Journal of King Saud University – Computer and Information Sciences 35 (2023) 757–774

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A comprehensive review on ensemble deep learning: Opportunities and challenges

Ammar Mohammed, Rania Kora

Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University, Cairo, Egypt

A R T I C L E   I N F O

Article history:
Received 18 November 2022
Revised 8 January 2023
Accepted 19 January 2023
Available online 1 February 2023

Keywords:
Ensemble learning
Ensemble methods
Machine learning
Deep learning
Ensemble deep learning

A B S T R A C T

In machine learning, two approaches outperform traditional algorithms: ensemble learning and deep learning. The former refers to methods that integrate multiple base models in the same framework to obtain a stronger model that outperforms them. The success of an ensemble method depends on several factors, including how the baseline models are trained and how they are combined. In the literature, there are common approaches to building an ensemble model successfully applied in several domains. On the other hand, deep learning-based models have improved the predictive accuracy of machine learning across a wide range of domains. Despite the diversity of deep learning architectures and their ability to deal with complex problems and the ability to extract features automatically, the main challenge in deep learning is that it requires a lot of expertise and experience to tune the optimal hyper-parameters, which makes it a tedious and time-consuming task. Numerous recent research efforts have been made to approach ensemble learning to deep learning to overcome this challenge. Most of these efforts focus on simple ensemble methods that have some limitations. Hence, this review paper provides comprehensive reviews of the various strategies for ensemble learning, especially in the case of deep learning. Also, it explains in detail the various features or factors that influence the success of ensemble methods. In addition, it presents and accurately categorized several research efforts that used ensemble learning in a wide range of domains.

Contents

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

E-mail addresses: ammar@cu.edu.eg (A. Mohammed), rania.kora@pg.cu.edu.eg (R. Kora)

https://doi.org/10.1016/j.jksuci.2023.01.014
1319-1578/

1. Introduction

In a world full of diverse and varied data sources. Machine learning has become one of the most important and dominant branches of artificial intelligence methods, which is applied in many fields. There are many different learning algorithms and methods. Each method's pitfalls and drawbacks are measured in terms of several factors, including performance and scalability. Based on a lot of research in machine learning, two methods dominate learning algorithms; namely deep learning (Deng et al., 2014) and ensemble learning (Polikar, 2012; Sagi and Rokach, 2018; Rokach, 2019). The deep learning techniques can scale and handle complex problems and offer an automatic feature extraction from unstructured data(Kamilaris and Prenafeta-Boldú, 2018). Also, deep learning methods contain several types of network architectures for different tasks, such as feeding forward neural networks (Bebis and Georgiopoulos, 1994), convolutional neural networks (Collobert and Weston, 2008), recurrent neural networks (Yu et al., 2019). Many others (Ain et al., 2017). However, the training process of deep learning models requires a massive effort, and tuning the optimal hyper-parameters requires expertise and extensive trial, which is a tedious and time-consuming task. Also, training more complex deep neural network increases the chance of overfitting.

Ensemble Learning, on the other hand, refers to a learning methodology that combines several baseline models to build a bigger single yet more powerful model than its constituents (Kumar et al., 2021). Also, ensemble learning can reduce the risk of overfitting thanks to the diversity of baseline models. Ensemble learning was successfully applied in various fields and domains and outperforms single models (Anwar et al., 2014; Shahzad and Lavesson, 2013; Prusa et al., 2015; Ekbal and Saha, 2011). There are several ensemble techniques varied in terms of how different baseline models are trained and combined. The most widely used ensemble techniques include averaging, bagging, random forest, stacking, and boosting. In the literature, there are many reviews about ensemble learning methods, and techniques (Krawczyk et al., 2017; Sagi and Rokach, 2018; Dong et al., 2020). Traditional ensemble learning is based on integrating traditional machine learning models and applying them in different fields (Tsai et al., 2011; Abellán and Mantas, 2014; Catal et al., 2015; Da Silva et al., 2014; Aburomman and Reaz, 2016). However, these efforts were limited to simple single models. In recent years, numerous attempts have been made to approach ensemble learning to deep learning (Haralabopoulos et al., 2020; Tasci et al., 2021; Alharbi et al., 2021; Ortiz et al., 2016; Can Malli et al., 2016; Xu et al., 2016). However, most of these attempts are articulated using the average voting method of baseline deep learning models. However, the ensemble process using average voting methods is biased towards weak baseline learners and is not a smart strategy for combing the baseline learners. Despite several strategies of combining baseline learners that can be applied to ensemble deep learning, these strategies have some limitations in terms of generalization, difficulties in training, and other issues (Tasci et al., 2021).In the literature, some review efforts have introduced the concept of deep ensemble learning(Dong et al., 2020; Sagi and Rokach, 2018). This effort, however, is restricted to the application of ensemble in particular domains with reviews on traditional ensemble approaches.

To this end, this paper tries to comprehensively review the different strategies for applying ensemble deep learning. It also presents several aspects that influence the success of ensemble methods, such as the type of utilized baseline learning models, the data samples techniques used in training, the diversity of employing different baseline classifiers, and the fusion methods of the baseline deep models. Additionally, it discusses the benefits and drawbacks are each strategy.

The contributions of this paper are highlighted as the following. First, we provide quantitative analytics insight into ensemble learning. Second, we introduce the fundamental concepts and general architecture of ensemble learning, strategies for generating diversity among the base classifiers, and the factors impacting any ensemble method. Additionally, we present the structure of each of the several ensemble methods and the advantages, disadvantages, and general classifications for each method. Moreover, we discuss the different strategies of ensemble deep learning models. Finally, we comprehensively survey numerous research efforts that used ensemble learning in various applications.

The remainder of this manuscript is organized as the following: Section 2 introduces quantitative analytics for research discussions on ensemble learning and deep learning techniques indexed in "Scopus." Section 3 introduces a comprehensive overview of the foundations of ensemble learning and the factors that influence any ensemble method. Section 4 provides an overview of various methods in ensemble learning and explains the general strategies of ensemble based on deep learning models. Section 5 discusses several criteria for evaluating different ensemble learning methods. Section 6 reviews several applications of ensemble learning in different domains. Finally, Section 7 concludes this paper and gives directions for future trends.

2. Trends of ensemble learning

Due to the strength and effectiveness of the ensemble learning system to improve the predictive performance of models. Ensemble learning has become an important research trend in recent years, which has led to an increase in the number of research used for ensemble learning in several domains of applications. Hence, this section presents this important trend in one of the most powerful databases, "Scopus". To show the extent to which the articles indexed published for ensemble learning increased each year and the different applied fields of ensemble learning from 2014 to 2021. The search query in this database is "Ensemble Learning" and "Ensemble Deep Learning." These were searched in the article titles, abstract, and keywords. Fig. 1 shows the number of articles published for the search term "Ensemble Learning" each year in the abovementioned period. The figure shows that the number of articles found using this term was estimated at 25,262, indicating an increase in the ensemble learning trend over several years. In addition, Fig. 2 shows the number of articles that discussed the search term "Ensemble Learning" in all fields. From the figure, it can be noted that the field of computer sciences has the highest estimated number of articles mentioned, estimated as 16,782 documents. Fig. 3 shows the number of articles published for the search term "Ensemble Deep Learning" each year in the abovementioned period. The figure shows that the number of articles found using this term was estimated as 6,173, indicating increased interest from researchers in this trend. Also, Fig. 4 shows the number of articles that discussed the search term "Ensemble Deep Learning in all fields. From the figure, it can be noted that the field of computer

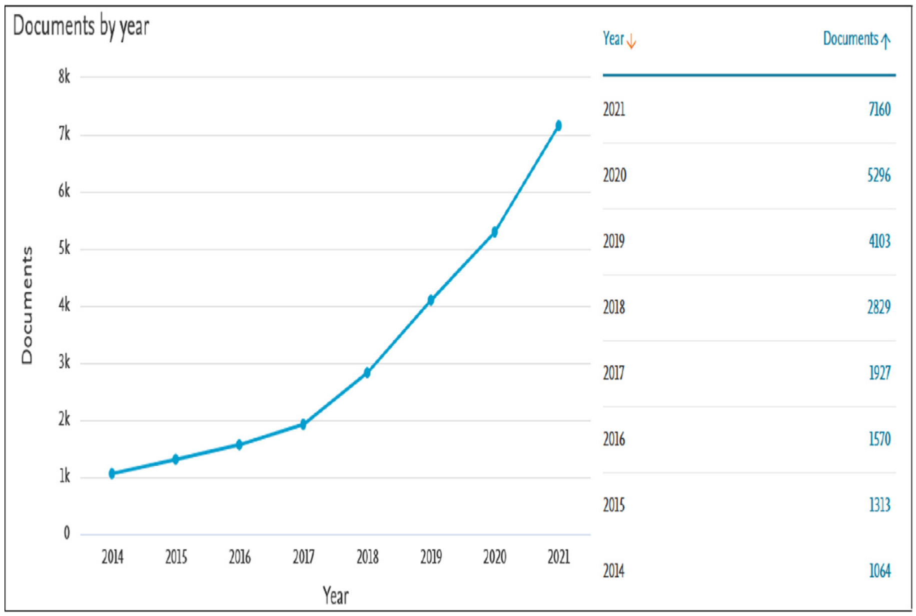

Fig. 1. The trends of search term "Ensemble Learning" in "Scopus" from 2014 to 2021 (Scopus, 2023).

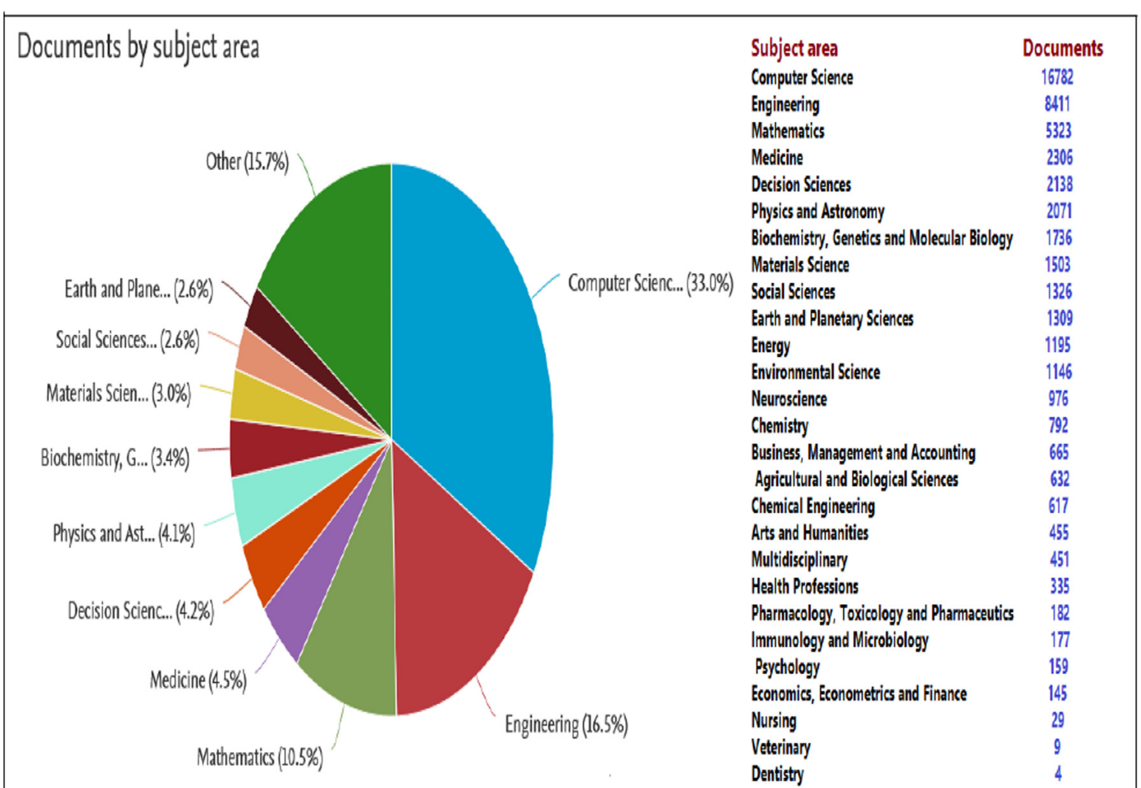

Fig. 2. The different fields of search term "Ensemble Learning" in "Scopus" from 2014 to 2021 (Scopus, 2023).

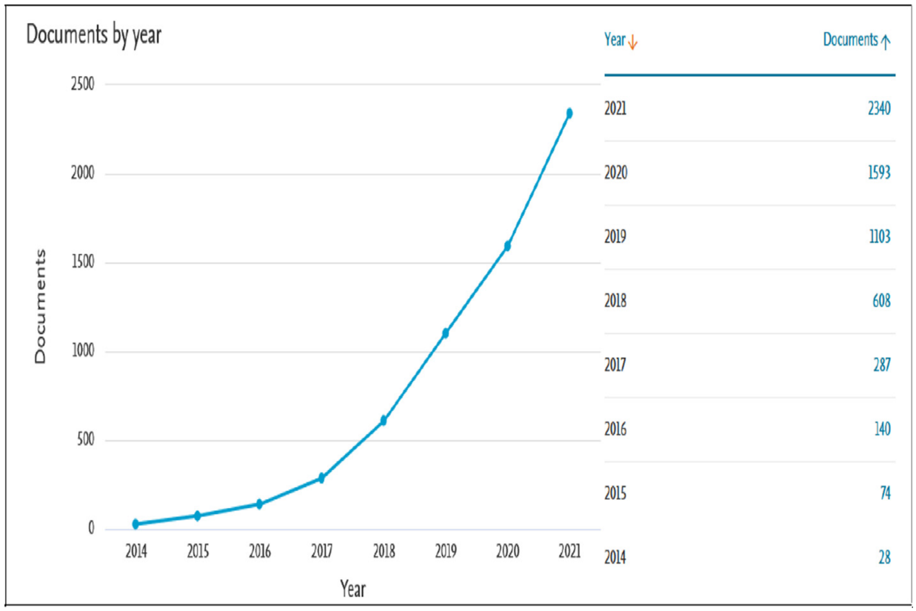

Fig. 3. The trends of search term "Ensemble Deep Learning" in "Scopus" from 2014 to 2021 (Scopus, 2023).

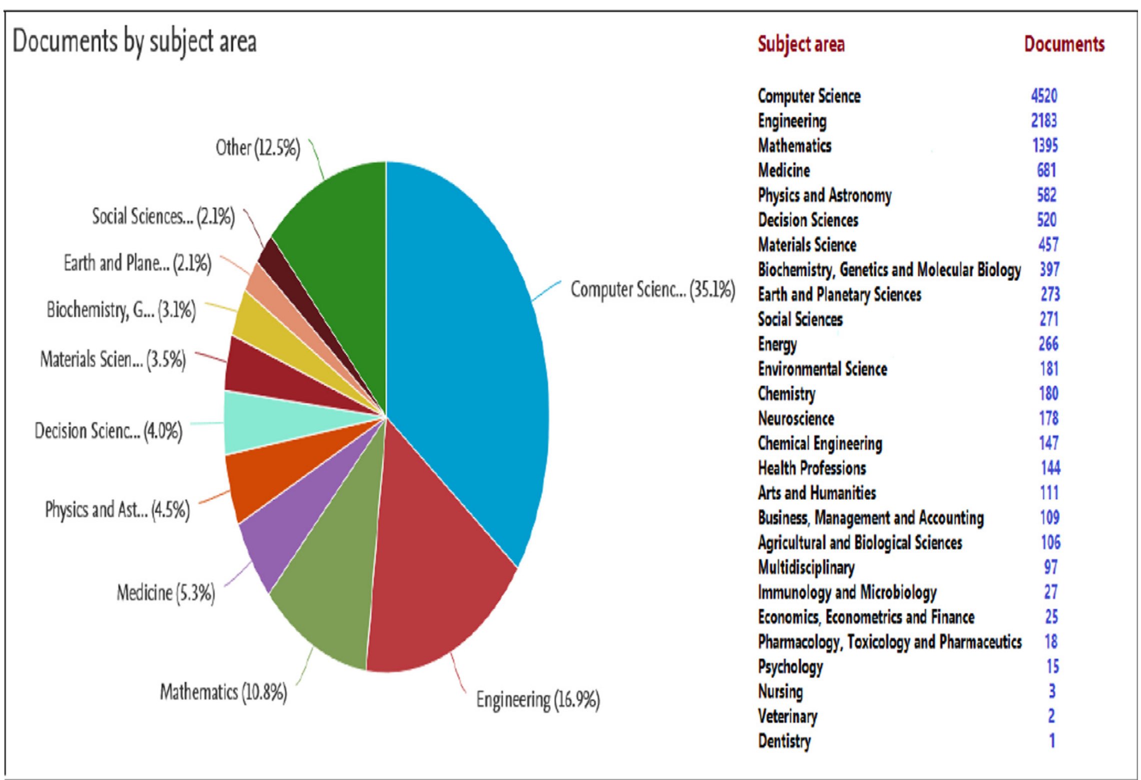

Fig. 4. The different fields of search term "Ensemble Deep Learning" in "Scopus" from 2014 to 2021(Scopus, 2023).

sciences has the highest estimated number of articles mentioned, estimated at 4520 documents.

According to the above statistical information, it is clear that research in ensemble learning and ensemble deep learning is growing faster each year due to its ability to improve prediction performance. According to estimates, the largest number of articles using "Ensemble Learning" and "Ensemble Deep Learning" in 2021 was estimated at 7160 and 2340 documents, respectively. In addition, ensemble learning and deep ensemble learning have been applied in several fields, especially computer science, with the highest utilization rate of ensemble learning and deep ensemble learning of 30% and 35.1%, respectively.

3. Foundations of ensemble learning

The general framework of any ensemble learning system is to use an aggregation function G to combine a set h of baseline classifiers, $c1, c2, \ldots, c_h$, towards predicting a single output. Given a dataset of size n and features of dimension

$m, D = \{(x_i, y_i)\}, 1 \leqslant i \leqslant n, x_i \in R^m$, the predication of the output based on this ensemble method is given by Eq. 1.

$$y_i = \phi(x_i) = G(c1, c2, \ldots, c_k) \quad (1)$$

Fig. 5 illustrates the general abstract framework of ensemble learning. All ensembles are made up of a collection of baseline classifiers (classifiers ensemble) that have been trained on input data that produce predictions that are combined to produce an aggregate prediction (Lakshminarayanan et al., 2017). Ensemble strategies differ on how to select the baseline classifiers that are trained. Two strategies generate diversity among the base classifiers based on their nature, either homogeneous or heterogeneous ensembles as shown in Fig. 6 (Seijo-Pardo et al., 2017). Homogeneous ensemble (da Conceição et al., 2015) consists of baseline classifiers of the same type, with each classifier based on different data. The feature selection method in this strategy is the same for different training data. The main difficulty in homogeneous form is the generation of diversity from the same learning algorithm. Whereas heterogeneous ensembles consist of different numbers of baseline classifiers, (da Conceição et al., 2016), as each classifier is based on the same data. In heterogeneous classifiers, the feature selection method is different for the same training data. Finally, homogeneous ensemble methods are more appealing to researchers since they are easier to understand and apply. Also, it is less costly to build homogeneous ensembles than heterogeneous ones (Hosni et al., 2019).

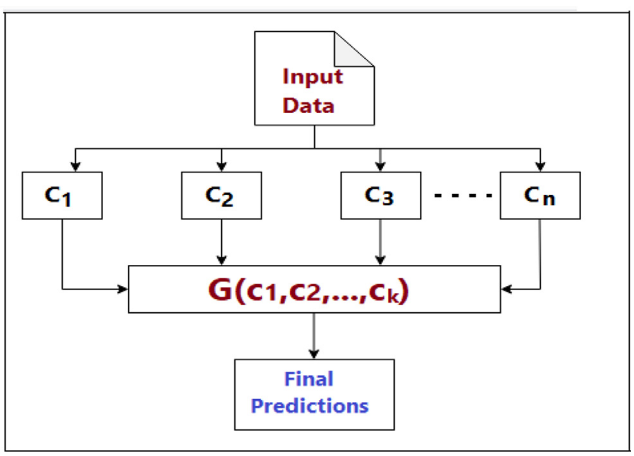

Fig. 5. General Framework of Ensemble.

Generally, any ensemble framework can be viewed and defined using three characteristics that affect its performance. The first one is the dependency on the trained baseline models, whether they are sequential or parallel. The second characteristic is the fusion methods, which involve choosing a suitable process for combining outputs of the baseline classifiers using different weight voting or meta-learning method. The third characteristic is the heterogeneity of the involved baseline classifiers, whether homogeneous or heterogeneous. Table 1 summarizes the characteristics of the popular ensemble methods. In what follows, those characteristics will be discussed in detail.

3.1. Data sampling

The selection of a data sampling method is one of the most important factors affecting the performance of the ensemble system. In the ensemble system, we need diversity in the data sampling decisions of the baseline classifiers. There are two strategies of the sampling methods from the training dataset in the ensemble system: the independent datasets strategy and the

Table 1
Categorization of ensemble methods.

Method	Dependent	Fusion method	Heterogeneity
Bagging	Parallel	Weight Voting	Homogenous
Random Forest	Parallel	Weight Voting	Homogenous
Boosting	Sequential	Weight Voting	Homogenous
AdaBoost	Sequential	Weight Voting	Homogenous
Gradient Boosting	Sequential	Weight Voting	Homogenous
Extreme Gradient Boosting	Sequential	Weight Voting	Homogenous
Stacking	Parallel	Meta Learning	Heterogeneous
Hybrid Ensemble	Both	Both	Heter/Homogeneous

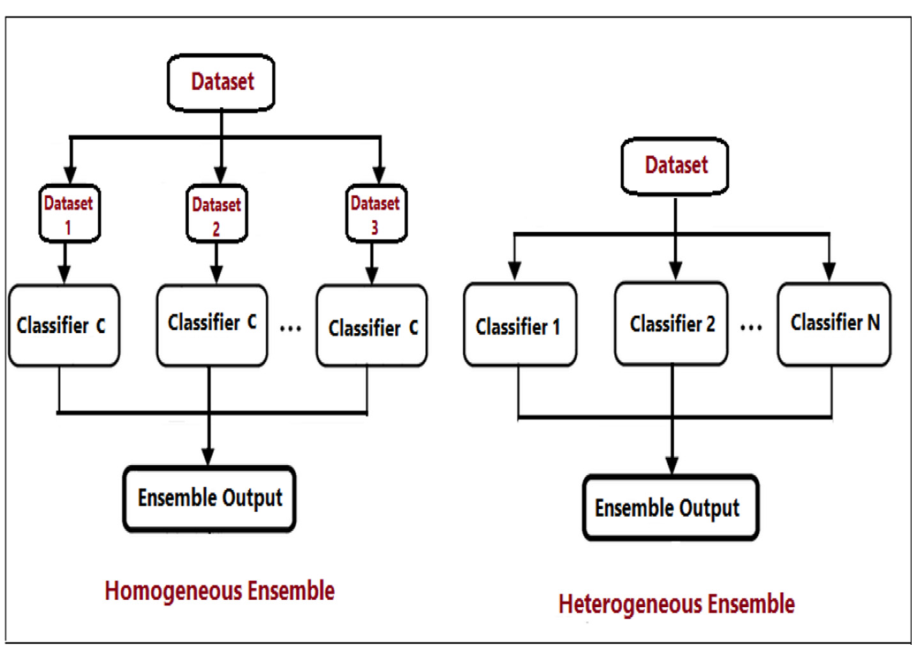

Fig. 6. General framework of homogeneous and heterogeneous ensemble.

dependent datasets strategy (Sagi and Rokach, 2018). In independent datasets strategy, (Ge et al., 2020), are those subsets that are not dependent on each other. By contrast, independent datasets strategy (Hassan et al., 2013) are subsets dependent on each other. The main advantage of using an independent datasets strategy is that its sub-data set is not affected by the performance of other sub-datasets, in contrast to using a dependent datasets strategy, where its sub-data set is affected by the results of the previous sub-data set. The difficulty of the data sampling method in both strategies is determining the optimal size of each data sample and the maximum number of samples. In addition, determining the appropriate strategy for data samples according to different ensemble methods(Lu and Van Roy, 2017).

3.2. Training baseline classifiers

The diversity of the baseline classifiers is the second influential factor in the ensemble system. At the core of any ensemble-based system are two techniques for training individual ensemble members: the sequential ensemble technique and the parallel ensemble technique (Huang et al., 2016). In sequential ensemble technique (Sultana et al., 2020), different learners learn sequentially because of data dependency. Thus, the errors made by the first model are sequentially corrected by the second model as shown in Fig. 7. So, the main advantage of sequential methods is to exploit the dependence between the base learners (Saeed et al., 2022). Whereas in parallel ensemble technique (Tang et al., 2020), base learners are generated simultaneously, as there is no data dependency. So, each data in the base learner is generated independently as shown in Fig. 8. This technique's basic advantage is exploiting the independence between base learners. Thus, the errors made by one model differ from those found in another independent model, allowing the ensemble model to calculate the average out the errors (Valle et al., 2010).

3.3. Fusion method

Output fusion refers to integrating the outputs of the baseline classifiers into a single output. There are two methods of fusion, the voting method, and the meta-learning method. We will explain in each method how to implement in integrating the outputs of baseline classifiers, their advantages, and the difficulty of applying them, as well as select the appropriate fusion method for each of the ensemble methods. The fusion methods can be used with independent or dependent data samples and can also be used with parallel or sequential baseline classifiers.

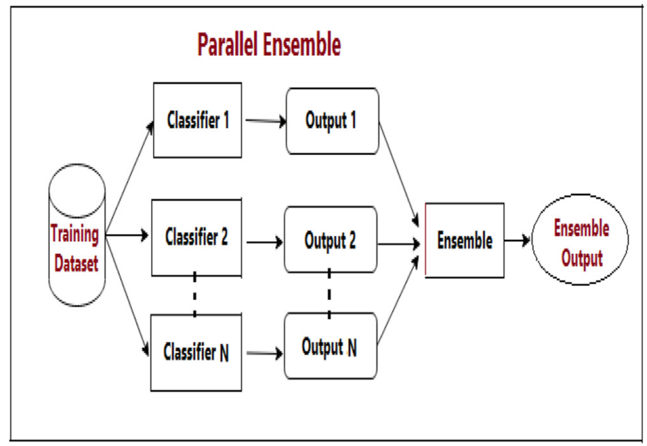

Fig. 8. General framework of parallel ensemble.

3.3.1. Voting method

Voting methods are generally used in classification or regression problems to improve predictive performance. In addition, voting methods are the appropriate integrating method for bagging and boosting methods. The first fusion method is a voting ensemble, which includes three methods: max voting, averaging voting, and weighted average voting. We will discuss in each voting method the nature of implementation and the advantages and drawbacks of implementation it.

1. **Max Voting:** The first and most popular voting method is the max voting (Kim et al., 2003) often, often known as majority voting or hard voting. The idea of max voting involves collecting predictions for each class label and predicting the class label with the most votes as shown in function (2). For example,

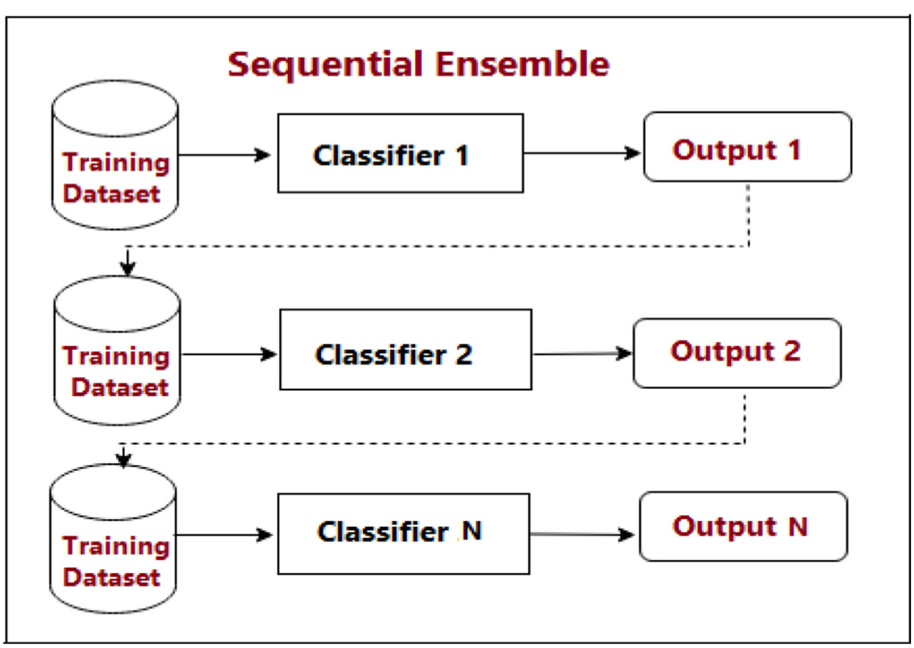

Fig. 7. General framework of sequential ensemble.

assuming we combine three classifiers, C1, C2, and C3, that assign the following classifications to a training sample: [0,0,1] becomes y^* =mode [0,0,1]=0. We would categorize the sample as "class 0". Max voting is often used in the bagging method. Another type of max voting is soft voting. Soft voting involves collecting predicted probabilities for each class label and predicting the class label with the largest probability as shown in function (3). Max voting is distinguished from soft voting in that once we know the prediction for any of the baseline classifiers, we do not need to store any other information about the probability distributions of the predictions. On the other hand, soft voting needs to store and use all the distribution values, making it more computationally and costly for storage. However, in soft voting, we can use various methods to calculate the prediction, such as calculating maximum or average probability values (Delgado, 2022). In general, the max voting method has the advantages of being simple to understand and the simplest method of voting. The drawbacks of the max voting method include the computational expense of using several baseline models. Additionally, max voting is useless when the baseline classifiers predictions are the same results and may not fit all problems (Nti et al., 2020).

$$y^* = \mathrm{mod}[C_1(x), C_2(x), .., C_n(x)] \tag{2}$$

Where y^* a predict the class label via majority (plurality) voting of each classifier C_n.

$$y^* = \underset{i}{argmax} \sum_{j=1}^{n} w_j P_{ij} \tag{3}$$

Where w_j is the weight that can be assigned to the j^{th} classifier.

2. **Averaging Voting:** The second voting method is the averaging voting (Montgomery et al., 2012). The idea of averaging voting is that predictions are extracted from multiple models, and an average of the predictions is used to make the final prediction. Average prediction is calculated using the arithmetic mean, which is the sum of the predictions divided by the total predictions made as shown in function (4). For instance, suppose the ensemble of classifiers contained three members: C1(x)=[0.9,0.1], C2(x)=[0.2,0.8], and C3(x)=[0.6,0.4]. The mean prediction would be as follows: to calculate the class 0 y_0^* [0.9 + 0.2 + 0.6/3] = 0.566. And to calculate the class 1 y_1^* [0.1 + 0.8 + 0.4 /3] = 0.433, would yield a prediction $y^* = 0$. The average voting method has the advantage of being the strongest from the point of view of predictive power. In addition, it is more accurate in performance than majority voting and reduces overfitting. Also, the average voting is a natural competitor to the max voting for bagging method. The drawbacks of the average voting method include being computationally more expensive than the max voting method, as it requires averaging the prediction results of all the baseline models. One limitation of the averaging voting method is that it assumes that all baseline models in the ensemble are equally effective. However, it is not the case as some models may be better than others (Hopkinson et al., 2020).

$$y^* = \underset{i}{argmax} \frac{1}{n} \sum_{j=1}^{m} w_{ij} \tag{4}$$

where w_{ij} is the probability of the i^{th} class label of the j^{th} classifier.

3. **A weighted Average Voting:** The third method of voting is the weighted average voting, which is a slightly modified version of averaging voting (Latif-Shabgahi, 2004). The idea of weighted average voting is different weights given to the baseline learners, indicating the importance of each model in prediction. By multiplying each prediction by the weight of the classifiers to produce a weighted sum and then dividing the result by the sum of the weights of the classifier, these weights may be used to calculate the weighted average for each class 0 or class 1 as shown in function (5). For instance, suppose the ensemble of classifiers contained three members: C1(x)=[97.2,2.8], C2(x)=[100.0,0], and C3(x)=[95.8,4.2]. It has constant weights for ensemble members [0.84, 0.87, 0.75]. To calculate the class 0 y_0^* = ((97.2 * 0.84) + (100.0 * 0.87) + (95.8 * 0.75))/ (0.84 + 0.87 + 0.75) =97.763. And to calculate the class 1 y_1^* = ((2.8 * 0.84) + (0 * 0.87) + (4.2 * 0.75))/ (0.84 + 0.87 + 0.75) =2.235, would yield a prediction $y^* = 0$. The weighted average voting method is more accurate than the simple average-voting method. The challenge in using a weighted average ensemble is choosing each member's relative weighting. Also, the computation is more expensive than the average voting method, as it requires calculating the weighted average of the prediction results of all the baseline models, which makes it of little application (Khan et al., 2020).

$$y^* = \frac{\sum_{j=1}^{m} w_j x_i}{\sum_{j=1}^{m} w_j} \tag{5}$$

where w weighted average, m is a number of terms to be averaged, weights applied to x values w_j, and data values to be averaged x_j.

3.3.2. Meta learning method

The second fusion method is meta-learning (Soares et al., 2004), also known as "learning to learn", which is the process of learning from learners. The term "meta-learning" covers learning based on previous experience with other tasks. Therefore, it is used to improve the performance and results of a learning algorithm by changing some aspects of the learning algorithm based on experiment results. The meta-learning method differs from traditional machine-learning models in that it involves more than one learning stage where the individual inducer outputs serve as an input to the meta-learner that generates the final output (Kuruvayil and Palaniswamy, 2021).

Over the past five years, interest in meta-learning has increased, especially after 2017. With the increased use of advanced machine learning algorithms, the difficulties of training these learning algorithms have led to an increased interest in meta-learning. Machine learning algorithms have many challenges, such as the high operational costs due to many experiments during the training phase, which takes a long time to find the best model that achieves the best performance for a certain dataset. Meta-learning helps to meet these challenges by improving learning algorithms and finding learning algorithms that perform better (Kuruvayil and Palaniswamy, 2022). In addition, the benefits of meta-learning include speeding up learning processes by reducing the number of experiments required, helping learning algorithms better adapt to changing conditions, and optimizing hyperparameters to achieve optimal results. Moreover, this method provides an opportunity to tackle many challenges of deep learning, including data size, computational complexities, and generalization. The challenge in meta-learning is to learn from experience in a systematic, data-driven manner (Hospedales et al., 2021). There are many meta-learning methods, the most common of which is stacking (Haghighi and Omranpour, 2021). To implement the meta-learning, there are several challenges represented in defining an appropriate meta-learning approach and the computation time complexity, whether through a large amount of available dataset or through multiple baseline models or multiple levels of meta-learning (Monteiro et al., 2021).

4. Ensemble methods

This section presents two aspects. The first aspect includes the structure of the most popular ensemble learning methods and lists each method's benefits, drawbacks, and implementation challenges separately. The second aspect presents the idea of deep ensemble learning and the advantages of its application compared to traditional ensemble learning. It also discusses the deep learning challenges that ensemble deep learning overcomes them. Moreover, it introduces the different strategies for applying ensemble deep learning and the advantages of each strategy with an explanation of the factors that can affect its performance.

4.1. Common ensemble methods

Three popular ensemble learning methods can be used to improve the machine learning process: bagging, boosting, and stacking. We will discuss the nature of each method's work and its characteristics regarding the nature of data generation, the nature of training of baseline classifiers, and the appropriate fusion methods. In addition, the benefits, drawbacks, and implementation challenges of each method will be covered.

4.1.1. Bagging

The bagging method (Breiman, 1996), also known as bootstrap aggregating, is a completely data-specific algorithm. It refers to creating multiple small subsets of data from the actual dataset. The goal of bagging is to create more diverse predictive models by adjusting a stochastic distribution of the training datasets, where small changes in the training data set will lead to significant changes in the model predictions. Bagging is shorthand for the combination of bootstrapping and aggregating. In bootstrapping, the training of the ensemble models on bootstrap replicates the training dataset. In aggregation, the final result is achieved by majority voting of the model's predictions performed to determine the final prediction. Bagging offers the advantage of reducing variance, thus eliminating overfitting. It also performs well on high-dimensional data. The drawback of bagging is that it is computationally expensive and has high bias, and it also leads to a loss of interpretability of a model (Bühlmann and Yu, 2002). Random Forests (RF) algorithm (Breiman, 2001) is a good example of bagging. There are several challenges to implementing the bagging method: determining the optimal number of base learners and subsets and the maximum number of bootstrap samples per subset. In addition, the determine of fusion method of integrating the outputs of the base classifiers from various voting methods. In summary, the bagging method uses parallel ensemble techniques where baseline learners are generated simultaneously, as there is no data dependency and the fusion methods depend on different voting methods. The function of bagging is shown as follows (6):

$$f(x) = \frac{1}{B}\sum_{B=1}^{B} f_{b(x)} \tag{6}$$

where $f_{b(x)}$ weak learners, $\frac{1}{B}$ generates bootstrapping sets.

4.1.2. Boosting

Boosting method was first presented by Freund and Schapire in the year 1997 (Freund et al., 1996), and is a sequential process where each subsequent model attempts to correct the errors of the previous model. Boosting consists of sequentially multiple weak learners in a very adaptive way, whereby each model in the sequence is fitted, giving more importance to observations in the dataset that the previous models in the sequences badly handled. Boosting, like bagging, can be used for regression and classification problems. Boost algorithms include three types, namely, Adaptive Boosting (AdaBoost) (Freund et al., 2003), Stochastic Gradient Boosting (SGB) (Friedman, 2001), and Extreme Gradient Boosting (XGB), also known as XGBoost(Friedman et al., 2000). Several studies have applied various types of boosting. For example, the AdaBoost algorithm is implemented in Sun et al. (2016) for noise detection and in Asbai and Amrouche (2017) for speech feature extraction. The XGB algorithm is implemented in Haumahu et al. (2021) for Fake news classification. The SGB algorithm is implemented in Shin (2019) for early prediction of safety accidents at construction sites. Boosting provides ease of interpretation of the model and helps reduce variance and bias in a machine learning ensemble. The drawback of boosting is that each classifier must fix the errors in the predecessors. To implement boosting, several challenges are represented by the difficulty of scaling sequential training in boosting. It is computationally costly and more vulnerable to overfitting when increasing the number of iterations. Finally, it can be noted that boosting algorithms can be slower to train when compared to bagging because a large number of parameters can also affect the behavior of the model. In summary, the boosting method uses sequential ensemble techniques where different learners learn sequentially, as there is data dependency and the fusion methods depend on different voting methods. The function of boosting is shown as follows (7):

$$f(x) = \sum_{t} \alpha_t h_t(x) \tag{7}$$

where creates a strong classifier $f(x)$ from several weak classifiers $h_t(x)$. This is done by building a model from the training data, then creating a second model that attempts to correct the errors from the first model α_t.

4.1.3. Stacking

Stacking method (Smyth and Wolpert, 1997), also known as Stacked Generalization, is a model ensembling technique used to combine information from multiple predictive models to generate a new model (meta-model). The architecture of a stacking model involves two or more base models, referred to as a level-0 model, and a meta-model that combines the predictions of the base models, referred to as a level-1 model. In level 0 models (base models), models fit on the training data and whose predictions are compiled. However, in the level 1 model (meta-model), the model learns how to combine the base models' predictions best. The outputs from the base models used as input to the meta-model may be probability values, or class labels in the case of classification (Ma et al., 2018). The stacking method typically performs better than all trained models. For instance, a stacking ensemble learning system was proposed by Divina et al. (2018) to forecast electric energy usage in Spain and Qiu et al. (2014) to forecast electric energy usage in Australia. Stacking has the benefit of a deeper comprehension of the data, making it more precise and effective. Overfitting is a major issue with model stacking because there are so many predictors that all predict the same target that is merged. In addition, multi-level stacking is costly to data (as lots of data needed to be trained) and time-consuming (as each layer adds multiple models) (Xiong et al., 2021). Xiong et al. (2021). To implement stacking, several challenges are represented by identifying the appropriate number of baseline models and the baseline models that can be relied upon to generate better predictions from datasets when designing a stacking ensemble from scratch. Also, the difficulty of interpreting the final model and the computation time complexity are added when the amount of available data grows exponentially. A highly complex model would take months to run. Finally, the problem of multi-label classification raises many issues, such as overfitting and the curse of dimensionality, from the high dimensionality of the data (Chatzimparmpas et al., 2020). In summary, the stacking method uses parallel ensemble techniques where baseline learners are generated simultaneously, as there is no data dependency and the fusion methods depend on the meta-learning method. The function of stacking is shown as follows (8):

$$f_s(x) = \sum_{i=1}^{n} a_i f_i(x) \tag{8}$$

A formal stacking concept: Here, we make predictions from several models $(m1, m2, m3..., mn)$ to build a new model, where the new model is used to make predictions on the test dataset. Stacking seeks to increase the predictive power of a model. The basic idea of stacking is to "stack" the predictions of $(m1, m2, m3..., mn)$ by a linear combination of weights $a_j, ..., (i = 1, 2, ..., n)$.

4.2. Ensemble deep learning

In recent years, deep learning or deep neural learning has led to a series of achievements in various tasks(Arel et al., 2010). Deep learning architectures have shown great success in almost all challenges related to machine learning across different areas, such as NLP (Mohammed and Kora, 2019; Elnagar et al., 2020), computer vision (Haque et al., 2020; Brunetti et al., 2018), speech recognition (Jaouedi et al., 2020; Noda et al., 2015). Machine translation (Popel et al., 2020; Popel et al., 2020). Deep neural network models are nonlinear methods that learn through a stochastic training algorithm. This means that it is highly flexible, able to learn the complex relationships between variables and approximate any mapping function. The downside to this flexibility is that the models need a higher variance. The high variance of the deep model can be addressed by ensemble deep learning approach opportunities by training multiple deep models for the problem and combining their predictions. Hence, ensemble deep learning methods refer to training several baseline deep models and combining some rules to make predictions. Ensemble deep learning aims to effectively combine the major benefits of several deep learning models with those of an ensemble learning system (Mohammed and Kora, 2021). Despite the power of ensemble deep learning system methods in improving prediction performance, most of the ensemble deep learning literature focuses on only applying a majority of voting algorithms to enhance the performance due to its simplicity.

Ensemble learning based on deep learning models is more difficult than ensemble learning based on traditional classifiers due to deep neural networks containing millions to billions of hyper-parameters that need a lot of time and space to train multiple base deep learners. Thus, hyper-parameters are challenges in the application of ensemble deep learning techniques. Ensemble learning strategies are formed in the context of manipulating the data level or the baseline model level. In manipulation at the level of data, by sampling data or cross-validation data (re-sampling) to create new training sets to train different base learners. In manipulation at the level of basic models, deep learning is distinguished by more diverse strategies than traditional or machine learning, which is the possibility of reducing the number of hyper-parameters used in the ensemble base deep models by selecting the same model and changing the hyper-parameters (Saleh et al., 2022). Fig. 9 shows four strategies through which deep learning can be conducted based on the ensemble represented by: (A) Applying many different basic models using the same data. (B) Applying different structures of the same basic model using the same data. (C) Applying many different basic models using many different data samples. (D) Applying different structures of the same basic model using many different data samples. Comparing these strategies shows that strategy *A* and strategy *C* are compatible with deep learning models and traditional learning techniques. Whereas strategy *B* and strategy *D* only apply to deep learning models and cannot be used with traditional learning techniques, making the ensemble deep learning strategies diverse. In addition, strategy *B* and strategy *D* enable ensemble deep learning to reduce the hyper-parameters of the baseline deep models by different structures of the same basic model by altering some of the hyper-parameters values. In addition to these strategies, the strength of the ensemble deep learning system depends on the ensemble system design, from identifying the most effective deep learning models to address the problem and determining the appropriate

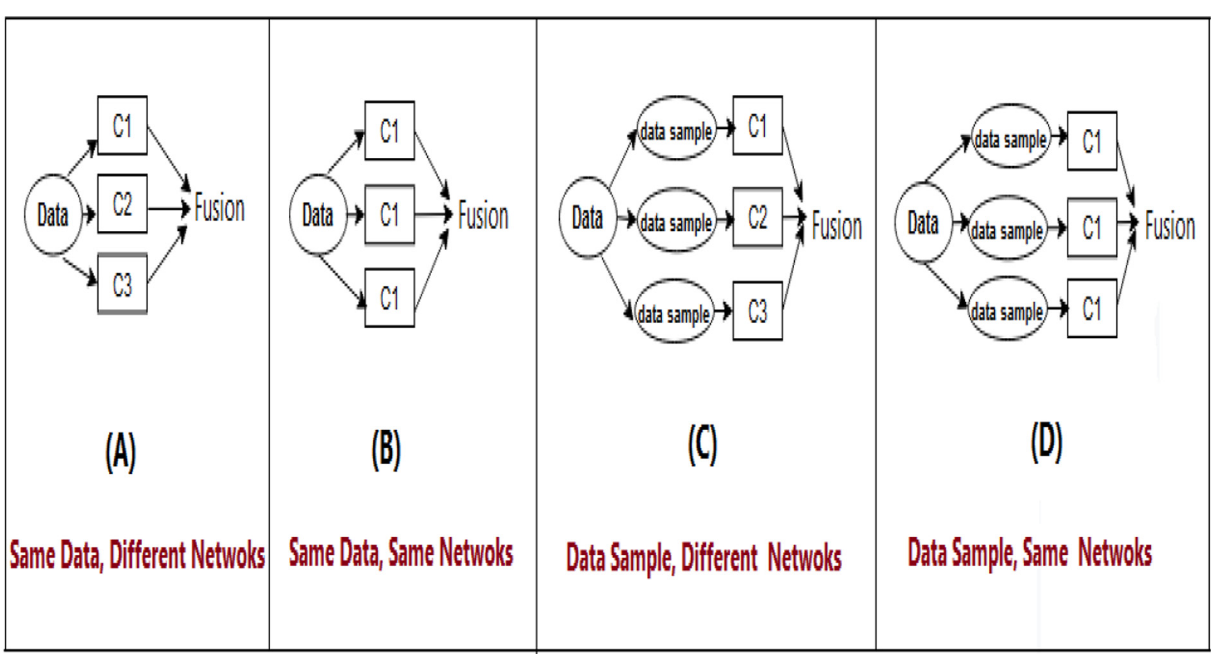

Fig. 9. Different cases of ensemble deep learning.

number of baseline deep learning models, such as three or more and also determining the optimal ratio for data splitting such as (80–20 or 70–30 or 60–40). Moreover, we consider factors that may affect the deep ensemble system, such as defining the nature of data generation, training deep baseline models, and deciding the most appropriate fusion method of combining outputs of the baseline classifiers, as previously mentioned. These three factors affect the general framework of the ensemble system.

5. Evaluating ensembles

With the emergence of ensemble learning approaches, lots of research has been conducted to evaluate the methods of ensemble (Hashino et al., 2007; Zhang et al., 2016; Das and Sengur, 2010; Hosni et al., 2019). The evaluation is crucial to determining the effectiveness of a certain ensemble method. There are several criteria for evaluation ensemble, including predictive performance. Other criteria, such as the computational complexity or the comprehensibility of the generated ensemble, can also be important. In the following, we summarize the different evaluation criteria of ensemble learning.

5.1. Predictive performance

Predictive performance metrics have always been the primary criterion for choosing the performance of classifiers. Also, predictive performance measures are considered objective and quantifiable, so they are often used to benchmark machine learning algorithms practically. The first step to applying predictive performance is to use a suitable dataset. The holdout technique is a typical approach for measuring predictive performance where the given dataset is randomly divided into two subsets: training and test sets. Other versions of the holdout method might be utilized. It is normal procedure to resample data, which means dividing it into training and test sets in different ways. Two common resampling methods include random subsampling, and n-fold cross-validation (Dai, 2013).

There are common measures for evaluating an ensemble model. Accuracy is one of the popular and simplest metrics, which as defined in Eq. 9:

$$Accuracy = \frac{number\ of\ true\ predictions}{total\ number\ of\ prediction} \tag{9}$$

In some cases, accuracy is insufficient and can be deceptive in evaluating an ensemble model with imbalanced class distributions. In the latter scenario, other measures can be used as alternative measures, such as Recall, Precision, Specificity, and F-Measure (Kadam et al., 2019).

Recall, also known as sensitivity, measures the ensemble model's capability to identify positive samples, which as defined in Eq. 10:

$$Recall = \frac{true\ positive}{positive} \tag{10}$$

where true positive denotes the number of true positive observations and positive denotes the number of positive observations.

Another well-known performance metric is precision. It quantifies how many instances classified as positive are actually positive. Formally,the precision equation is defined as 11:

$$Precision = \frac{true\ positive}{true\ positive\ + false\ positive} \tag{11}$$

Likewise, specificity measures how well the model identifies negative samples. The equation is defined as 12:

$$Specificity = \frac{true\ negative}{negative} \tag{12}$$

where true negative denotes the number of true negative observations and negative denotes the number of negative observations.

There is commonly a trade-off between precision and recall metrics. Attempting to enhance one measure often results in the fall of the second. Thus, F-Measure quantifies this trade-off by calculating the harmonic mean of both precision and recall. More specifically, this measure is defined in Eq. 13:

$$F - Measure = \frac{2\ x\ Precision\ x\ Recall}{Precision\ +\ Recall} \tag{13}$$

5.2. Computational complexity

The computational complexity of the ensemble approach is an additional essential aspect to consider. Generally, the computational cost refers to the amount of CPU time required by each ensemble model. The computational cost is distributed on two complexity metrics: The computational cost of training and creating the ensemble model and the computational cost of predicting a new instance: The computational cost of the prediction is relatively small compared to the computation cost of the training ensemble. Thus, this metric should be addressed. In terms of memory, a smaller ensemble model needs less memory to keep its components. Furthermore, smaller ensembles perform faster prediction.

5.3. Other criteria

In addition to computational complexity and prediction accuracy, other considerations may be made when selecting the best ensemble method. These criteria include Interpretability, Scalability, usability, and robustness of the ensemble model. Interpretability (Carvalho et al., 2019) refers to the ability of a user to understand the ensemble outcomes. However, interpretability is typically a subjective metric. One of the many quantitative metrics and indicators that can help us evaluate this criterion is the compactness metric. Compactness in the ensemble can be evaluated using the number of classifiers involved and the complexity of each classifier.

On the other hand, scalability refers to the capacity of the ensemble approach to construct a classification model given large amounts of data. Independent ensemble methods are considered more scalable than dependent methods, as the classifier involved in the ensemble approach can be trained in parallel. Usability is another metric that assesses the user's preference for comprehending how to adjust the ensemble models they employ. Broadly speaking, a good ensemble method should contain a comprehensive set of control parameters that can be easily adjusted.

6. Application domains

This section highlights applications of ensemble learning across different domains, using either traditional or deep learning as baseline classifiers. In general, we briefly summarize the baseline classifiers applied, the ensemble techniques used, and the domain used in their experiments.

6.1. Applications of traditional ensemble learning

This part discusses applications of traditional ensemble learning in various domains, including image classification, natural language processing (NLP), and others. Table 2 summarizes some works that presented ensemble learning methods in machine

Table 2
Applications of ensemble learning in machine learning approach.

Studies	Baseline Classifiers	Fusion Method	Domain
Shipp and Kuncheva (2002)	NB	Voting	Medical Image
Stamatatos and Widmer (2002)	SVM	Voting	Music Recognition
Cho and Won (2003)	SVM,KNN	Voting	Medical Image
Wilson et al. (2006)	DT	Boosting	English Sentiment
Tsutsumi et al. (2007)	SVM, ME	Stacking	English Sentiment
Abbasi et al. (2008b)	SVM	Boosting	Arabic Sentiment
Li et al. (2010)	SVM, LR	Voting	English Sentiment
Lu and Tsou (2010)	NB, ME, SVM	Stacking	Chinese Sentiment
Xia et al. (2011)	NB, ME, SVM	Stacking	English Sentiment
Ekbal and Saha (2011)	SVM, NB, ME	Voting	Named Entity Recognition
Li et al. (2012)	SVM, KNN	Stacking	Chinese Sentiment
Su et al. (2012)	ME, SVM	Voting, Stacking	Chinese Sentiment
Hassan et al. (2013)	SVM	Boosting	English Sentiment
Rodriguez-Penagos et al. (2013)	SVM	Voting	English Sentiment
Clark and Wicentwoski (2013)	NB	Voting	English Sentiment
Anifowose et al. (2013)	RF	Bagging	Petroleum Reservoir
Shahzad and Lavesson (2013)	NB, DT, KNN	Voting	Malware Detection
Wang et al. (2013)	SVM	Voting	Image Classification
Cortes et al. (2014)	DT	AdaBoost	Medical Image
Kuznetsov et al. (2014)	DT, LR	AdaBoost	Medical Image
Fersini et al. (2014)	ME, SVM, NB	Voting,Bagging	English Sentiment
Wang et al. (2014)	SVM, KNN, DT, ME, NB	Bagging, Boosting	English Sentiment
Da Silva et al. (2014)	SVM, RF,LR	Voting	English Sentiment
Anwar et al. (2014)	KNN, DT, RF, LR	Bagging	Medical Image
Bharathidason and Venkataeswaran (2014)	RF	Voting, Bagging	Medical Image
Zareapoor and Shamsolmoali (2015)	NB, KNN, SVM	Bagging	Credit Card Fraud Detection
Kanakaraj and Guddeti (2015)	NB, SVM	Bagging, Boosting	English Sentiment
Prusa et al. (2015)	KNN, SVM, LR	Bagging, Boosting	English Sentiment
Bashir et al. (2015)	SVM, LR	Voting, Bagging	Medical Image
Bashir et al. (2015)	SVM, DT	Voting	Medical Image
Mishra and Mishra (2015)	NB	Voting	Medical Image
Kang et al. (2015)	SVM	Bagging, Boosting	Medical Image
Xia et al. (2016)	SVM, LR	Voting	English Sentiment
Perikos and Hatzilygeroudis (2016)	NB, ME	Bagging	English Sentiment
Fersini et al. (2016)	NB, DT, SVM	Voting	English Sentiment
Onan et al. (2016)	BLR,NB,LDA,LR, SVM	Stacking, AdaBoost,Bagging	English Sentiment
Araque et al. (2017)	NB, ME,SVM	Voting	English Sentiment
Dedhia and Ramteke (2017)	NB, SVM, ME	Adaboost	English Sentiment
Oussous et al. (2018)	MNB, SVM, ME	Voting, Stacking	Moroccan Dialect Sentiment
Saleena et al (2018)	SVM, RF,NB, LR	Voting	English Sentiment
Sharma et al. (2018)	SVM	Bagging	English Sentiment
Fouad et al. (2018)	SVM,NB,LR	Voting	English Sentiment
Kulkarni et al. (2018)	SVM,NB,RF	Voting	Text Classification
Livieris et al. (2019)	KNN, DT	Voting,Bagging	Medical Image
Chen et al. (2019)	FLDA	Bagging	Groundwater Potential Analysis
Erdoğan and Namlı (2019)	SVM	Voting, Stacking	A living environment Analysis
Seker and Ocak (2019)	RF, LR, Linear R	Bagging	Roadheaders Performance Analysis
Alrehili and Albalawi (2019)	NB, SVM	Voting, Bagging,Boosting	English Sentiment
Pasupulety et al. (2019)	SVM, RF	Stacking	India's Sentiment
Cai et al. (2020)	SVM, LR	Voting	Chloride Concentration Analysis
Saeed et al. (2022)	SVM, NB, LR, DT, KNN	Voting, Stacking	Arabic Sentiment

learning in different fields. In the image classification domain, the researchers in Wang et al. (2013) applied voting based on SVM for image retrieval using COREL images database (Liu et al., 2011). In particular, in medical image classification, the researchers in Cortes et al. (2014) suggested boosting based on deep decision tree (DT) for image classification using several breast cancer datasets. The researchers in Kuznetsov et al. (2014) used AdaBoost based on DT for multi-class classification using 8 UCI datasets (Fernández-Delgado et al., 2014). The researchers in Livieris et al. (2019) applied voting and bagging based on kNN and DT to classify lung abnormalities from chest X-rays using three benchmark datasets (Kermany et al., 2018). The researchers in Anwar et al. (2014) proposed bagging based on many classifiers (KNN, DT, RF, and LR) using seven datasets from various diseases (such as Cancer, Diabetes, Heart disease, Sonar, etc.). The researchers in Bharathidason and Venkataeswaran (2014) applied voting and bagging based on RF using heart disease dataset (Makhtar et al., 2012). The researchers in Shipp and Kuncheva (2002) proposed voting based on NB using Breast Cancer dataset (Antoniou et al., 2000). The researchers in Mishra and Mishra (2015) applied voting-based NB using six medical image benchmark datasets (Leukemia, Breast cancer, Lung cancer, Hepatitis, Lymphoma, and Embryonal tumors). The researchers in Cho and Won (2003) applied voting based on SVM and KNN using three Leukemia cancer datasets. In Bashir et al. (2015) applied voting and bagging based on SVM and LR using five heart disease datasets. That same year, Bashir et al. (2015) applied voting based on SVM and DT using breast cancer diagnosis datasets. The researchers in Kang et al. (2015) proposed two ensemble methods (bagging and boosting) based on SVMs for the treatment of patients' diabetes using dataset (Li and Maguire, 2010).

In addition, in the NLP domain for the English language, the authors in Wang et al. (2014) used two popular ensemble methods (Bagging, Boosting) based on five base learners (NB, ME, DT, KNN, SVM) by ten public sentiment analysis datasets. The authors in Xia et al. (2011) used stacking based on three algorithms, namely NB, ME, and SVM, by five datasets. The authors in Li et al. (2010), Xia et al. (2016) applied a voting method based on both LR and SVM using reviews extracted from Amazon.com.(Rushdi-Saleh et al., 2011). The authors in Araque et al. (2017) applied voting methods based on different machine classifiers (NB, ME, and SVM) by even public datasets from movie reviews. The authors in Alrehili and Albalawi (2019) suggested three ensemble methods (voting, bagging, and boosting) based on NB and SVM using English customer reviews datasets (Alrehili and Albalawi, 2019). The authors in Saleena et al (2018) applied voting based on different baseline classifiers (SVM, RF, NB, and LR) by several English tweets datasets. The authors in Dedhia and Ramteke (2017) used Adaboost based on three classifiers (NB, SVM, and ME) using several English tweets datasets. The authors in Perikos and Hatzilygeroudis (2016) applied bagging based on NB and ME using different English news portals datasets. The authors in Fersini et al. (2016) used voting based on NB, DT, and SVM by English Movie Reviews datasets (Chen et al., 2012). The authors in Onan et al. (2016) proposed three ensemble methods (bagging, AdaBoost, and stacking) based on five classifiers (BLR, NB, LDA, LR, and SVM) using nine public English sentiment analysis datasets from different domains (Whitehead and Yaeger, 2009). The authors in Kanakaraj and Guddeti (2015) suggested bagging and boosting based on both NB and SVM using English movie review (Pang and Lee, 2005). The authors in Fersini et al. (2014) proposed voting and bagging based on different baseline classifiers (ME, SVM, and NB) by several English movie and product reviews datasets (Täckström and McDonald, 2011; Pang and Lee, 2005. The authors in Prusa et al. (2015) applied KNN, SVM, and LR based on both bagging and boosting using English sentiment140 corpus (Go et al., 2009). The authors in Wilson et al. (2006) introduced boosting based on a DT classifier by English MPQA Corpus (Wiebe et al., 2005). The authors in Tsutsumi et al. (2007) applied stacking based on two classifiers (SVM and ME) using the English movie review dataset (Pang and Lee, 2005). The authors in Hassan et al. (2013) proposed boosting based on SVM using three English product review forum datasets (Abbasi et al., 2010; and Abbasi et al., 2008a. The authors in Fouad et al. (2018) compared the performance of a voting method based on three classifiers (SVM, NB, and LR) using several English tweets datasets. The authors in Rodriguez-Penagos et al. (2013) introduced voting based on SVM by English SemEval 2013 dataset (Dzikovska et al., 2013). The authors in Clark and Wicentwoski (2013) suggested voting based on NB using the English SemEval-2013 dataset (Nakov et al., 2016). The authors in Da Silva et al. (2014) applied voting-based four baseline classifiers (SVM, RF, and LR) using several English tweets datasets. But, in multiclass sentiment classification, (Sharma et al., 2018) proposed a bagging based on SVM using several English movie review datasets. In contrast, in the Arabic language, the authors in Saeed et al. (2022) applied both voting and stacking for spam detection based on five baseline classifiers (SVM, NB, LR, DT, KNN) using two datasets from Opinion Spam Corpus (Li et al., 2011). Besides, in the different dialects, the authors in Su et al. (2012) applied both voting and stacking based on two algorithms (ME and SVM) using two datasets for three domains of Chinese reviews (book, hotel, and notebook). The authors in Li et al. (2012) suggested stacking based on SVM and KNN using several Chinese food review datasets. The authors in Lu and Tsou (2010) applied stacking based on three classifiers NB, ME, and SVM, using the Chinese dataset (Seki et al., 2008). The authors in Pasupulety et al. (2019) introduced stacking based on two baseline classifiers (SVM and RF) for predicting stock prices of companies using India's National Stock Exchange (NSE) datasets (Kumar and Misra, 2018). The authors in Oussous et al. (2018) proposed voting and stacking based on three baseline classifiers (MNB, SVM, and ME) using the Moroccan tweets dataset (Tratz et al., 2013). The authors in Ekbal and Saha (2011) suggested voting based on diverse classification methods such as SVM, ME, and RF for named entity recognition using three Indian languages (Bengali, Hindi, and Telugu) by using Bengali news corpus (Ekbal and Bandyopadhyay, 2008). The authors in Abbasi et al. (2008b) proposed a boosting based on SVM using several middle eastern web forums.

Moreover, in the diverse fields, in Stamatatos and Widmer (2002) used a voting method based on SVM for music performer recognition using several pianists playing datasets. In Chen et al. (2019) applied the bagging method based on Fisher's linear discriminant function (FLDA) for potential groundwater assessment at the Ningtiaota area in Shaanxi, China. They used using a database with 66 groundwater spring locations. In Zareapoor and Shamsolmoali (2015) suggested Bagging based on three machine algorithms: SVM, NB, and KNN for credit card fraud predicting. They use 100,000 records of credit card transactions dataset (Hormozi et al., 2013). In Shahzad and Lavesson (2013) proposed voting based on NB, DT, and KNN for malware detection using three datasets of malicious threat (Shahzad et al., 2010). In Anifowose et al. (2013) applied bagging RF to predict petroleum reservoir properties using six datasets from a giant carbonate reservoir in the Middle East and a drilling site in the Northern Marion platform of North America (Helmy et al., 2010). In Kulkarni et al. (2018) suggested voting based on SVM, NB, and RF for a crop recommendation system using the input soil dataset into the recommendable crop type, Kharif and Rabi. In Erdoğan and Namlı (2019), applied voting and stacking based on SVM for a living environment prediction. In Cai et al. (2020), voting based on SVM and LR was applied to predict surface chloride concentration. In Seker and Ocak (2019) proposed a bagging based on three classifiers (RF, LR, and Linear R) to predict road headers using several datasets.

6.2. Applications of ensemble deep learning

Ensemble learning methods in deep learning applications outperform traditional ensemble learning in many domains, including image classification, natural language processing (NLP), and others. Table 3 summarizes some works that presented ensemble learning methods in deep learning in different fields. In the image classification domain, in Wang et al. (2020) applied stacking method based on multiple CNNs using CIFAR-10 dataset (Pandit and Kumar, 2020). Also, in Zhang et al. (2019) applied of stacking method based on multiple CNNs used for Image Deblurring. They used GoPro dataset (Marques et al., 2021) and the Video Deblurring dataset (Wu et al., 2020). In Waltner et al. (2019) proposed boosting method based on CNN used for image retrieval by the biggest available retrieval datasets. In Chen et al. (2019) and Chen et al. (2018) proposed the deep boosting framework by integrating the CNN into the boosting algorithm. They used two benchmark datasets (Set12 and BSD68) (Thakur et al., 2019). In Can Malli et al. (2016) suggested voting based on CNNs for apparent age estimation "face detection" using IMDB-WIKI dataset (Russakovsky et al., 2015). In Opitz et al. (2017) applied Boosting CNNs using several image retrieval datasets(Liu et al., 2016). In Mosca and Magoulas (2016) applied boosting CNN by using two image datasets; namely, MNIST (LeCun, 1998), and CIFAR-10 (Pandit and Kumar, 2020). In Walach and Wolf (2016) proposed boosting CNNs for object counting in images using different image datasets, namely mall crowd counting (Chen et al., 2012). UCF 50 crowd counting (Idrees et al., 2013), UCSD (Chan et al., 2008). In Moghimi et al. (2016) applied boosting CNNs using several image datasets, namely (Cars (Krause et al., 2013) and Aircrafts (Gosselin et al., 2014)). In Yang et al. (2015) proposed boosting CNNs for face detection using imageNet dataset (Krizhevsky et al., 2012). In Li et al. (2015) suggested stacking based on simpli-

Table 3
Applications of ensemble learning in deep learning approach.

Studies	Baseline Classifiers	Fusion Method	Domain
Tur et al. (2012)	DCN	Stacking	Semantic Utterance Classification
Deng et al. (2012)	DCN	Stacking	Spoken Language Understanding
Liu et al. (2014)	DNN	Boosting	Facial Expression Recognition
Palangi et al. (2014)	RNN	Stacking	Speech Recognition
Deng and Platt (2014)	RNN, CNN	Stacking	Speech Recognition
Yang et al. (2015)	CNN	Boosting	Face Detection
Li et al. (2015)	SNNM	Stacking	Image Classification
Ortiz et al. (2016)	DBN	Voting	Medical Image
Can Malli et al. (2016)	CNN	Voting	Image Classification
Xu et al. (2016)	CNN,LSTM	Voting	English Sentiment
Deriu et al. (2016)	CNN	Stacking	English Sentiment
Walach and Wolf (2016)	CNN	Boosting	Image Classification
Kumar et al. (2016)	CNN	Stacking	Image Classification
Moghimi et al. (2016)	CNN	Boosting	Image Classification
Han et al. (2016)	CNN	Boosting	Facial Recognition
Liu et al. (2017)	BPNN	Stacking	Flood Forecasting
Codella et al. (2017)	CNNs, DRN	Voting	Medical Image
Chen et al. (2017)	CNN_RNN	Voting	Text Classification
Opitz et al. (2017)	CNN	Boosting	Image Retrieval
Mosca and Magoulas (2016)	CNN	Boosting	Image Classification
Akhtyamova et al. (2017)	CNN	Voting	English Sentiment
Araque et al. (2017)	CNN,LSTM,GRU	Voting, Stacking	English Sentiment
Chen et al. (2018)	CNN	Boosting	Image Denoising
Heikal et al. (2018)	CNN, LSTM	Voting	Arabic Sentiment
Zhang et al. (2019)	CNN	Stacking	Deblurring Image
Waltner et al. (2019)	CNN	Boosting	Image Retrieval
Chen et al. (2019)	CNN	Boosting	Image Denoising
Wang et al. (2019)	DNN	Adaboost	Security Level Classification
Alshazly et al. (2019)	CNN	Voting	Medical Image
Cha et al. (2019)	CNN	Voting	Medical Image
Al-Omari et al. (2019)	Bi_LSTM	Voting	Fake News
Nguyen and Le Nguyen (2019)	CNN, LSTM	Voting	English Sentiment
Ali et al. (2020)	DNN	Boosting	Medical Image
Guo et al. (2020)	CNN,RetinaNet,Deep SVDD	Voting	Medical Image
Khamparia et al. (2020)	CNN	Voting	Medical Image
Zhang et al. (2020)	CNN, LSTM	Boosting	Computer Vision,NLP
Zhang et al. (2020)	GNet, SNet	Boosting, Stacking	Robotic arm control "Reinforcement"
Wang et al. (2020)	CNN	Stacking	Image Classification
Haralabopoulos et al. (2020)	LSTM,GRU,CNN,RCNN,DNN	Voting, Stacking	English Sentiment
Mohammed and Kora (2021)	6 Models	Hybrid Ensemble	Multilingual Text Classification
Tasci et al. (2021)	CNN	Voting	Medical Image
Alharbi et al. (2021)	LSTM, GRU	Voting	Arabic Sentiment
Livieris et al. (2020)	CNN	Bagging, Boosting	English Text
Mohammadi and Shaverizade (2021)	CNN,LSTM, GRU, Bi_LSTM	Stacking	English Sentiment

fied neural network module (SNNM) using four face image datasets (Jiang et al., 2013). In Zhang et al. (2020) applied a boosting method on CIFAR-10 dataset (Pandit and Kumar, 2020) containing 60000 colored images to train the CNN. In particular, in medical image classification, the authors of Ali et al. (2020) applied a smart healthcare system for heart disease prediction using ensemble deep learning and feature fusion approaches. The proposed system achieved an accuracy of 98.5%. The authors of Alshazly et al. (2019) suggested voting based on CNNs for visual recognition tasks (ear recognition) using several ear datasets.

The authors of Ortiz et al. (2016) applied voting based on deep belief networks using a large dataset from the Alzheimer's disease Neuroimaging Initiative (ADNI) (Hinrichs et al., 2009). The authors of Codella et al. (2017) proposed voting based on residual networks (DRN) and CNNs for melanoma recognition in dermoscopy images. The voting method achieved an accuracy of 76% by using the dermoscopic images dataset (containing 1279 images) (Mendonca et al., 2015). The authors of Tasci et al. (2021) applied voting based on CNNs for tuberculosis detection by two TB CXR image datasets (Sharma et al., 2017). The voting method achieved an accuracy of 97.5% and 97.69% accuracy rates on datasets, respectively. The authors of Cha et al. (2019) suggested voting based on nine CNNs to classify eardrum and external auditory canal features. The voting achieved an average accuracy of 93.67% by using a large database of 910,544 images(Locketz et al., 2016). The authors of Guo et al. (2020) proposed a voting method for automated cervical precancer screening using 30,000 images from several datasets. The voting method combined the assessment of three deep learning architectures, RetinaNet, Deep SVDD, and CNN. The average accuracy and F-score of 91.6% and 0.89%, respectively. The authors of Khamparia et al. (2020) applied a voting method based on CNNs for disease prediction related to neuromuscular disorders using two neuromuscular disorder datasets (Bakay et al., 2006).

In addition, in the NLP domain, in Mohammed and Kora (2021) proposed a novel ensemble for multilingual text classification using six benchmark datasets. Also, compare the performance of the proposed and other ensemble methods. The results prove that the proposed method outperforms the state-of-art ensemble methods. In Deng et al. (2012) suggested a stacking method based on deep convex network (DCN) to spoken language understanding (SLU) problems. The stacking method achieved an accuracy of 91.88% by using the ATIS dataset (consists of 5871 sentences) (Wen et al., 2005). The authors in Xu et al. (2016) proposed a soft voting ensemble based on CNN and LSTM using SemEval 2013 dataset (Dzikovska et al., 2013). In Chen et al. (2017) presented voting based on the CNN_RNN model using a large documents dataset (Lewis et al., 2004). In Akhtyamova et al. (2017) suggested a voting method based on CNNs for predicting drug safety using

English reviews from health forums (Karimi et al., 2015). In Araque et al. (2017) applied both voting and stacking based on several deep learning models, namely CNN, LSTM, and GRU, using seven English movie review datasets. In Al-Omari et al. (2019) applied voting based on Bi_LSTM for English fake news detection using NLP4IF 2019 (Barrón-Cedeno et al., 2019). In Nguyen and Le Nguyen (2019) applied voting based on CNN and LSTM using five English datasets from movie reviews (Koh et al., 2010). In Livieris et al. (2020) proposed CNNs based on bagging and stacking using several English review datasets. In Haralabopoulos et al. (2020) applied both voting and stacking based on several deep learning models, namely LSTM, GRU, CNN, RCNN, and DNN, using two English tweets datasets (SemEval (Bethard et al., 2016), Toxic Comment (van Aken et al., 2018)). In Mohammadi and Shaverizade (2021) applied stacking based on four deep learning models, namely CNN, LSTM, GRU, and BiLSTM using English review dataset (SemEval) (Bethard et al., 2016). In Deriu et al. (2016) proposed stacking ensemble based on CNN for English tweets classification by using SemEval-2016 dataset (Bethard et al., 2016). In contrast, in Heikal et al. (2018) applied voting based on the combination of CNN and LSTM models using Arabic dataset (ASTD) (Nabil et al., 2015). In Alharbi et al. (2021) applied a voting method based on LSTM and GRU using five datasets from Arabic tweets.

Moreover, in the diverse fields, in Zhang et al. (2020) proposed a system that jointly learns the grasping and the stacking policies through the grasping for stacking network (GSNet) for enables a robotic arm to correctly pick boxes from a table and put it on a platform. In Wang et al. (2019) proposed an Adaboost method based on DNN for security level classification. The dataset is the assessment results of 100 Android terminals (including smartphones, smart bracelets, tablet PC) and from schools, hospitals, factories, and other environments. In Liu et al. (2014) applied boosted deep belief network for facial expression recognition/shape changes based on the CK + database (contains 327 expression images) (Seyyedsalehi and Seyyedsalehi, 2014). The authors of Deng and Platt (2014) applied the stacking method based on both RNN and CNN for speech recognition using TIMIT dataset (Garofolo et al., 1993). The authors of Liu et al. (2017) applied stacking based on back propagation neural networks (BPNN) for flood forecasting. Han et al. (2016) applied boosting CNNs for recognizing facial action units. In Tur et al. (2012) applied a stacking method based on deep convex networks (DCNs) to semantic utterance classification by the dataset of utterances from the users of a spoken dialog system. In Palangi et al. (2014) applied stacking RNN for speech recognition systems based on TIMIT dataset (Garofolo et al., 1993).

7. Conclusion

In machine learning, reducing the bias and the variance of models is one of the key factors determining the success of the learning process. In the literature, it has been proven that merging the output of different classification algorithms might decrease the generalization error without increasing the variance of the model. The previous is the key essence of the so-called ensemble learning. Numerous research efforts have preferred ensemble learning over single-model learning in various domains. The main advantage of ensemble learning is combining several individual models to improve prediction performance and obtain a stronger model that outperforms them. In the literature, there are several ensemble techniques to boost classification algorithms. The main difference between any two ensemble methods is training the baseline models and how to combine them. Several research efforts introduced ensemble learning into deep learning models to remedy the problems appearing during the learning process of deep learning models. Usually, the main challenge of deep learning models is that they need a lot of knowledge and experience to tune the optimal hyperparameters aiming at reaching a global minimum error. However, finding the optimal hyperparameters requires an exhausting technique in the search space, which in turn becomes a tedious and time-consuming task. Thus, several research efforts have applied deep ensemble learning in many fields, and most of these efforts are articulated around simple ensemble methods. This paper provided a comprehensive review of the various strategies for ensemble learning, especially in the case of deep learning. The paper also illustrated the recent trends in ensemble learning using quantitative analysis of several research papers. Moreover, the paper offered various factors that influence ensemble methods' success, including sampling the training data, training the baseline models, and the fusion techniques of the baseline models. Also, the papers discussed the pros and cons of each ensemble method. Additionally, the paper extensively introduced and presented several research efforts that used ensemble learning in a wide range of domains and categorized these efforts into either traditional machine or deep learning models as baseline classifiers. It is worth noting that an ensemble of deep learning models using simple averaging methods is not a smart choice and is very sensitive to biased baseline models. On the other hand, Injecting diversity in ensemble deep learning can become robust to the biased baseline models. The diversity can be achieved by training different baseline deep learning architectures over several data samples. The diversity, however, is limited by the computation cost and the availability of suitable data to be sampled.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbasi, A., Chen, H., Salem, A., 2008a. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Trans. Informat. Syst. (TOIS) 26 (3), 1–34.

Abbasi, A., Chen, H., Thoms, S., Fu, T., 2008b. Affect analysis of web forums and blogs using correlation ensembles. IEEE Trans. Knowledge Data Eng. 20 (9), 1168–1180.

Abbasi, A., France, S., Zhang, Z., Chen, H., 2010. Selecting attributes for sentiment classification using feature relation networks. IEEE Trans. Knowl. Data Eng. 23 (3), 447–462.

Abellán, J., Mantas, C.J., 2014. Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. Expert Syst. Appl. 41 (8), 3825–3830.

Aburomman, A.A., Reaz, M.B.I., 2016. A novel svm-knn-pso ensemble method for intrusion detection system. Appl. Soft Comput. 38, 360–372.

Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. Int. J. Adv. Comput. Sci. Appl. 8 (6), 424.

Akhtyamova, L., Ignatov, A., Cardiff, J., 2017. A large-scale cnn ensemble for medication safety analysis In: International Conference on Applications of Natural Language to Information Systems. Springer, pp. 247–253.

Alharbi, A., Kalkatawi, M., Taileb, M., 2021. Arabic sentiment analysis using deep learning and ensemble methods. Arabian J. Sci. Eng. 46 (9), 8913–8923.

Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M., Kwak, K.-S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Informat. Fusion 63, 208–222.

Al-Omari, H., Abdullah, M., AlTiti, O., Shaikh, S., 2019. Justdeep at nlp4if 2019 task 1: Propaganda detection using ensemble deep learning models. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, pp. 113–118.

Alrehili, A., Albalawi, K., 2019. Sentiment analysis of customer reviews using ensemble method, pp. 1–6.

Alshazly, H., Linse, C., Barth, E., Martinetz, T., 2019. Ensembles of deep learning models and transfer learning for ear recognition. Sensors 19 (19), 4139.

Anifowose, F., Labadin, J., Abdulraheem, A., 2013. Ensemble model of artificial neural networks with randomized number of hidden neurons. In: 2013 8th International Conference on Information Technology in Asia (CITA). IEEE, pp. 1–5.

Antoniou, A.C., Gayther, S.A., Stratton, J.F., Ponder, B.A., Easton, D.F., 2000. Risk models for familial ovarian and breast cancer. Genetic Epidemiol.: Off. Publ. Int. Genetic Epidemiol. Soc. 18 (2), 173–190.
Anwar, H., Qamar, U., Muzaffar Qureshi, A.W., 2014. Global optimization ensemble model for classification methods. Sci. World J. 2014.
Araque, O., Corcuera-Platas, I., Sánchez-Rada, J.F., Iglesias, C.A., 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Syst. Appl. 77, 236–246.
Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. IEEE Comput. Intell. Mag. 5 (4), 13–18.
Asbai, N., Amrouche, A., 2017. Boosting scores fusion approach using front-end diversity and adaboost algorithm, for speaker verification. Comput. Electr. Eng. 62, 648–662.
Bakay, M., Wang, Z., Melcon, G., Schiltz, L., Xuan, J., Zhao, P., Sartorelli, V., Seo, J., Pegoraro, E., Angelini, C., et al., 2006. Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of rb–myod pathways in muscle regeneration. Brain 129 (4), 996–1013.
Barrón-Cedeno, A., Da San Martino, G., Jaradat, I., Nakov, P., 2019. Proppy: A system to unmask propaganda in online news. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 9847–9848.
Bashir, S., Qamar, U., Khan, F.H., 2015. Bagmoov: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Austral. Phys. Eng. Sci. Med. 38 (2), 305–323.
Bashir, S., Qamar, U., Khan, F.H., 2015. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. Quality Quantity 49 (5), 2061–2076.
Bebis, G., Georgiopoulos, M., 1994. Feed-forward neural networks. IEEE Potentials 13 (4), 27–31.
Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., Verhagen, M., 2016. Semeval-2016 task 12: Clinical tempeval. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 1052–1062.
Bharathidason, S., Venkataeswaran, C.J., 2014. Improving classification accuracy based on random forest model with uncorrelated high performing trees. Int. J. Comput. Appl 101 (13), 26–30.
Breiman, L., 1996. Bagging predictors. Machine Learn. 24 (2), 123–140.
Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.
Brunetti, A., Buongiorno, D., Trotta, G.F., Bevilacqua, V., 2018. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing 300, 17–33.
Bühlmann, P., Yu, B., 2002. Analyzing bagging. Annals Stat. 30 (4), 927–961.
Cai, R., Han, T., Liao, W., Huang, J., Li, D., Kumar, A., Ma, H., 2020. Prediction of surface chloride concentration of marine concrete using ensemble machine learning. Cem. Concr. Res. 136, 106164.
Can Malli, R., Aygun, M., Kemal Ekenel, H., 2016. Apparent age estimation using ensemble of deep learning models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 9–16.
Carvalho, D.V., Pereira, E.M., Cardoso, J.S., 2019. Machine learning interpretability: A survey on methods and metrics. Electronics 8 (8), 832.
Catal, C., Tufekci, S., Pirmit, E., Kocabag, G., 2015. On the use of ensemble of classifiers for accelerometer-based activity recognition. Appl. Soft Comput. 37, 1018–1022.
Cha, D., Pae, C., Seong, S.-B., Choi, J.Y., Park, H.-J., 2019. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 45, 606–614.
Chan, A.B., Liang, Z.-S.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–7.
Chatzimparmpas, A., Martins, R.M., Kucher, K., Kerren, A., 2020. Stackgenvis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. IEEE Trans. Visual Comput. Graphics 27 (2), 1547–1557.
Chen, L., Wang, W., Nagarajan, M., Wang, S., Sheth, A., 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 6, no. 1, pp. 50–57.
Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting. Bmvc 1 (2), 3.
Chen, G., Ye, D., Xing, Z., Chen, J., Cambria, E., 2017. Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 2377–2383.
Chen, C., Xiong, Z., Tian, X., Wu, F., 2018. Deep boosting for image denoising. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–18.
Chen, W., Pradhan, B., Li, S., Shahabi, H., Rizeei, H.M., Hou, E., Wang, S., 2019. Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis. Nat. Resour. Res. 28 (4), 1239–1258.
Chen, C., Xiong, Z., Tian, X., Zha, Z.-J., Wu, F., 2019. Real-world image denoising with deep boosting. IEEE Trans. Pattern Anal. Machine Intell. 42 (12), 3071–3087.
Cho, S.-B., Won, H.-H., 2003. Machine learning in dna microarray analysis for cancer classification. In: Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003-Volume 19, pp. 189–198.
Clark, S., Wicentwoski, R., 2013. Swatcs: Combining simple classifiers with estimated accuracy. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 425–429.
Codella, N.C., Nguyen, Q.-B., Pankanti, S., Gutman, D.A., Helba, B., Halpern, A.C., Smith, J.R., 2017. Deep learning ensembles for melanoma recognition in dermoscopy images. IBM J. Res. Dev. 61 (4/5), pp. 5–1.
Collobert, R., Weston, J., 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167.
Cortes, C., Mohri, M., Syed, U., 2014. Deep boosting. In: International Conference on Machine Learning. PMLR, pp. 1179–1187.
da Conceição, L.R., da Costa, C.E., Rocha, G.N.d., Pereira-Filho, E.R., Zamian, J.R., 2015. Ethanolysis optimisation of jupati (raphia taedigera mart.) oil to biodiesel using response surface methodology. J. Brazil. Chem. Soc.26, 1321–1330.
da Conceição, L.R.V., Carneiro, L.M., Rivaldi, J.D., de Castro, H.F., 2016. Solid acid as catalyst for biodiesel production via simultaneous esterification and transesterification of macaw palm oil. Ind. Crops Prod. 89, 416–424.
Dai, Q., 2013. A competitive ensemble pruning approach based on cross-validation technique. Knowl.-Based Syst. 37, 394–414.
Das, R., Sengur, A., 2010. Evaluation of ensemble methods for diagnosing of valvular heart disease. Expert Syst. Appl. 37 (7), 5110–5115.
Da Silva, N.F., Hruschka, E.R., Hruschka Jr, E.R., 2014. Tweet sentiment analysis with classifier ensembles. Decis. Support Syst. 66, 170–179.
Dedhia, C., Ramteke, J., 2017. Ensemble model for twitter sentiment analysis. In 2017 International Conference on Inventive Systems and Control (ICISC). IEEE, pp. 1–5.
Delgado, R., 2022. A semi-hard voting combiner scheme to ensemble multi-class probabilistic classifiers. Appl. Intell. 52 (4), 3653–3677.
Deng, L., Platt, J., 2014. Ensemble deep learning for speech recognition. In: Proc. Interspeech.
Deng, L., Tur, G., He, X., Hakkani-Tur, D., 2012. Use of kernel deep convex networks and end-to-end learning for spoken language understanding. 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 210–215.
Deng, L., Yu, D., et al., 2014. Deep learning: methods and applications. Found. Trends Signal Process. 7(3–4), 197–387.
Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., Luca, V.D., Jaggi, M., 2016. Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In: Proceedings of the 10th international workshop on semantic evaluation, no. CONF, pp. 1124–1128.
Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., Torres, J.F., 2018. Stacking ensemble learning for short-term electricity consumption forecasting. Energies 11 (4), 949.
Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14 (2), 241–258.
Dzikovska, M.O., Nielsen, R.D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H.T., 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. NORTH TEXAS STATE UNIV DENTON, Tech. Rep.
Ekbal, A., Bandyopadhyay, S., 2008. Web-based bengali news corpus for lexicon development and pos tagging. Polibits 37, 21–30.
Ekbal, A., Saha, S., 2011. A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in indian languages as case studies. Expert Syst. Appl. 38(12), 14 760–14 772.
Elnagar, A., Al-Debsi, R., Einea, O., 2020. Arabic text classification using deep learning models. Informat. Process. Manage. 57 (1), 102121.
Erdoğan, Z., Namlı, E., 2019. "A living environment prediction model using ensemble machine learning techniques based on quality of life index. J. Ambient Intell. Humanized Comput., 1–17
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Machine Learn. Res. 15 (1), 3133–3181.
Fersini, E., Messina, E., Pozzi, F.A., 2014. Sentiment analysis: Bayesian ensemble learning. Decision Support Syst. 68, 26–38.
Fersini, E., Messina, E., Pozzi, F.A., 2016. Expressive signals in social media languages to improve polarity detection. Informat. Process. Manage. 52 (1), 20–35.
Fouad, M.M., Gharib, T.F., Mashat, A.S., 2018. Efficient twitter sentiment analysis system with feature selection and classifier ensemble. In: International Conference on Advanced Machine Learning Technologies and Applications. Springer, pp. 516–527.
Freund, Y., Schapire, R.E., et al. 1996. Experiments with a new boosting algorithm. 96, pp. 148–156.
Freund, Y., Iyer, R., Schapire, R.E., Singer, Y., 2003. An efficient boosting algorithm for combining preferences. J. Machine Learn. Res. 4 (Nov), 933–969.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals Stat. 1189–1232.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Annals Stat. 28 (2), 337–407.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1–1.1. NASA STI/Recon Technical Report N 93, 27403.
Ge, R., Feng, G., Jing, X., Zhang, R., Wang, P., Wu, Q., 2020. Enacp: An ensemble learning model for identification of anticancer peptides. Front. Genet. 11, 760.
Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N project report, Stanford, vol. 1, no. 12, p. 2009.

Gosselin, P.-H., Murray, N., Jégou, H., Perronnin, F., 2014. Revisiting the fisher vector for fine-grained classification. Pattern Recognit. Lett. 49, 92–98.
Guo, P., Xue, Z., Mtema, Z., Yeates, K., Ginsburg, O., Demarco, M., Long, L.R., Schiffman, M., Antani, S., 2020. Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. Diagnostics 10 (7), 451.
Haghighi, F., Omranpour, H., 2021. Stacking ensemble model of deep learning and its application to persian/arabic handwritten digits recognition. Knowl.-Based Syst. 220, 106940.
Han, S., Meng, Z., Khan, A.-S., Tong, Y., 2016. Incremental boosting convolutional neural network for facial action unit recognition. Adv. Neural Informat. Process. Syst. 29, 109–117.
Haque, A., Milstein, A., Fei-Fei, L., 2020. Illuminating the dark spaces of healthcare with ambient intelligence. Nature 585 (7824), 193–202.
Haralabopoulos, G., Anagnostopoulos, I., McAuley, D., 2020. Ensemble deep learning for multilabel binary classification of user-generated content. Algorithms 13 (4), 83.
Hashino, T., Bradley, A., Schwartz, S., 2007. Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol. Earth Syst. Sci. 11 (2), 939–950.
Hassan, A., Abbasi, A., Zeng, D., 2013. Twitter sentiment analysis: A bootstrap ensemble framework. 2013 International Conference on Social Computing. IEEE, pp. 357–364.
Haumahu, J., Permana, S., Yaddarabullah, Y., 2021. Fake news classification for indonesian news using extreme gradient boosting (xgboost). IOP Conference Series: Materials Science and Engineering, vol. 1098, no. 5. IOP Publishing, p. 052081.
Heikal, M., Torki, M., El-Makky, N., 2018. Sentiment analysis of arabic tweets using deep learning. Proc. Comput. Sci. 142, 114–122.
Helmy, T., Fatai, A., Faisal, K., 2010. Hybrid computational models for the characterization of oil and gas reservoirs. Expert Syst. Appl. 37 (7), 5353–5363.
Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M.K., Johnson, S.C., Initiative, A.D.N., et al., 2009. Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. Neuroimage 48 (1), 138–149.
Hopkinson, B.M., King, A.C., Owen, D.P., Johnson-Roberson, M., Long, M.H., Bhandarkar, S.M., 2020. Automated classification of three-dimensional reconstructions of coral reefs using convolutional neural networks. PloS One 15 (3), e0230671.
Hormozi, E., Akbari, M.K., Hormozi, H., Javan, M.S., 2013. Accuracy evaluation of a credit card fraud detection system on hadoop mapreduce. The 5th Conference on Information and Knowledge Technology. IEEE, pp. 35–39.
Hosni, M., Abnane, I., Idri, A., de Gea, J.M.C., Alemán, J.L.F., 2019. Reviewing ensemble classification methods in breast cancer. Comput. Methods Programs Biomed. 177, 89–112.
Hospedales, T., Antoniou, A., Micaelli, P., Storkey, A., 2021. Meta-learning in neural networks: A survey. IEEE Trans. Pattern Anal. Machine Intell. 44 (9), 5149–5169.
Huang, S., Wang, B., Qiu, J., Yao, J., Wang, G., Yu, G., 2016. Parallel ensemble of online sequential extreme learning machine based on mapreduce. Neurocomputing 174, 352–367.
Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2547–2554.
Jaouedi, N., Boujnah, N., Bouhlel, M.S., 2020. A new hybrid deep learning model for human action recognition. J. King Saud Univ.-Comput. Informat. Sci. 32 (4), 447–453.
Jiang, Z., Lin, Z., Davis, L.S., 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. IEEE Trans. Pattern Anal. Machine Intell. 35 (11), 2651–2664.
Kadam, V.J., Jadhav, S.M., Vijayakumar, K., 2019. Breast cancer diagnosis using feature ensemble learning based on stacked sparse autoencoders and softmax regression. J. Medical Syst. 43 (8), 1–11.
Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Comput. Electron. Agric. 147, 70–90.
Kanakaraj, M., Guddeti, R.M.R., 2015. Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015). IEEE, pp. 169–170.
Kang, S., Kang, P., Ko, T., Cho, S., Rhee, S.-J., Yu, K.-S., 2015. An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction. Expert Syst. Appl. 42 (9), 4265–4273.
Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C., 2015. Cadec: A corpus of adverse drug event annotations. J. Biomed. Informat. 55, 73–81.
Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 172 (5), 1122–1131.
Khamparia, A., Singh, A., Anand, D., Gupta, D., Khanna, A., Arun Kumar, N., Tan, J., 2020. A novel deep learning-based multi-model ensemble method for the prediction of neuromuscular disorders. Neural Comput. Appl. 32 (15), 11-083–11-095.
Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H., Alfakeeh, A.S., 2020. "Stock market prediction using machine learning classifiers and social media, news. J. Ambient Intell. Humanized Comput., 1–24
Kim, H.-C., Pang, S., Je, H.-M., Kim, D., Bang, S.Y., 2003. Constructing support vector machine ensemble. Pattern Recognit. 36 (12), 2757–2767.
Koh, N.S., Hu, N., Clemons, E.K., 2010. Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures. Electron. Commer. Res. Appl. 9 (5), 374–385.
Krause, J., Stark, M., Deng, J., Fei-Fei, L., 2013. 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 554–561.
Krawczyk, B., Minku, L.L., Gama, J., Stefanowski, J., Woźniak, M., 2017. Ensemble learning for data stream analysis: A survey. Informat. Fusion 37, 132–156.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Informat. Process. Syst. 25, 1097–1105.
Kulkarni, N.H., Srinivasan, G., Sagar, B., Cauvery, N., 2018. Improving crop productivity through a crop recommendation system using ensembling technique. In: 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). IEEE, pp. 114–119.
Kumar, G., Misra, A.K., 2018. Commonality in liquidity: Evidence from india's national stock exchange. J. Asian Econ. 59, 1–15.
Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D., 2016. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J. Biomed. Health Informat. 21 (1), 31–40.
Kumar, V., Aydav, P.S.S., Minz, S., 2021. Multi-view ensemble learning using multi-objective particle swarm optimization for high dimensional data classification. J. King Saud Univ.-Comput. Informat. Sci.
Kuruvayil, S., Palaniswamy, S., 2021. Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. J. King Saud Univ.-Comput. Informat. Sci.
Kuruvayil, S., Palaniswamy, S., 2022. Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning. J. King Saud Univ.-Comput. Informat. Sci. 34 (9), 7271–7282.
Kuznetsov, V., Mohri, M., Syed, U., 2014. Multi-class deep boosting.
Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv. Neural Informat. Process. Syst. 30.
Latif-Shabgahi, G.-R., 2004. A novel algorithm for weighted average voting used in fault tolerant computing systems. Microprocess. Microsyst. 28 (7), 357–361.
LeCun, Y., 1998. The mnist database of handwritten digits, http://yann.lecun.com/exdb/mnist/.
Lewis, D.D., Yang, Y., Russell-Rose, T., Li, F., 2004. Rcv1: A new benchmark collection for text categorization research. J. Machine Learn. Res. 5 (Apr), 361–397.
Li, Y., Maguire, L., 2010. Selecting critical patterns based on local geometrical and statistical information. IEEE Trans. Pattern Anal. Machine Intell. 33 (6), 1189–1201.
Li, S, Lee, S.Y., Chen, Y., Huang, C.-R., Zhou, G., 2010. Sentiment classification and polarity shifting. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 635–643.
Li, F.H., Huang, M., Yang, Y., Zhu, X., 2011. Learning to identify review spam. In: Twenty-second International Joint Conference on Artificial Intelligence.
Li, W., Wang, W., Chen, Y., 2012. Heterogeneous ensemble learning for chinese sentiment classification. J. Informat. Comput. Sci. 9 (15), 4551–4558.
Li, J., Chang, H., Yang, J., 2015. Sparse deep stacking network for image classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29, no. 1.
Liu, G.-H., Li, Z.-Y., Zhang, L., Xu, Y., 2011. Image retrieval based on micro-structure descriptor. Pattern Recogn. 44 (9), 2123–2133.
Liu, P., Han, S., Meng, Z., Tong, Y., 2014. Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1805–1812.
Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104.
Liu, F., Xu, F., Yang, S., 2017. A flood forecasting model based on deep learning algorithm via integrating stacked autoencoders with bp neural network. 2017 IEEE third International conference on multimedia big data (BigMM). Ieee, pp. 58–61.
Livieris, I.E., Kanavos, A., Tampakas, V., Pintelas, P., 2019. A weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from x-rays. Algorithms 12 (3), 64.
Livieris, I.E., Iliadis, L., Pintelas, P., 2020. On ensemble techniques of weight-constrained neural networks. Evolv. Syst., 1–13
Locketz, G.D., Li, P.M., Fischbein, N.J., Holdsworth, S.J., Blevins, N.H., 2016. Fusion of computed tomography and propeller diffusion-weighted magnetic resonance imaging for the detection and localization of middle ear cholesteatoma. JAMA Otolaryngol.-Head Neck Surg. 142 (10), 947–953.
Lu, B., Tsou, B.K., 2010. Combining a large sentiment lexicon and machine learning for subjectivity classification. 2010 International Conference on Machine Learning and Cybernetics, vol. 6. IEEE, pp. 3311–3316.
Lu, X., Van Roy, B., 2017. Ensemble sampling. Adv. Neural Informat. Process. Syst. 30.
Ma, Z., Wang, P., Gao, Z., Wang, R., Khalighi, K., 2018. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. PloS One 13 (10), e0205872.
Makhtar, M., Yang, L., Neagu, D., Ridley, M., 2012. Optimisation of classifier ensemble for predictive toxicology applications. In: 2012 UKSim 14th International Conference on Computer Modelling and Simulation. IEEE, pp. 236–241.

Marques, J., Alves, R.M.F., Oliveira, H.C., Mendonca, M., Souza, J.R., 2021. An evaluation of machine learning methods for speed-bump detection on a gopro dataset. Anais da Academia Brasileira de Ciencias 93 (1), e20190734.
Mendonca, T., Celebi, M., Mendonca, T., Marques, J., 2015. Ph2: A public database for the analysis of dermoscopic images. Dermoscopy image analysis.
Mishra, S., Mishra, D., 2015. Adaptive multi-classifier fusion approach for gene expression dataset based on probabilistic theory. J. Korean Stat. Soc. 44 (2), 247–260.
Moghimi, M., Belongie, S.J., Saberian, M.J., Yang, J., Vasconcelos, N., Li, L.-J., 2016. Boosted convolutional neural networks. In: BMVC, vol. 5, p. 6.
Mohammadi, A., Shaverizade, A., 2021. Ensemble deep learning for aspect-based sentiment analysis. Int. J. Nonlinear Anal. Appl. 12, 29–38.
Mohammed, A., Kora, R., 2019. Deep learning approaches for arabic sentiment analysis. Social Network Anal. Min. 9 (1), 1–12.
Mohammed, A., Kora, R., 2021. An effective ensemble deep learning framework for text classification. J. King Saud Univ.-Comput. Informat. Sci. 2021.
Monteiro, J.P., Ramos, D., Carneiro, D., Duarte, F., Fernandes, J.M., Novais, P., 2021. Meta-learning and the new challenges of machine learning. Int. J. Intell. Syst. 36 (11), 6240–6272.
Montgomery, J.M., Hollenbach, F.M., Ward, M.D., 2012. Improving predictions using ensemble bayesian model averaging. Polit. Anal. 20 (3), 271–291.
Mosca, A., Magoulas, G.D., 2016. Deep incremental boosting. in: GCAI, pp. 293–302.
Nabil, M., Aly, M., Atiya, A., 2015. Astd: Arabic sentiment tweets dataset. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2515–2519.
Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S.M., Kozareva, Z., Ritter, A., Stoyanov, V., Zhu, X., 2016. Developing a successful semeval task in sentiment analysis of twitter and other social media texts. Language Resourc. Eval. 50 (1), 35–65.
Nguyen, H.T., Le Nguyen, M., 2019. An ensemble method with sentiment features and clustering support. Neurocomputing 370, 155–165.
Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T., 2015. Audio-visual speech recognition using deep learning. Appl. Intell. 42 (4), 722–737.
Nti, I.K., Adekoya, A.F., Weyori, B.A., 2020. A comprehensive evaluation of ensemble learning for stock-market prediction. J. Big Data 7 (1), 1–40.
Onan, A., Korukoğlu, S., Bulut, H., 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. Expert Syst. Appl. 62, 1–16.
Opitz, M., Waltner, G., Possegger, H., Bischof, H., 2017. Bier-boosting independent embeddings robustly. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5189–5198.
Ortiz, A., Munilla, J., Gorriz, J.M., Ramirez, J., 2016. Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease. Int. J. Neural Syst. 26 (07), 1650025.
Oussous, A., Lahcen, A.A., Belfkih, S., 2018. Improving sentiment analysis of moroccan tweets using ensemble learning. In: International Conference on Big Data, Cloud and Applications. Springer, pp. 91–104.
Palangi, H., Deng, L., Ward, R.K., 2014. Recurrent deep-stacking networks for sequence classification. In: 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP). IEEE, pp. 510–514.
Pandit, S., Kumar, S., 2020. Improvement in convolutional neural network for cifar-10 dataset image classification. Int. J. Comput. Appl. 176, 25–29.
Pang, B., Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: ACL.
Pasupulety, U., Anees, A.A., Anmol, S., Mohan, B.R., 2019. Predicting stock prices using ensemble learning and sentiment analysis. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). IEEE, pp. 215–222.
Perikos, I., Hatzilygeroudis, I., 2016. Recognizing emotions in text using ensemble of classifiers. Eng. Appl. Artif. Intell. 51, 191–201.
Polikar, R., 2012. Ensemble learning. In: Ensemble Machine Learning. Springer, pp. 1–34.
Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., Žabokrtskỳ, Z., 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nat. Commun. 11 (1), 1–15.
Prusa, J., Khoshgoftaar, T.M., Dittman, D.J., 2015. Using ensemble learners to improve classifier performance on tweet sentiment data. 2015 IEEE International Conference on Information Reuse and Integration. IEEE, pp. 252–257.
Qiu, X., Zhang, L., Ren, Y., Suganthan, P.N., Amaratunga, G., 2014. Ensemble deep learning for regression and time series forecasting. 2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL). IEEE, pp. 1–6.
Rodriguez-Penagos, C., Atserias, J., Codina-Filba, J., García-Narbona, D., Grivolla, J., Lambert, P., Saurí, R., 2013. Fbm: Combining lexicon-based ml and heuristics for social media polarities. In: Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 483–489.
Rokach, L., 2019. Ensemble learning: Pattern classification using ensemble methods. World Sci. 85.
Rushdi-Saleh, M., Martín-Valdivia, M.T., Ureña-López, L.A., Perea-Ortega, J.M., 2011. Oca: Opinion corpus for arabic. J. Am. Soc. Informat. Sci. Technol. 62(10), 2045–2054.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115 (3), 211–252.
Saeed, R.M., Rady, S., Gharib, T.F., 2022. An ensemble approach for spam detection in arabic opinion texts. J. King Saud Univ.-Comput. Informat. Sci. 34 (1), 1407–1416.
Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscip. Rev.: Data Min. Knowledge Discov. 8 (4), e1249.
Saleena et al. N. 2018. An ensemble classification system for twitter sentiment analysis. Proc. Comput. Sci. 132, 937–946.
Saleh, H., Mostafa, S., Alharbi, A., El-Sappagh, S., Alkhalifah, T., 2022. Heterogeneous ensemble deep learning model for enhanced arabic sentiment analysis. Sensors 22 (10), 3707.
Scopus, 2023. scopus preview, https://scopus.com/.
Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., Alonso-Betanzos, A., 2017. Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl.-Based Syst. 118, 124–139.
Seker, S.E., Ocak, I., 2019. Performance prediction of roadheaders using ensemble machine learning techniques. Neural Comput. Appl. 31 (4), 1103–1116.
7Seki, Y., Evans, D.K., Ku, L.-W., L.S. 0001, Chen, H.-H., Kando, N., 2008. Overview of multilingual opinion analysis task at ntcir-7. In: NTCIR. Citeseer, pp. 185–203.
Seyyedsalehi, S.Z., Seyyedsalehi, S.A., 2014. Simultaneous learning of nonlinear manifolds based on the bottleneck neural network. Neural Proces. Lett. 40 (2), 191–209.
Shahzad, R.K., Lavesson, N., 2013. Comparative analysis of voting schemes for ensemble-based malware detection. J. Wireless Mobile Netw., Ubiquitous Comput. Dependable Appl. 4 (1), 98–117.
Shahzad, R.K., Haider, S.I., Lavesson, N., 2010. Detection of spyware by mining executable files. In: 2010 International Conference on Availability, Reliability and Security. IEEE, pp. 295–302.
Sharma, A., Raju, D., Ranjan, S., 2017. Detection of pneumonia clouds in chest x-ray using image processing approach. In: 2017 Nirma University International Conference on Engineering (NUiCONE). IEEE, pp. 1–4.
Sharma, S., Srivastava, S., Kumar, A., Dangi, A., 2018. Multi-class sentiment analysis comparison using support vector machine (svm) and bagging technique-an ensemble method. In: 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). IEEE, pp. 1–6.
Shin, Y., 2019. Application of stochastic gradient boosting approach to early prediction of safety accidents at construction site. Adv. Civil Eng. 2019.
Shipp, C.A., Kuncheva, L.I., 2002. Relationships between combination methods and measures of diversity in combining classifiers. Informat. Fus. 3 (2), 135–148.
Smyth, P., Wolpert, D., 1997. Stacked density estimation. Adv. Neural Informat. Process. Syst. 10.
Soares, C., Brazdil, P.B., Kuba, P., 2004. A meta-learning method to select the kernel width in support vector regression. Machine Learn. 54 (3), 195–209.
Stamatatos, E., Widmer, G., 2002. Music performer recognition using an ensemble of simple classifiers. ECAI, 335–339.
Su, Y., Zhang, Y., Ji, D., Wang, Y., Wu, H., 2012. Ensemble learning for sentiment classification. In: Workshop on Chinese Lexical Semantics. Springer, pp. 84–93.
Sultana, N., Sharma, N., Sharma, K.P., Verma, S., 2020. A sequential ensemble model for communicable disease forecasting. Curr. Bioinform. 15 (4), 309–317.
Sun, B., Chen, S., Wang, J., Chen, H., 2016. A robust multi-class adaboost algorithm for mislabeled noisy data. Knowl.-Based Syst. 102, 87–102.
Täckström, O., McDonald, R., 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.
Tang, J., Su, Q., Su, B., Fong, S., Cao, W., Gong, X., 2020. Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition. Comput. Methods Programs Biomed. 197, 105622.
Tasci, E., Uluturk, C., Ugur, A., 2021. A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. Neural Comput. Appl., 1–15
Thakur, R.S., Yadav, R.N., Gupta, L., 2019. State-of-art analysis of image denoising methods using convolutional neural networks. IET Image Proc. 13 (13), 2367–2380.
Tratz, S., Briesch, D., Laoudi, J., Voss, C., Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pp. 135–139.
Tsai, C.-F., Lin, Y.-C., Yen, D.C., Chen, Y.-M., 2011. Predicting stock returns by classifier ensembles. Appl. Soft Comput. 11 (2), 2452–2459.
Tsutsumi, K., Shimada, K., Endo, T., 2007. Movie review classification based on a multiple classifier. In: Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation, pp. 481–488.
Tur, G., Deng, L., Hakkani-Tür, D., He, X., 2012. Towards deeper understanding: Deep convex networks for semantic utterance classification. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5045–5048.
Valle, C., Saravia, F., Allende, H., Monge, R., Fernández, C., 2010. Parallel approach for ensemble learning with locally coupled neural networks. Neural Process. Lett. 32 (3), 277–291.
van Aken, B., Risch, J., Krestel, R., Löser, A., 2018. Challenges for toxic comment classification: An in-depth error analysis. In: ALW.
Walach, E., Wolf, L., 2016. Learning to count with cnn boosting. In: European Conference on Computer Vision. Springer, pp. 660–676.
Waltner, G., Opitz, M., Possegger, H., Bischof, H., 2019. Hibster: Hierarchical boosted deep metric learning for image retrieval. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 599–608.

Wang, X.-Y., Zhang, B.-B., Yang, H.-Y., 2013. Active svm-based relevance feedback using multiple classifiers ensemble and features reweighting. Eng. Appl. Artif. Intell. 26 (1), 368–381.
Wang, G., Sun, J., Ma, J., Xu, K., Gu, J., 2014. Sentiment classification: The contribution of ensemble learning. Decision Support Syst. 57, 77–93.
Wang, F., Jiang, D., Wen, H., Song, H., 2019. Adaboost-based security level classification of mobile intelligent terminals. J. Supercomput. 75 (11), 7460–7478.
Wang, B., Xue, B., Zhang, M., 2020. Particle swarm optimisation for evolving deep neural networks for image classification by evolving and stacking transferable blocks. 2020 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp. 1–8.
Wen, Y.-H., Lee, T.-T., CHO, H.-J., 2005. Missing data treatment and data fusion toward travel time estimation for atis. J. Eastern Asia Soc. Transport. Stud. 6, 2546–2560.
Whitehead, M., Yaeger, L., 2009. Building a general purpose cross-domain sentiment mining model. 2009 WRI World Congress on Computer Science and Information Engineering, vol. 4. IEEE, pp. 472–476.
Wiebe, J., Wilson, T., Cardie, C., 2005. Annotating expressions of opinions and emotions in language. Language Resourc. Eval. 39 (2), 165–210.
Wilson, T., Wiebe, J., Hwa, R., 2006. Recognizing strong and weak opinion clauses. Comput. Intell. 22 (2), 73–99.
Wu, J., Yu, X., Liu, D., Chandraker, M., Wang, Z., 2020. David: Dual-attentional video deblurring. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 2365–2374.
Xia, R., Zong, C., Li, S., 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Informat. Sci. 181 (6), 1138–1152.
Xia, R., Xu, F., Yu, J., Qi, Y., Cambria, E., 2016. Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis. Informat. Process. Manage. 52 (1), 36–45.
Xiong, Y., Ye, M., Wu, C., 2021. Cancer classification with a cost-sensitive naive bayes stacking ensemble. Comput. Mathe. Methods Med. 2021.
Xu, S., Liang, H., Baldwin, T., 2016. Unimelb at semeval-2016 tasks 4a and 4b: An ensemble of neural networks and a word2vec based model for sentiment classification. In: Proceedings of the 10th international Workshop on Semantic Evaluation (SemEval-2016), pp. 183–189.
Yang, B., Yan, J., Lei, Z., Li, S.Z., 2015. Convolutional channel features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 82–90.
Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: Lstm cells and network architectures. Neural Comput. 31 (7), 1235–1270.
Zareapoor, M., Shamsolmoali, P., et al., 2015. Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Comput. Sci. 48 (2015), 679–685.
Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., Xiao, W., 2016. Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing 173, 979–987.
Zhang, H., Dai, Y., Li, H, Koniusz, P., 2019. Deep stacked hierarchical multi-patch network for image deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5978–5986.
Zhang, W., Jiang, J., Shao, Y., Cui, B., 2020. Snapshot boosting: a fast ensemble framework for deep neural networks. Science China Informat. Sci. 63 (1), 1–12.
Zhang, J., Zhang, W., Song, R., Ma, L., Li, Y., 2020. Grasp for stacking via deep reinforcement learning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 2543–2549.