applied sciences

MDPI

Article

Use of Ensemble Learning to Improve Performance of Known Convolutional Neural Networks for Mammography Classification

Mayra C. Berrones-Reyes [1,*], M. Angélica Salazar-Aguilar [1,*] and Cristian Castillo-Olea [2,*]

[1] Facultad de Ingeniería Mecánica y Eléctrica, Universidad Autónoma de Nuevo León, Av. Universidad s/n, Cd. Universitaria, San Nicolás de los Garza 66455, Mexico
[2] Facultad de Ingeniería, CETYS Universidad Campus Mexicali, Calzada CETYS s/n, Colonia Rivera, Mexicali 21259, Mexico
* Correspondence: mayra.berronesrys@uanl.edu.mx (M.C.B.-R.); maria.salazaragl@uanl.edu.mx (M.A.S.-A.); cristian.castillo@cetys.mx (C.C.-O.)

Abstract: Convolutional neural networks and deep learning models represent the gold standard in medical image classification. Their innovative architectures have led to notable breakthroughs in image classification and feature extraction performance. However, these advancements often remain underutilized in the medical imaging field due to the scarcity of sufficient labeled data which are needed to leverage these new features fully. While many methodologies exhibit stellar performance on benchmark data sets like DDSM or Minimias, their efficacy drastically decreases when applied to real-world data sets. This study aims to develop a tool to streamline mammogram classification that maintains high reliability across different data sources. We use images from the DDSM data set and a proprietary data set, YERAL, which comprises 943 mammograms from Mexican patients. We evaluate the performance of ensemble learning algorithms combined with prevalent deep learning models such as Alexnet, VGG-16, and Inception. The computational results demonstrate the effectiveness of the proposed methodology, with models achieving 82% accuracy without overtaxing our hardware capabilities, and they also highlight the efficiency of ensemble algorithms in enhancing accuracy across all test cases.

Keywords: convolutional neural networks; ensemble learning; deep learning; transfer learning; image classification; medical imaging; mammography

Citation: Berrones-Reyes, M.C.; Salazar-Aguilar, M.A.; Castillo-Olea, C. Use of Ensemble Learning to Improve Performance of Known Convolutional Neural Networks for Mammography Classification. *Appl. Sci.* **2023**, *13*, 9639. https://doi.org/10.3390/app13179639

Academic Editor: Je-Keun Rhee

Received: 31 July 2023
Revised: 16 August 2023
Accepted: 18 August 2023
Published: 25 August 2023

1. Introduction

Machine learning and artificial intelligence have come a long way in advancing practical solutions to our everyday problems. Due to their remarkable progress, these methods have become the gold standard when solving more complex issues, such as image recognition [1–3], automatized tasks [4,5], and optimization problems [6], among many others. In medical imaging, machine learning (ML) has been vital to building more modern versions of computer-aided diagnostic tools (CADs).

Deep learning has now become the computer standard in medical imaging analysis; important areas requiring a high level of specialization, such as radiology, dermatology, pathology, and ophthalmology, have found excellent performance levels when using deep learning [7]. Processes that took significant time and human resources are now improved by increasing accuracy and speeding up the diagnosis. In some cases, investigations have reported a level matching or exceeding human performance, which has caused a massive wave of ML models applied to medical problems [8].

However, the surge in ML models for complex medical imaging problems has raised concerns among medical professionals. According to [9,10], these concerns primarily center around the applicability of these models in real-life clinical settings, given the intricate requirements of deep learning and ML models. Deep learning models often need a vast number of data for optimal performance. With specialized computer-aided systems,

medical imaging requires rigorous expert review to label each image type appropriately. As a result, there is a scarcity of free-to-use data sets, creating a bottleneck for the further exploration of deep learning applications.

Other concerns include the limited representational scope of these free-to-use data sets in the practical application of an ML model. Studies have indicated that algorithms that perform well on benchmark data sets often falter in real-world clinical scenarios [11]. A significant contributor to this discrepancy is data set bias, which frequently occurs when the training set's population distribution differs from the target set's, commonly known as the test set [12].

Image annotation, despite being time-consuming and requiring a large team of specialists, poses an even greater challenge when constructing models with limited labeled images. Models need more data to converge to a good performance. This demand for resources is particularly detrimental in low-income and middle-income countries, where the need for diagnosing medical images outstrips the available specialist capacity [8].

In this work, we examine a mammography classification case from a data set called YERAL, obtained from a specialized Mexican Oncology center. YERAL represents our target data set for optimizing the diagnosis process using ML models. The driving factor behind focusing on this particular medical issue is the staggering volume of data accumulating each year, outpacing the number of medical specialists available to handle it, despite the Mexican government's best efforts to create awareness.

We introduce a novel approach by leveraging ensemble learning in scenarios where the state of the art, like deep learning and cutting-edge methods, does not fit because of data and computational resource constraints. While others might hit a roadblock with limited resources, we have merged elements of deep learning with traditional machine learning tools. This hybrid approach allows us to break down the data, making them more digestible. Our innovative use of ensemble learning in this context not only streamlines training complexity but also serves as a pioneering contribution to the field, demonstrating new possibilities for those facing similar limitations.

The article's organization is as follows: We first discuss the problem we found about medical imaging results with existing deep learning and ML models that, when applied to our data set, received lower performance levels. Then, we discuss the steps we followed to solve this issue and why we opted to look into ensemble learning algorithms to improve our metrics performance. We highlight in the Background section why machine learning techniques in combination with deep learning can increase accuracy when deep learning alone produces low-level performance models. Ultimately, we show improved results and discuss the pros and cons we found when working with these ML models.

2. Background

In recent years, AI applications have rapidly developed and improved due to substantial public interest. Market forces stand to gain from applying deep learning in natural language translation, photo captioning, speech recognition, and self-driving cars [8].

The continuing success in applying these models has often caused the desire to project these same results into other areas. Medical imaging stands out in this context. The correct and prompt interpretation of diagnostic information is pivotal and, as such, remains heavily reliant on human expertise [13]. Such a paradigm often presents challenges, especially considering the voluminous medical images requiring diagnoses and the limited number of specialists available for interpretation.

While numerous studies have demonstrated that deep learning models can achieve expert-level performance in medical imaging, there is still room for improvement in reporting standards. Comprehensive surveys, such as those by [8,14], have highlighted a common concern: methodologies and studies often lack completeness and do not conform to a standardized approach for presenting results. Such inconsistencies in reporting often compromise the reliability of interpretations and challenge the replication of results with different data sets.

A notable observation from these surveys is that a large proportion of publications emphasize the application of convolutional neural networks (CNNs). Within these works, reference is frequently made to renowned architectures like Alexnet [15], VGG-16 [16], ResNet [17], and Inception [18]. CNNs have gained substantial acclaim in the medical domain due to their innate capability to automatically learn features that help to distinguish between classes of various computer vision tasks, and there are many examples where they achieve noteworthy performances [19,20]. As the field has evolved, an array of architectures has emerged, each presenting its own set of advantages and limitations. This diversity offers users the flexibility to select the parameters best suited to their needs.

The evolution of CNN architectures has been marked by significant milestones, one of which is the GoogLeNet or Inception network [18]. This architecture heralded a new era in computer vision, thanks to its distinct inception block layer. Due to its features, it has become a preferred choice for transfer learning in breast imaging, encompassing a broad spectrum of images beyond just mammography [21].

Subsequently, the ResNet architecture [17] made another big step for complex classification problems by introducing the concept of residual learning. With an impressive structure of over 100 layers and more than 11 million parameters, ResNet underscores the belief that deeper CNNs generally produce better outcomes. This assertion was further accentuated by identifying bottlenecks in the VGG-16 network, where increased data complexity led to a diminished generalization capability.

In a similar path to CNN architectures, the practice of transfer learning (TF) has also garnered considerable attention [7,22,23]. One of the main issues attached to medical applications for deep learning is that there is usually a limited number of annotated data, where the problem of overfitting arises [24]. This challenge comes from a limited number of available annotated training samples. Transfer learning helps with this problem by using pretrained models from nonmedical images, fine-tuning the network parameters to fit our data later, or using it to perform feature extraction [25].

A literature review about transfer learning for medical imaging classification [26] describes the backbone models for transfer learning, leaving Alexnet and VGG-16 as shallow and linear model types and Inception and ResNet as deep model types. The authors detail the importance of identifying the eligibility of these networks for different types of data.

Research has demonstrated that feature extraction by CNNs can notably boost classification tasks [27]. Other studies highlight the use of this ensemble framework in combination with the power of neural networks to enhance the performance of standard detection techniques [28].

At its core, ensemble learning is about bringing together multiple estimators to better address a machine learning task. While it falls under the umbrella of artificial intelligence, it leans more toward statistical learning. Drawing inspiration from natural behaviors, the ensemble approach mirrors the human tendency to weigh multiple perspectives when making intricate decisions. Various surveys delve into the effectiveness of ensemble methods, pointing out their respective merits and challenges and identifying contexts where they are particularly beneficial.

Ensemble methods generally bifurcate into averaging and boosting methods. There is a substantial body of work that discusses the application of ensemble algorithms in medical imaging [29–31]. These studies often face the issue of not being able to use deep learning tools because of their reduced numbers of data.

In the realm of CAD (computational-aided diagnostic) systems tailored for mammograms, several investigations akin to ours employ a methodology that encompasses image preprocessing, feature extraction using CNNs, and the application of a stacking ensemble method [32].

However, a notable difference lies in the data sets employed. While many studies like the one by [32] utilize the publicly available MIAS data set, they often do not tap into genuine clinical data sets. This discrepancy was highlighted in numerous medical

surveys critiquing such computational tools. There are also mentions of ensemble learning's commendable results in other works like [33], especially given its enhanced performance outcomes. Nonetheless, these studies often differ in the type of data they use as input for their models.

Several well-established machine learning algorithms like logistic regression, random forest, *k*-neighbors, and gradient boosting are highlighted in certain health computer-aided tools [34]. While these algorithms are prevalent in the broader domain of machine learning, they often fall short for medical imaging tasks, given that they are not primarily tailored for such data types.

This work considers the heavy remarks raised in medical journals about the prevalent issue of insufficient reporting on diagnostic accuracy in deep learning applications [8]. These shortcomings underscore the need for comprehensive documentation and the rigorous evaluation of the methodologies employed. In the next section, we delve into the results we achieved using deep learning models. Furthermore, we mention the steps and methods we employed to enhance their performance, specifically by integrating deep learning models with conventional ensemble learning ML algorithms.

In essence, by combining traditional machine learning tools with advanced deep learning models, this research endeavors to elevate the diagnostic precision and robustness of medical imaging solutions, ensuring that they are both effective and reproducible.

3. Materials and Methods

In the realm of medical imaging, artificial intelligence (AI) tools are frequently combined with computer vision algorithms. These tools and algorithms are evaluated using a consistent set of metrics and standards. Based on their performance results, one can ascertain the most suitable AI model for the specific problem at hand.

Take, for instance, the application of deep learning algorithms for image recognition tasks. The predominant metrics used to evaluate their efficacy include accuracy, precision, recall, and F1 score, with accuracy often being the most cited metric in medical imaging literature [35].

Furthermore, the practice of data set division, borrowed from computational vision, is pivotal in model training and evaluation. Conventionally, in ML models, data are divided into two segments: one reserved for training and the other for evaluating or testing the model's outcomes. The majority of the data, typically ranging between 70% and 80%, are allocated to the training set, leaving the remainder for the test set. This proportional division is strategic, ensuring that there are ample data for the model to be trained effectively and therefore mitigating the risk of underfitting.

3.1. Methods of Improvement

In the literature, machine learning and deep learning are used interchangeably. However, as previously established in this work, there are situations where the application of deep learning could be likened to using too big a hammer, especially when faced with limited data or constrained computational resources. In such scenarios, ensemble methods emerge as a potent alternative, offering the potential for constructing more resilient models [36].

Ensemble learning is a widely adopted technique, aimed at enhancing the predictive accuracy of machine learning algorithms by pooling the predictions from multiple models [37]. To visualize this concept in a real-world context, consider the approach of a medical team diagnosing a patient. Initially, they gather all the available patient data for diagnosis. These data are then shared among several specialists, each assessing them independently and arriving at their own diagnostic conclusions. Ultimately, a senior expert or the team leader reviews all the individual opinions, integrates them, and, coupled with their own insights, finalizes the diagnosis.

In essence, an *ensemble learning algorithm* is an ML strategy designed to amplify performance by leveraging the strengths of multiple individual estimators.

3.1.1. Bagging or Bootstrap Aggregation

Bootstrap aggregation, commonly known as bagging, is an ensemble learning method aiming to promote diversity amongst ensemble members by manipulating the training data. This method uses a statistical approach to estimate a population derived by averaging results from numerous small data samples. These samples are created by selecting observations from a larger data set and then returning them, a procedure termed sampling with replacement.

As depicted in Figure 1, bootstrap aggregation is represented graphically. Several subsets, identical in size and selected with replacement, are extracted from the primary data. A CNN of consistent architecture is applied to each subset. The results from these individual models are then collated and voted upon to produce a singular prediction. The Alexnet experiment kept the hyperparameters as specified in its original publication, while for VGG-16 and Inception, we employed pretrained weights from IMAGENET. It is important to note that all experiments were conducted in a manner that did not elevate the computational demands, keeping up with the simplicity of this ensemble approach.

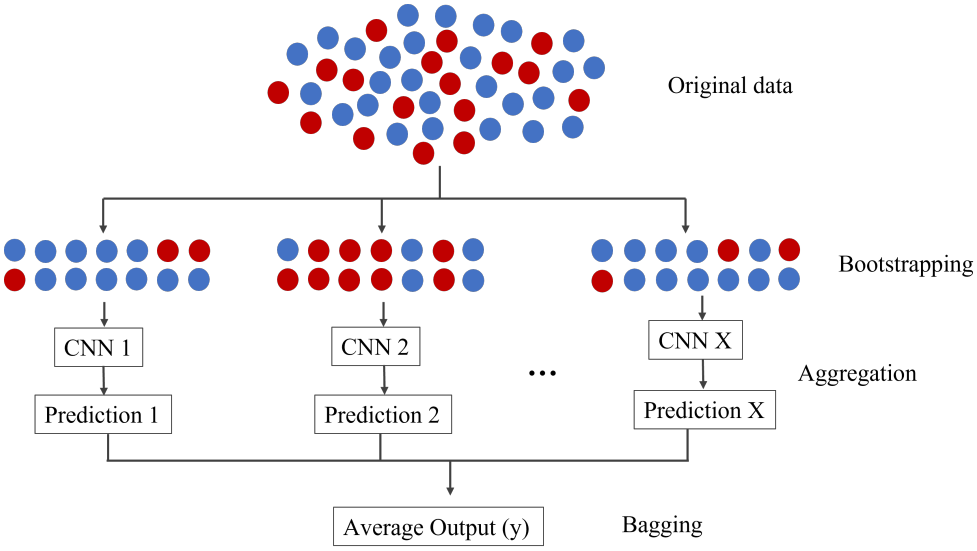


Figure 1. Example of the bagging algorithm where the blue and red dots represent the normal and abnormal images, with the CNN structures in the first experiment Alexnet, then VGG-16, and lastly with Inception. With the results, there is an averaging of predictions that results in the final voting.

The essence of the bootstrap method is estimation. It samples small portions, computes statistics for each, and then averages them. It is imperative that data preparation occurs within the sample data's loop, especially before model fitting or hyperparameter tuning. Such a step prevents data leakage, a scenario where the model, having complete access to the entire data set, inadvertently optimizes itself and causes itself to overfit.

In the case of bagging ensemble learning, averaging the predictions across the models typically results in better predictions than a single model fit on the training data set directly.

3.1.2. Stacking Ensemble Learning

Stacking is a technique that leverages multiple machine learning models, or estimators, to generate predictions. Unlike mere averaging, stacking feeds these predictions into a new model which subsequently forms its own predictions based on the earlier results. Within this framework, models have specific designations: those used in the primary ensemble step are termed 'zero-level models' or 'weak learners', while the subsequent model that consolidates these predictions is the 'first-level model'. Typically, stacking follows a two-tier hierarchy, though more layers can be introduced.

In the context of deep learning, transfer learning is an adaptation of stacking. As touched upon in the Background section, transfer learning is frequently cited in the litera-

ture as a means to utilize pretrained weights and architectures, which diminishes computational demands and paves the way for the implementation of intricate architectures.

However, given that these transfer learning architectures are originally trained on natural images rather than medical ones, their efficacy in medical image classification remains a topic of debate. In this study, the initial phase of the stacking process employs transfer learning using the VGG-16 architecture for feature extraction. This approach taps into the capabilities of a deep learning algorithm while synergizing it with other models. Subsequently, as an alternative to VGG-16, transfer learning is executed with the Inception architecture to assess its comparative performance.

Notably, akin to the scenario with the bagging ensemble, Alexnet was excluded from the transfer learning process. This is due to its relatively simpler architecture, which generally produces superior outcomes when trained from scratch.

Using pretrained weights to derive features from images generates a numerical vector. This vector, representative of distinct image attributes, offers a format readily interpretable by conventional machine learning algorithms. Despite stacking potentially increasing processing time, initial feature extraction renders the data more manageable.

As illustrated in Figure 2, following feature extraction, prevalent machine learning algorithms such as decision tree, random forest, *k*-neighbors, and support vector classifier are employed to interpret the transformed data. These algorithms are accessible within the *sklearn* Python library [38].

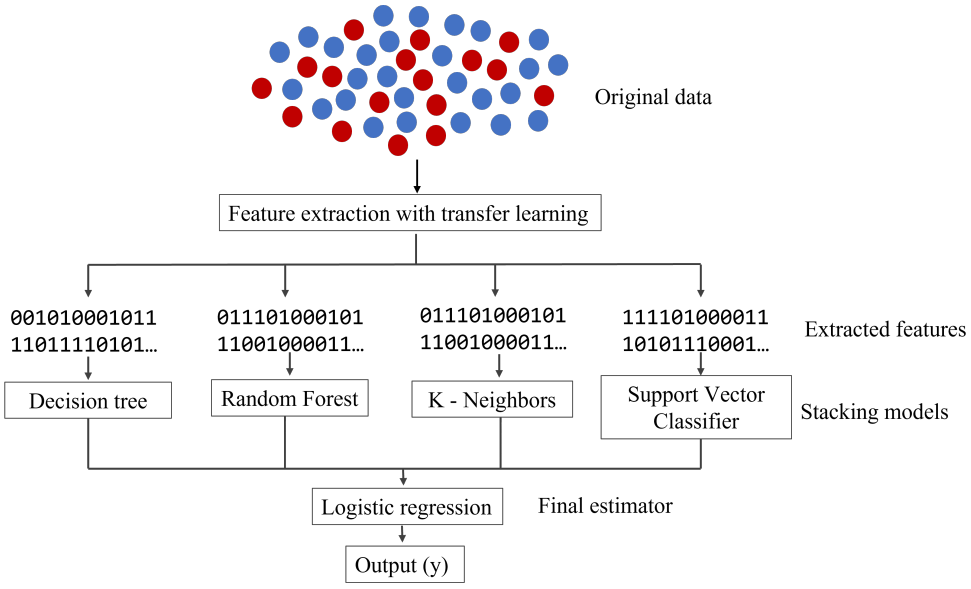


Figure 2. Example of the stacking ensemble where the blue and red dots represent the normal and abnormal images. The first part consists of a transfer learning method as a feature extractor that transforms the data set into a manageable format for machine learning algorithms. The second one contains the weak learners used for intermediate predictions, and the third one corresponds to the final estimator given by logistic regression.

We use deep learning at the beginning of the stacking and combine various machine learning models used for classification. A more in-depth explanation of each of the algorithms can be found in [36,39,40]. Nevertheless, in the context of the stacking ensemble, the models can be summarized as follows:

- **Decision tree:** As a base learner, a decision tree can be quick to train and has the advantage of simplicity. However, it might be prone to overfitting on its own.
- **Random forest:** This classifier is more robust than a single decision tree. It can reduce overfitting by averaging the results of individual trees. It is commonly used as a base learner in stacking due to its efficiency and high accuracy.

- **K-nearest neighbor:** It can capture complex patterns in the data without requiring explicit model training. It can be used as a base learner in stacking, especially when the data set has complex, nonlinear boundaries.
- **Support vector classifier:** As a base learner, SVC can capture complex relationships, especially when equipped with nonlinear kernels. It can be computationally intensive, so its use in stacking would depend on the data set size and computational constraints.

Finally, instead of averaging the results, we use a final logistic regression estimator that returns the final prediction. Due to its simplicity, regularization, and flexibility properties, logistic regression is a common and often effective choice as a metalearner in stacking for classification problems.

Stacked generalization is a method for combining estimators to reduce their biases. Since we have a balanced data set, the logistic estimator is used to average the solution for balanced performance and explainability.

3.1.3. Boosting Algorithms

Boosting is distinct from both bagging and stacking ensemble techniques. As illustrated in Figure 3, in boosting, models are sequentially integrated into the ensemble. Each subsequent model strives to rectify the predictions of its predecessor. The overarching aim of this method is to evolve a robust learner through successive iterations. What differentiates boosting from techniques such as bagging is its inherent capacity to learn iteratively from prior classifiers, progressively focusing on misclassified elements. Contrarily, in bagging, each iteration uses a separate set, thus lacking this accumulative 'learning' aspect.

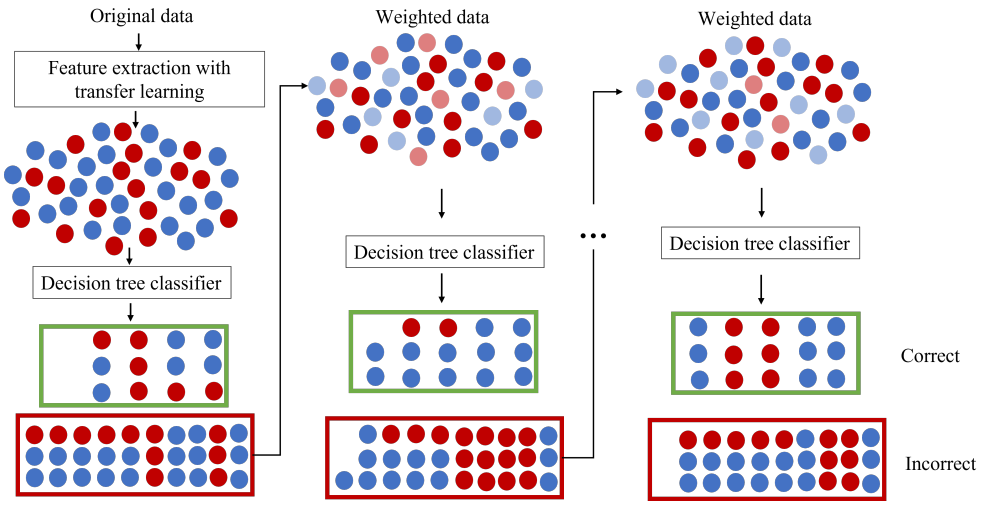


Figure 3. Example of the boosting algorithm, where the red and blue dots represent the normal and abnormal cases, and the green and red boxes represent the correct and incorrect predictions, respectively. As the model progresses, the decision tree classifier represents the weak learners, and the faded blue and red colors represent the images that have less weight because they have already been predicted correctly. In the same fashion as the stacking method, the first part represents transfer learning for feature extraction so the data set can be transformed into a more manageable state for ML algorithms. This process will iterate until the complete training data fit without error or to a specified number of estimators, which in our case is set to 200.

In the context of our work, the blending of boosting with deep learning mirrors our approach in stacking. Firstly, feature extraction through transfer learning is carried out before introducing the models into the boosting framework. As with previous methods, VGG-16 and Inception are the chosen architectures for this feature extraction. While we have previously touched on the strategy of amalgamating weak learners to cultivate a stronger one, here, our attention is on adaptive boosting (AdaBoost), which holds the distinction of being among the earliest boosting techniques effectively utilized in classification problems.

Referencing Figure 3, AdaBoost employs a decision tree as its weak classifier. This choice is not arbitrary; the decision tree is a frequently favored weak classifier for AdaBoost. During the training phase, the user is required to define the iteration count. In our study, this count is set to 200, meaning that the process will iterate until either the entire training data set fits without error or until it reaches this maximum count.

3.2. Hardware Specifications

One of the most notorious things about deep learning algorithms, outside of needing a large number of data, is that they often require hardware setups that can withstand the complexity of their models. This work began on a regular computer where the Alexnet and VGG-16 architectures were trained and tested. Alexnet was trained and tested without trouble, but VGG-16 resulted in a time-costly architecture for our features. When trying the Inception net, it proved too complex for the computer. So, the experiments were later performed in a better-equipped computer. The features of the computer used for all of the experiments shown in this work are as follows:

- CPU: AMD Ryzen 7 5800× 3.80 GHz;
- GPU: AMD Radeon RX6900XT AsRock;
- Motherboard: GIGABYTE Aorus Elite;
- OS: Windows 10.

3.3. Data Sets

In surveys of medical imaging [8,14], the most common benchmark database for mammography is the DDSM (Digital Database for Screening Mammography) [41]. It consists of 2620 scanned film mammography studies containing normal, benign, and malignant cases with verified pathology information. Notice that the database maintained by the University of South Florida is currently obsolete since the images are compressed with lossless JPEG format (.LJPEG), an encoding generated by broken software. This problem was fixed by implementing the improved DDSM, with all images now in a .PNG format [42].

The other database used for this work is the private data set YERAL, which comes from a Mexican hospital and was revised by the FUCAM (Fundacion de Cancer de Mama), a private nonprofit institution in Mexico and Latin America. FUCAM offers comprehensive treatment and specialized breast cancer follow-up through its highly specialized hospital unit in Mexico City. This set has 641 images with confirmed anomalies and 302 images without anomalies.

In medical imaging studies that use deep learning, one remarkable observation from various surveys is that a vast majority of these studies do not present results validated externally with genuine clinical data. In those few studies that do use a validation set, authors often gauge the model's performance against the identical sample [8,12]. To circumvent this shortcoming, our study prioritized the appropriate use of data sets for training, development, and testing evaluations.

We refer to our last validation set as the test set; it contains only images from the target data set YERAL since this work focuses primarily on giving the hospital experts a good computer-aided diagnosis tool. The standard practice of distributing the training and developing sets was considered. We adopted the standard procedure of segregating the training and development sets. Many machine learning sources maintain that the training and development sets should invariably include at least a subset of the images the model will eventually classify [39,40,43,44]. Due to data limitations, our test set comprises 50 images showcasing anomalies and another 50 without, all sourced from the YERAL data set.

Combining the updated DDSM data set with the remaining images from the YERAL data set, we obtained 2594 images without anomalies and 8401 with anomalies. This imbalance gave way to a noticeable bias and too many false positives on the images without anomalies. Since accuracy and F1 score are the most popular metrics in medical imaging,

it was essential to balance the data set correctly to avoid any bias and improve overall accuracy [12,45]. Some anomaly images (only from the DDSM data set) were randomly removed from the data set. In the end, the balanced data set consisted of 2594 normal images and 2900 images with anomalies. Table 1 shows the impact of a balanced data set in popular metrics. For future reference, the balanced data set for the training and developing sets will be referred to as DDSM_YERAL, and the additional validation test set as YERAL.

Table 1. Comparison of the most popular metrics from the validation data set of an unbalanced data set to a balanced one. We used a simple Alexnet architecture using the same weights described in [15].

	Unbalanced Data Set		**Balanced Data Set**	
	Normal	Anomaly	Normal	Anomaly
Precision	64%	90%	82%	86%
Recall	82%	78%	83%	85%
F1 score	76%	73%	86%	82%
Accuracy	79%		84%	

As mentioned before, we used various state-of-the-art methodologies to solve different image recognition problems in the medical field. Some studies show the importance of adequately evaluating results and choosing suitable metrics to score them [12,35].

Finally, a critical decision in any ML model is the percentage of data used for the training and validation sets. There is no standardized way to choose the percentage, so in this work, the training set contains 80% of the total of images included in DDSM_YERAL, and the validation set contains 20%.

3.4. Image Processing

In medical imaging, preprocessing operations are usually required before data analysis and feature extraction; there is no exception in this work. The mammograms in the YERAL data set contained a black mask with information about the patients, so it was essential to remove all of it. We removed the black mask using Python libraries. (See https://github.com/mayraberrones94, accessed on 19 May 2023, for code details).

Notice that the resulting images were all mismatched in size, and CNN architectures require them to be all the same size for the input process. We used OpenCV, a popular Python library in computer vision, to normalize the image size to 224 × 224 pixels. We chose this size because when using popular CNN architectures for transfer learning, they all require the input images to be of this size.

While training the CNN models, we used the Keras library to perform data augmentation. Data augmentation is a well-known practice to improve the accuracy of the training models and reduce overfitting [46]. The images are temporarily stored in the computer's memory during this process. When a training iteration ends, those newly generated images are discarded to make room for the new batch of images. This practice is highly recommended for cases where hardware is limited. It is worth mentioning that data augmentation was only used in the training and developing set and not in the test set. The intention was to avoid oversampling our validation set.

Finally, feature extraction is another common practice in image processing by convolutional neural networks because it allows us to use the power of a deep learning tool and apply its results to more traditional ML algorithms. The dimensionality reduction on the data, using a pretrained network as feature extraction, allows the input data to move forward in the net and stop at a prespecified layer. The output of this process will not be a prediction of the image, but the learned features from the CNN, from which we can train a standard ML model. Therefore, this step becomes essential because we use weak learners for the stacking and boosting ensemble algorithms. Notice that we used transfer

learning with the VGG-16 net and Inception net for feature extraction; the comparison of both results can be seen in the Results section.

4. Results

In the literature on deep learning methods for medical imaging, it was noticed that the most common methods for breast cancer were CNN architectures, most notably VGG-16, Alexnet, and Inception V3. The VGG-16 and Inception architectures are often utilized with the help of transfer learning. The experiment showed promising results that went from 80% accuracy to 95%. Table 2 shows that the results obtained in the training and developing sets, using the balanced DDSM_YERAL data, are similar to those shown in other articles. However, in the test section where we use the YERAL validation set, the accuracy and all the other metrics have a more significant disparity.

Table 2 also shows the loss on training and developing sets. These results highlight that the behavior of the training was expected and did not overfit the data. After these results, there was a revision of the other parameters that could have potentially caused the low accuracy in the test set. Specialists from the FUCAM confirmed that the YERAL data set had no issue and that all images were labeled correctly. Moreover, comparing the hardware requirements from the few articles that had them, they were not far from the ones used in this work.

In the Background section, we divided CNN architectures into two broad categories. The first is the shallow and linear models, typified by Alexnet and VGG-16. The second encompasses the deeper model types like Inception and other architectures, which incorporate built-in modules, making them inherently more intricate.

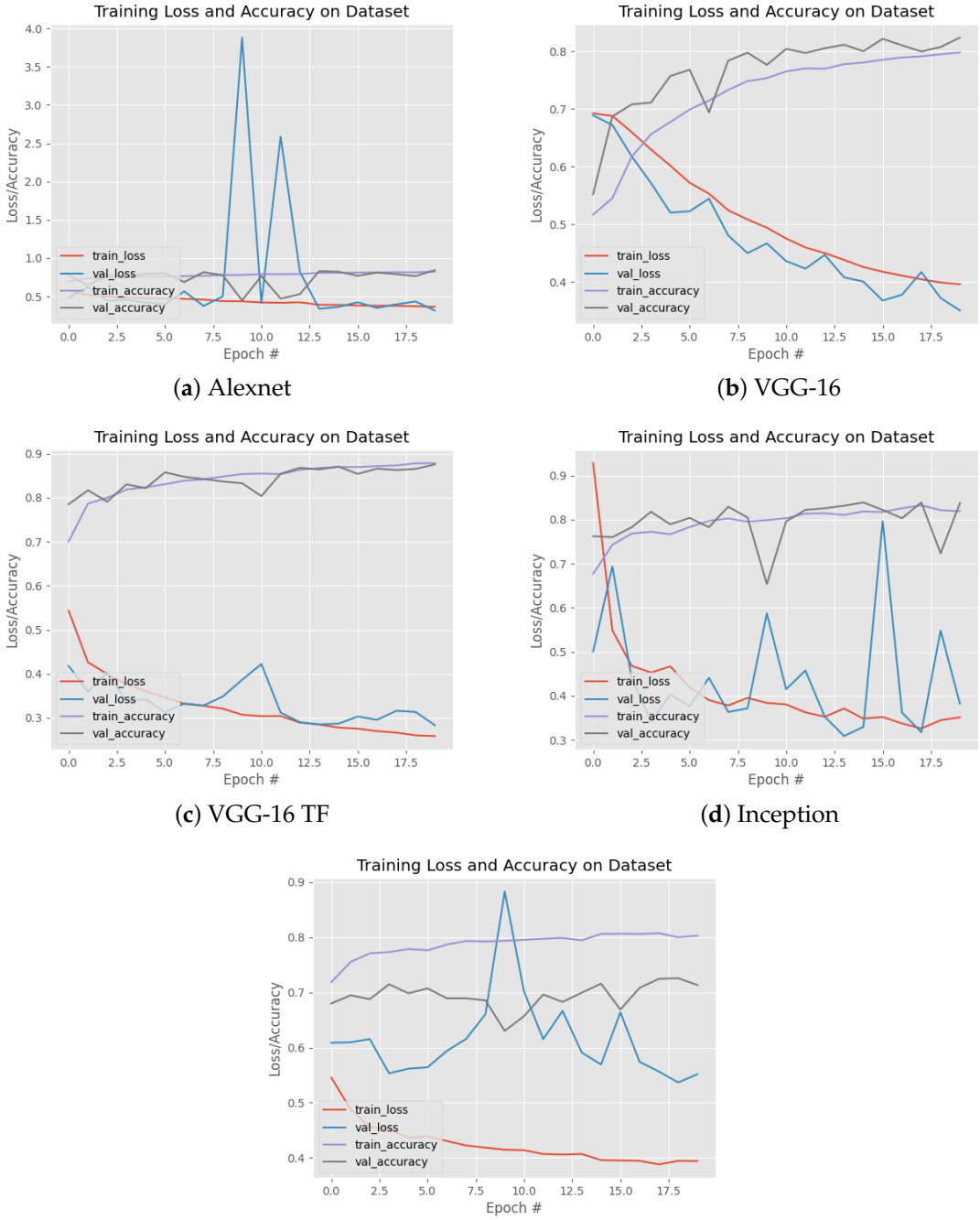
A close inspection of Table 2 reveals that the VGG-16, when incorporated with transfer learning, stands out by maintaining only a nominal gap between the accuracy of the developing and test sets. Contrarily, the Inception architecture, while registering impressive metrics on the training and developing sets, unfortunately fails when it comes to the test set. This pattern suggests that while deeper networks have their merits, they might not always be the optimal choice, and conventional ML algorithms still hold substantial value.

Moreover, the interrelation between Table 2 and Figure 4 provides deeper insights into the training process. The training was executed on the balanced data set amalgamating the DDSM and YERAL sets. Despite the training and developing sets exhibiting standard accuracy and loss rates as shown in Table 2, the test set from the YERAL data set depicted in Figure 4 does not quite match up to the anticipated accuracy outcomes.

Furthermore, Figure 4 delineates the training trajectory for a range of popular CNN architectures. Remarkably, both models rooted in VGG-16 (including its transfer learning counterpart) display the steadiest results. A noteworthy difference between them lies in the training duration. The standalone VGG-16 model took roughly 45 min per epoch, whereas its transfer learning variant achieved the same in under 10 min.

Table 2. Results of popular CNN architectures. Rows 1, 2, and 4 represent the architecture introduced in the literature. Rows 3 and 5 represent the architectures combined with transfer learning, keeping the last pooling layers unfrozen. The gray highlighted section shows the accuracy of our target data set and the metric we seek to improve.

		DDSM_YERAL				YERAL			
		Training Set		Developing Set		Test Set			
	Model	Loss	Acc	Loss	Acc	Acc	Prec	Rec	F1
1	Alexnet	0.36	82%	0.31	84%	60%	57%	76%	65%
2	VGG-16	0.39	79%	0.35	82%	52%	51%	90%	65%
3	VGG-16 TF	0.25	87%	0.28	87%	79%	82%	74%	77%
4	Inception	0.35	81%	0.38	83%	57%	53%	93%	69%
5	Inception TF	0.39	80%	0.55	71%	55%	52%	90%	66%



(**a**) Alexnet (**b**) VGG-16

(**c**) VGG-16 TF (**d**) Inception

(**e**) Inception TF

Figure 4. Detailed performance of the popular CNN architectures used in this study. Each image shows the accuracy and loss for each epoch in the training and validation process. Subfigures (**a**–**c**) are related to the classical arquitectures, while subfigures (**d**,**e**) correspond to the ones with transfer learning.

Table 3 shows the results from all three ensemble algorithms explained in the Methodology section, bagging, stacking, and boosting. As one can see in Figure 1, each bagging iteration uses the CNN several times, from scratch, with a different distribution of images. Therefore, due to their simple architecture and performance, we implemented the Alexnet architecture and VGG-16 with transfer learning for the bagging algorithm. In addition, since Inception performed poorly when using all the images, it would hardly be better if it received even fewer images.

For the stacking, we carried out a comparison between VGG-16 and Inception with transfer learning. Alexnet was not considered for transfer learning because of its simplicity. The same thing was repeated for the boosting ensemble since it also needs feature extraction to feed the algorithm.

Table 3. Results of the expected accuracy of the ensemble algorithms and the results of the test set predictions of that model. The highlighted gray area represents the accuracy of our test data set and it is where one can see the improvement from previous results.

	DDSM_YERAL Developing Set	**YERAL Test Set**			
Model	Acc	Acc	Prec	Rec	F1
Alexnet_Bagging	84%	70%	67%	76%	71%
VGG-16 TF_Bagging	91%	82%	79%	86%	82%
Stacking VGG-16 TF	84%	78%	85%	78%	77%
Stacking Inception TF	82%	74%	78%	74%	73%
AdaBoost VGG-16 TF	83%	76%	82%	76%	75%
AdaBoost Inception TF	80%	71%	75%	71%	70%

As one can see, VGG-16 reaches the best performance in each feature extraction ensemble with transfer learning. The best overall performance was obtained with the bagging algorithm using the VGG-16 with transfer learning, and it reached 85% accuracy. Comparing the results from Table 3 with the first experiment shown in Table 2, we see that VGG-16 with transfer learning achieved an accuracy of 79%. After combining it with an ensemble algorithm, it improved by 6%. Similar behavior is observed in the remaining models; if we compare Tables 2 and 3, we can see improvements in all of them.

5. Conclusions

Deep learning continues to evolve rapidly, with the promise of even more advanced methodologies on the horizon. The core challenge in medical imaging is not just about the sophistication of models but about their applicability and robustness across diverse data sets. A significant opportunity in current research is the limited generalizability of many state-of-the-art models, which, while performing excellently on benchmark data sets, do not perform properly on real-world, practical data sets like YERAL. Addressing this gap was a central theme of our study. The computational science of medical imaging demands stringent standardization in experimentation and result representation. Furthermore, it is vital to distinguish between standard computational vision algorithms and medical imaging, recognizing their unique challenges. As demonstrated in this paper, the way forward may not necessarily lie in exclusively using the latest deep learning tools. Instead, our research emphasizes the potential of adapting and molding these tools to fit specific requirements. We have showcased that by integrating advanced deep learning techniques with what many consider a traditional approach, ensemble learning algorithms can achieve superior performance across diverse data sets.

In our research journey, we initiated our experimentation by assessing the efficacy of the most prevalent methodologies in medical image classification. However, what set our work apart was our commitment to delivering an efficient and accurate tool tailored for specialists, particularly in light of the unique challenges presented by the YERAL data set. This dedication led us to revisit and repurpose certain strategies, which, although perceived as dated by contemporary standards, unveiled significant potential in enhancing the precision of modern techniques.

Highlighting our findings, the bagging ensemble model, integrated with VGG-16 using transfer learning, emerged as a game changer, far surpassing the standalone performance of VGG-16 with transfer learning. What is equally noteworthy is the promising accuracy achieved by both the stacking model and the boosting algorithm. Their incorporation into the realm of image classification, especially when used with the potent feature extraction

capabilities of transfer learning, suggests a paradigm shift, challenging conventional notions and offering renewed perspectives on the subject.

The relationship between traditional machine learning algorithms and select deep learning methodologies has come as an unexpected yet potent combination, often outperforming the interest of more intricate networks. Notably, these more elaborate architectures present their own challenges, particularly when users find themselves with hardware limitations, insufficient training data, or difficulties typical of medical images.

One hypothesis that our work highlights is the challenge of generalizing medical imaging problems solely using deep learning. The importance of this challenge often stems from the necessity of turning to benchmark data sets to boost data volumes. These data sets frequently hail from a distribution or demographic that is significantly different from the primary data set intended for diagnostic purposes. While diversifying data set distributions might be an asset in generic computer vision scenarios, it introduces potential pitfalls in medical imaging. Such diversity can inadvertently inject noise, complicating the modeling process.

In light of our findings and the challenges highlighted, future exploration would involve developing our novel methodology on data sets from varied hospitals, each serving a different demographic than YERAL. Such a pursuit would gauge the adaptability and robustness of our approach, potentially reaffirming its promise as a solution to the generalizability issue that plagues many current models in the medical imaging domain.

Author Contributions: Conceptualization, M.C.B.-R., M.A.S.-A. and C.C.-O.; methodology, M.C.B.-R.; validation, M.C.B.-R.; formal analysis, M.C.B.-R.; investigation, M.C.B.-R.; data curation, C.C.-O.; original draft preparation, M.C.B.-R. and M.A.S.-A.; review and editing, M.A.S.-A. and C.C.-O.; supervision, M.A.S.-A. and C.C.-O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Part of the data used in this study is publicly available at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=22516629 (accessed on 1 August 2023), and the YERAL data set is not publicly available.

Acknowledgments: We want to thank the FUCAM for their support in reviewing our work and being available to answer our medical questions. The first author thanks CONACYT for the scholarship to carry out her doctorate studies. studies. Finally, a very special thanks to Satu Elisa Schaeffer and Sara Elena Garza Villarreal for their valuable feedback while carrying out this project.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	convolutional neural networks
ML	machine learning
DL	deep learning
TF	transfer learning

References

1. Fujiyoshi, H.; Hirakawa, T.; Yamashita, T. Deep learning-based image recognition for autonomous driving. *IATSS Res.* **2019**, *43*, 244–252. [CrossRef]
2. Li, Y. Research and Application of Deep Learning in Image Recognition. In Proceedings of the 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), Beijing, China, 5–9 January 2022. [CrossRef]
3. Winston, J.J.; Hemanth, D.J.; Angelopoulou, A.; Kapetanios, E. Hybrid deep convolutional neural models for iris image recognition. *Multimed. Tools Appl.* **2021**, *81*, 9481–9503. [CrossRef]

4. García-Holgado, A.; Vázquez-Ingelmo, A.; Alonso-Sánchez, J.; García-Peñalvo, F.J.; Therón, R.; Sampedro-Gómez, J.; Sánchez-Puente, A.; Vicente-Palacios, V.; Dorado-Díaz, P.I.; Sánchez, P.L. KoopaML, a Machine Learning platform for medical data analysis. *J. Interact. Syst.* **2022**, *13*, 154–165. [CrossRef]
5. Rüttgers, M.; Waldmann, M.; Schröder, W.; Lintermann, A. A machine-learning-based method for automatizing lattice-Boltzmann simulations of respiratory flows. *Appl. Intell.* **2022**, *52*, 9080–9100. [CrossRef]
6. Gambella, C.; Ghaddar, B.; Naoum-Sawaya, J. Optimization problems for machine learning: A survey. *Eur. J. Oper. Res.* **2021**, *290*, 807–828.
7. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* **2021**, *13*, 1590. [CrossRef]
8. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297. [CrossRef]
9. Snyder, M.; Zhou, W. Big data and health. *Lancet Digit. Health* **2019**, *1*, e252–e254. [CrossRef]
10. Health, T.L.D. Walking the tightrope of artificial intelligence guidelines in clinical practice. *Lancet Digit. Health* **2019**, *1*, e100. [CrossRef]
11. Zendel, O.; Murschitz, M.; Humenberger, M.; Herzner, W. How Good Is My Test Data? Introducing Safety Analysis for Computer Vision. *Int. J. Comput. Vis.* **2017**, *125*, 95–109. [CrossRef]
12. Varoquaux, G.; Cheplygina, V. How I failed machine learning in medical imaging–shortcomings and recommendations. *arXiv* **2021**, arXiv:2103.10292 .
13. Zhang, L.; Wang, H.; Li, Q.; Zhao, M.H.; Zhan, Q.M. Big data and medical research in China. *BMJ* **2018**, *5*, j5910. [CrossRef] [PubMed]
14. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 5386. [CrossRef]
16. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Brussels, Belgium, 14–17 June 2016; pp. 770–778. [CrossRef]
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
19. Bi, L.; Feng, D.D.; Fulham, M.; Kim, J. Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. *Pattern Recognit.* **2020**, *107*, 107502. [CrossRef]
20. Lin, C.H.; Lin, C.J.; Li, Y.C.; Wang, S.H. Using generative adversarial networks and parameter optimization of convolutional neural networks for lung tumor classification. *Appl. Sci.* **2021**, *11*, 480. [CrossRef]
21. Morid, M.A.; Borjali, A.; Del Fiol, G. A scoping review of transfer learning research on medical image analysis using ImageNet. *Comput. Biol. Med.* **2021**, *128*, 104115. [CrossRef]
22. Kora, P.; Ooi, C.P.; Faust, O.; Raghavendra, U.; Gudigar, A.; Chan, W.Y.; Meenakshi, K.; Swaraja, K.; Plawiak, P.; Acharya, U.R. Transfer learning techniques for medical image analysis: A review. *Biocybern. Biomed. Eng.* **2021**, *128*, 104115.
23. Rahman, T.; Chowdhury, M.E.; Khandakar, A.; Islam, K.R.; Islam, K.F.; Mahbub, Z.B.; Kadir, M.A.; Kashem, S. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Appl. Sci.* **2020**, *10*, 3233. [CrossRef]
24. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y.; Oleiwi, S.R. Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study. *Appl. Sci.* **2020**, *10*, 4523. [CrossRef]
25. Yang, A.; Yang, X.; Wu, W.; Liu, H.; Zhuansun, Y. Research on feature extraction of tumor image based on convolutional neural network. *IEEE Access* **2019**, *7*, 24204–24213. [CrossRef]
26. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **2022**, *22*, 69.
27. Hakak, S.; Alazab, M.; Khan, S.; Gadekallu, T.R.; Maddikunta, P.K.R.; Khan, W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Gener. Comput. Syst.* **2021**, *117*, 47–58.
28. Chakraborty, D.; Narayanan, V.; Ghosh, A. Integration of deep feature extraction and ensemble learning for outlier detection. *Pattern Recognit.* **2019**, *89*, 161–171. [CrossRef]
29. Taspinar, Y.S.; Cinar, I.; Koklu, M. Classification by a stacking model using CNN features for COVID-19 infection diagnosis. *J. X-ray Sci. Technol.* **2022**, *30*, 73–88. [CrossRef]
30. Müller, D.; Soto-Rey, I.; Kramer, F. An Analysis on Ensemble Learning optimized Medical Image Classification with Deep Convolutional Neural Networks. *arXiv* **2022**, arXiv:2201.11440.
31. Kandel, I.; Castelli, M.; Popovič, A. Comparing stacking ensemble techniques to improve musculoskeletal fracture image classification. *J. Imaging* **2021**, *7*, 100.
32. Haq , I.U.; Ali, H.; Wang, H.Y.; Lei, C.; Ali, H. Feature fusion and Ensemble learning-based CNN model for mammographic image classification. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 3310–3318. [CrossRef]

33. Das, A.; Mohanty, M.N.; Mallick, P.K.; Tiwari, P.; Muhammad, K.; Zhu, H. Breast cancer detection using an ensemble deep learning method. *Biomed. Signal Process. Control.* **2021**, *70*, 103009. [CrossRef]
34. Barhoom, A.M.; Almasri, A.; Abu-Nasser, B.S.; Abu-Naser, S.S. Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. *Int. J. Eng. Inf. Syst.* **2022**, *6*, 2972.
35. Reinke, A.; Eisenmann, M.; Tizabi, M.D.; Sudre, C.H.; Rädsch, T.; Antonelli, M.; Arbel, T.; Bakas, S.; Cardoso, M.J.; Cheplygina, V.; et al. Common limitations of image processing metrics: A picture story. *arXiv* **2021**, arXiv:2104.05642.
36. Kunapuli, G. *Ensemble Methods for Machine Learning*; Simon and Schuster: New York, NY, USA, 2022.
37. Sagi, O.; Rokach, L. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, 1249. [CrossRef]
38. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
39. Raschka, S.; Mirjalili, V. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*; Packt Publishing Ltd.: Birmingham, UK, 2019.
40. Ng, A. Machine Learning Yearning: Technical Strategy for AI Engineers in the Era of Deep Learning. 2019. Available online: https://www.mlyearning.org (accessed on 15 August 2022).
41. Lee, R.S.; Gimenez, F.; Hoogi, A.; Miyake, K.K.; Gorovoy, M.; Rubin, D.L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **2017**, *4*, 177. [CrossRef]
42. Lekamlage, C.D.; Afzal, F.; Westerberg, E.; Chaddad, A. Mini-DDSM: Mammography-based Automatic Age Estimation. In Proceedings of the ACM 2020 3rd International Conference on Digital Medicine and Image Processing, Berlin, Germany, 12–16 November 2020. [CrossRef]
43. Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York, NY, USA, 2021.
44. Brownlee, J. *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*; Machine Learning Mastery: Vermont, VIC, Australia, 2016.
45. Ng, A. Machine Learning Yearning Volume 139. 2017. Available online : https://github.com/ajaymache/machine-learning-yearning (accessed on 1 August 2023).
46. Hepsağ, P.U.; Özel, S.A.; Yazıcı, A. Using deep learning for mammography classification. In Proceedings of the IEEE 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 418–423.