887. DOI: https://doi.org/10.52096/jsrbs.10.21.32.

16. Zhang, X., Lu, X. & Zhang, X. “A review of ensemble learning algorithms used in remote sensing applications”. *Appl. Sci*. 2022; 12 (17): 8654. DOI: https://doi.org/10.3390/app12178654.

17. “Autistic spectrum disorder screening data for adult”. – Available from: https://www.kaggle.com/datasets/faizunnabi/autism-screening. – [Accessed: Jun 2024].

18. “HELOC”. – Available from: https://www.kaggle.com/datasets/averkiyoliabev/home-equity-line-of-creditheloc?select=heloc_dataset_v1+%281%29.csv – [Accessed: Jun 2024].

19. “RAVDESS”. – Available from: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio. – [Accessed: Jun 2024].

20. “GTZAN”. – Available from: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification. – [Accessed: Aug 2024].

21. “The Fake-or-Real (FoR) Dataset (deepfake audio)”. – Available from: https://www.kaggle.com/datasets/mohammedabdeldayem/the-fake-or-real-dataset/data. – [Accessed: Jun 2024].

22. McFee, B., Raffel, C., Liang, D., et al. “librosa: Audio and music signal analysis in python.” In *Proceedings of the 14th Python in Science Conference*. 2015. p. 18–25. DOI: https://doi.org/10.25080/Majora-7b98e3ed-003.

23. Thukroo, I. A. Bashir, R. & Giri, K. J. “A comparison of cepstral and spectral featuresusing recurrent neural network for spoken language identification”. *Comput. Artif. Intell*. 2024; 2 (1): 440. DOI: https://doi.org/10.59400/cai.v2i1.440.

24. Sharma G., Umapathy K. & Krishnan S. “Trends in audio signal feature extraction methods”. *Applied Acoustics*. 2020; 158: 107020. DOI: https://doi.org/10.1016/j.apacoust.2019.107020.

25. McFee, B., McVicar, M., Faronbi, D., et al. “librosa/librosa: 0.10.2.post1 (0.10.2.post1)”. *Zenodo*. 2024. DOI: https://doi.org/10.5281/zenodo.11192913.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 30.07.2025
Received after revision 16.09.2025
Accepted 23.09.2025

DOI: https://doi.org/10.15276/hait.08.2025.18
УДК 004.1

Метод побудови ансамблевих класифікаторів структурованих та неструктурованих даних на базі уніфікованого підходу

Арсірій Олена Олександрівна[1)]
ORCID: https://orcid.org/0000−0001−8130−9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Андронаті Олександр Кирилович[1)]
ORCID: https://orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID: 58677655800
[1)] Національний університет “Одеська Політехніка”, пр. Шевченка, 1. Одеса, 65044, Україна

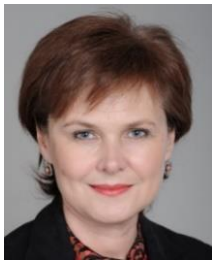
АНОТАЦІЯ

Ефективна класифікація гетерогенних типів даних, включаючи структуровані табличні дані та неструктуровані аудіосигнали, є надзвичайно важливою в таких різноманітних галузях, як охорона здоров'я, фінанси, інформаційна безпека та аналіз аудіоконтенту. Мета цього дослідження – розробити уніфікований підхід до побудови ансамблевих класифікаторів, здатних обробляти різноманітні формати даних в рамках єдиної структури, підвищуючи точність і надійність класифікації. Методологія інтегрує техніки попередньої обробки даних, вилучення ознак та стратегії нормалізації, які перетворюють гетерогенні набори даних у стандартизований числовий формат, придатний для ансамблевого навчання. Вісім базових

класифікаторів, що охоплюють традиційні алгоритми машинного навчання та глибокі нейронні мережі, включаючи K-найближчих сусідів, машини опорних векторів, випадковий ліс, XGBoost, логістичну регресію, багатошаровий перцептрон, згорткові нейронні мережі, мережі з довгою короткочасною пам'яттю, були навчені та оптимізовані за допомогою Grid Search Cross Validation. Для формування ансамблю було використано стекування з різними типами агрегації, такими як hard voting, soft voting та soft voting з нечітким ранжуванням Гомперца, для ефективного поєднання прогнозів моделі з урахуванням невизначеності та шуму. Експериментальна оцінка п'яти еталонних наборів даних – від медичної діагностики та оцінки кредитного ризику до розпізнавання емоцій у мовленні, класифікації музичних жанрів та виявлення синтетичного мовлення – демонструє постійне поліпшення точності та показників F1-score, з приростом до 8 відсотків порівняно з найкращими індивідуальними класифікаторами. Цей підхід виявляється особливо ефективним для неструктурованих аудіоданих, де часові та спектральні залежності становлять значні виклики. Результати підкреслюють універсальність і практичну цінність запропонованої уніфікованої методології ансамблю в вирішенні проблем дисбалансу класів, шуму та розмірності, пропонуючи масштабоване рішення, яке можна адаптувати до різних областей. Робота сприяє створенню комплексної структури, що полегшує розробку надійних класифікаторів для складних реальних даних, і відкриває шлях для майбутніх досліджень, що інтегрують гетерогенні джерела даних у цілісні прогнозні моделі.

Ключові слова: ансамблеві ласифікатори; машинне навчання; hard voting; soft voting; Gompertz function

ABOUT THE AUTHORS



Olena O. Arsirii - Doctor of Engineering Sciences, Professor, Head of the Department of Information Systems. Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID 54419480900
***Research field*:** Information technology; artificial intelligence; decision support systems; machine learning; neural networks

Арсірій Олена Олександрівна - доктор технічних наук, професор, завідувач кафедри Інформаційних систем. Національний університет "Одеська Політехніка", пр. Шевченка, 1. Одеса, 65044, Україна

Oleksandr K. Andronati - graduate student, Department of Information Systems. Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0009-0009-1794-5864; alex.andronati@gmail.com. Scopus Author ID 58677655800
***Research field*:** data science; machine learning; ensemble classfication

Андронаті Олександр Кирилович - аспірант кафедри Інформаційних систем. Національний університет "Одеська Політехніка", пр. Шевченка, 1. Одеса, 65044, Україна