Enhancing Diabetic Retinopathy Detection Using CNNs with Dimensionality Reduction Techniques and K-Nearest Neighbors Ensembles

Chaymaa Lahmar[1] and Ali Idri[1,2]
[1]*Software Project Management Research Team, ENSIAS, Mohammed V University in Rabat, Morocco*
[2]*Mohammed VI Polytechnic University Benguerir, Morocco*

Keywords: Diabetic Retinopathy, Homogeneous Ensembles, Machine Learning, Deep Learning, Feature Selection, Fundus Images.

Abstract: Diabetic Retinopathy (DR) is the most frequent cause of blindness and visual impairment among working-age adults in the world. Machine learning (ML) and deep learning (DL) techniques are playing an important role in the early detection of DR. This paper proposes a new homogeneous ensemble approach constructed using a set of hybrid architectures, as base learners, and two combination rules (hard and weighted voting) for referable DR detection using fundus images over the Kaggle DR, APTOS and Messidor-2 datasets. The hybrid architectures are created using seven deep feature extractors (DenseNet201, InceptionResNetV2, MobileNetV2, InceptionV3, VGG16, VGG19, and ResNet50), six dimensionality reduction techniques (Principal component analysis, Select from model feature selection, Recursive feature elimination with cross-validation, Factor analysis, Chi-Square test, and Low variance filter), and k-nearest neighbors algorithm (KNN) for classification. The results showed the importance of the proposed approach considering that it outperformed its base learners, and achieved an accuracy value of 92.47% for the Kaggle DR dataset, 89.59% for the APTOS dataset, and 82.03% for the Messidor-2 dataset. The experimental results demonstrated that the proposed approach is impactful for the detection of referable DR, and thus represents a promising tool to assist ophthalmologists in the diagnosis of DR.

1 INTRODUCTION

By 2040, diabetes mellitus (DM), generally known as diabetes, is predicted to affect 642 million individuals worldwide, with nearly 75% living in low and middle-income countries (Shaw, Sicree, & Zimmet, 2010). Over time, poorly controlled diabetes can harm the kidneys, blood vessels, heart, and eyes. Diabetic retinopathy (DR) is a serious sight-threatening complication of diabetes; it is the most frequent cause of blindness in working-age adults, affecting one in every three people with diabetes (Wong & Sabanayagam, 2020).

Machine learning (ML) and deep learning (DL) techniques have attained excellent diagnostic performance in identifying serious medical disorders (Islam, Yang, Poly, Jian, & (Jack) Li, 2020; Lahmar & Idri, 2022; Litjens et al., 2019). In particular, the use of ML and DL techniques in diagnosing various ophthalmic diseases, such as diabetic retinopathy, is drawing enormous interest (Han, 2022). ML techniques proved their potential by helping ophthalmologists obtain accurate diagnoses by detecting retinal anomalies when using fundus images (Islam et al., 2020). In addition, DL methods have recently gained popularity as one of the most efficient techniques for enhancing performance in medical image analysis (Islam et al., 2020; Litjens et al., 2019). Multiple studies used hybrid architectures that incorporated the advantages of DL techniques for feature extraction and ML techniques for classification (Lahmar & Idri, 2023; Zhang et al., 2019). A common method to improve the performance of hybrid architectures is to make use of ensemble learning methods (Bellemo et al., 2019; Gurcan, Beyca, & Dogan, 2021; Jinfeng, Qummar, Junming, Ruxian, & Khan, 2020).

Ensemble learning methods have been used in many studies in the medical field to improve the predictions given by the base learners or the single models which can be a DL model, a ML model, or a

Lahmar, C. and Idri, A.
Enhancing Diabetic Retinopathy Detection Using CNNs with Dimensionality Reduction Techniques and K-Nearest Neighbors Ensembles.
DOI: 10.5220/0012191900003598
In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2023) - Volume 1: KDIR*, pages 315-322
ISBN: 978-989-758-671-2; ISSN: 2184-3228

hybrid architecture (Sagi & Rokach, 2018). Ensemble methods aim to combine multiple base learners that are accurate and diverse in order to overcome their weaknesses and merge their advantages (Bellemo et al., 2019; Gurcan et al., 2021; Jinfeng et al., 2020). In general, to build an ensemble model, we start by selecting the base learners, then we use each base learner to predict the results, and finally, we use a combination rule to aggregate the predictions. More specifically, the ensemble model can be built using the combination rules after the base learners have been created. Hard voting, also known as majority voting, and weighted voting, often known as soft voting, are two of the most common combinations utilized for creating ensemble models. In the case of hard voting, the base learners vote for one class, and the final output is the class label that obtain more than half of the votes (Hosni, Abnane, Idri, Carrillo de Gea, & Fernández Alemán, 2019). While using weighted voting, base learners are given weights, with the best base learner receiving the highest weight.

In this paper, we propose a new homogeneous ensemble approach constructed using a set of hybrid architectures, as base learners, and two combination rules (hard and weighted voting). We used 5-fold cross validation, four performance criteria (recall, precision, F1-score and accuracy), the Scott Knott (SK) statistical test (Worsley, 2010) and the Borda Count voting method (García-Lapresta & Martínez-Panero, 2002) to assess the proposed ensembles. In order to create the hybrid architectures used as base learners, we used seven DL techniques for feature extraction (InceptionV3, DenseNet201, ResNet50, MobileNetV2, InceptionResNetV2, VGG16 and VGG19), six dimensionality reduction techniques to reduce the size of features, and the KNN classifier. As for the dimensionality reduction techniques, we used the principal component analysis (PCA), Select from model (SFM) feature selection, Recursive feature elimination with cross-validation (RFE-CV), Factor analysis (FA), Chi-Square test (Chi2), and Low variance filter (LVF). Note that the SFM and RFE-CV feature selection techniques needs to be used alongside an estimator that assigns importance weights to features; therefore, we implemented these two techniques based on SVM since they were widely used with this estimator (Remeseiro & Bolon-Canedo, 2019). To create the proposed ensembles: (1) The hybrid architectures, were compared in order to identify the best-performing ones (2) Using the Borda Count voting method, the hybrid architectures selected from the previous step, were ranked according to the four-performance metrics, (3) The Top ranked 2, 3, 4, 5, 6 and 7 hybrid architectures were used to create the ensembles. Hence, we obtain 12 ensembles for each dataset (6 ensembles with weighted voting + 6 ensembles with hard voting) and 36 ensembles across the three datasets (12 ensembles * 3 datasets). The DL techniques used in this study as well as the KNN classifier were selected since they provide high classification accuracy values in DR detection (Islam et al., 2020; Lahmar & Idri, 2021). As for the dimensionality reduction techniques, they were chosen since they are highly applied in medical image analysis (Remeseiro & Bolon-Canedo, 2019). This study addresses four research questions (RQs) to that end:

- **(RQ1): What is the overall performance of dimensionality reduction techniques in DR detection?**
- **(RQ2): Does the proposed ensembles perform better than their singles?**
- **(RQ3): Does increasing the number of base learners affect the classification performance of the proposed ensembles?**
- **(RQ4): Out of the two combination rules, which one is the best performing?**

The primary contributions of this study are as follows:

1) Proposing a new homogeneous ensemble approach using the most recent DL techniques for feature extraction, dimensionality reduction techniques to reduce the size of features and the KNN classifier with two combination rules.
2) Assessing whether the proposed ensembles outperform their base learners.
3) Assessing whether the proposed ensembles created using the weighted voting outperform the ones using hard voting.

This paper is organized as follows: Section 2 presents some related studies using ensemble learning for DR detection. The data preparation process is summarized in Section 3. Section 4 presents the empirical methodology followed in this study. Section 5 discusses the empirical findings. The study's threats of validity are presented in Section 6. Finally, Section 7 presents the conclusion and future works.

2 RELATED WORKS

Motivated by the success of ensemble learning methods, many studies in the medical field used

ensemble-based architectures to improve the predictions given by the base learners. For instance, the study (Bellemo et al., 2019) proposed an ensemble model combining a ResNet architecture and a VGGNet architecture using weighted voting for DR classification. The ensemble model was trained on a private dataset and the results showed the potential of ensemble learning for diabetic retinopathy detection. In (Jinfeng et al., 2020), the authors created an homogeneous ensemble model based on bagging using DenseNet121 with different configurations to create 3 base learners for diabetic retinopathy classification, the ensemble model was evaluated using the Kaggle DR dataset. The proposed ensemble reached an accuracy equal to 80%. Finally, in the study (Zhang et al., 2019), the authors developed an ensemble approach, for DR identification and grading. To create the ensemble, they used three models as base learners and they averaged the softmax scores of all models. For the ensemble used for the identification, they used Xception, Inception, InceptionResNet as feature extractors and a standard deep neural network (SDNN) for the classification part. And for the grading, they used DenseNet169, DenseNet201 and Resnet50 as feature extractor and SDNNS for the classification part. Experiment results showed the effectiveness of the proposed model since it provides reliable detection results with high sensitivity and specificity values.

3 DATA PREPARATION

In this study, we used three public datasets: (1) The APTOS dataset which includes 3662 fundus photographs. (2) The Kaggle DR dataset which contains 35,126 fundus photographs. Note that we used 5000 images from the Kaggle DR dataset in order to train and evaluate our models. And (3) the Messidor-2 dataset which includes 1748 fundus photographs. Besides, it is noteworthy that the grades of DR in the three datasets are on a scale of 0 to 4 representing the 5 grades of DR. And since the referable DR is the study's target variable, we relabeled the data from a 0 to 4 scale to a 0 to 1 scale, where 0 is defined as "no referable DR" and 1 is defined as "referable DR" (Lahmar & Idri, 2023). Knowing that referable DR is represented as mild non-proliferative DR or worse, and/or diabetic macular edema. And as the quality of fundus images is crucial for any automatic method, we used a variety of preprocessing techniques to enhance the quality of the images. Figure 1 shows the preprocessing techniques used in this study. Finally, we used data augmentation techniques to generate two or three new images from each fundus image since the number of images in the three datasets was unbalanced.

4 EMPIRICAL DESIGN

In this section, we present the experimental process. Then, the experiment configuration is described. And finally, we provide the abbreviations that are used to refer to the proposed ensembles.

4.1 Experimental Process

In this subsection, we present the methodology used to conduct the experiment's empirical evaluations. It involves six steps that consist of:

1. Design 42 hybrid architectures using the KNN classifier, 7 feature extractors (DenseNet201, MobileNet_V2, Inception_V3, VGG16, VGG19, InceptionResNet_V2, and ResNet50), and 6 dimensionality reduction techniques (PCA, FA, LVF, Chi2, SFM, and RFECV) for each dataset.
2. For each dataset and each feature extraction technique, select the best performing dimensionality reduction technique using the SK test and Borda count voting method. To identify the dimensionality reduction techniques that will be used alongside the architectures used as base learners.

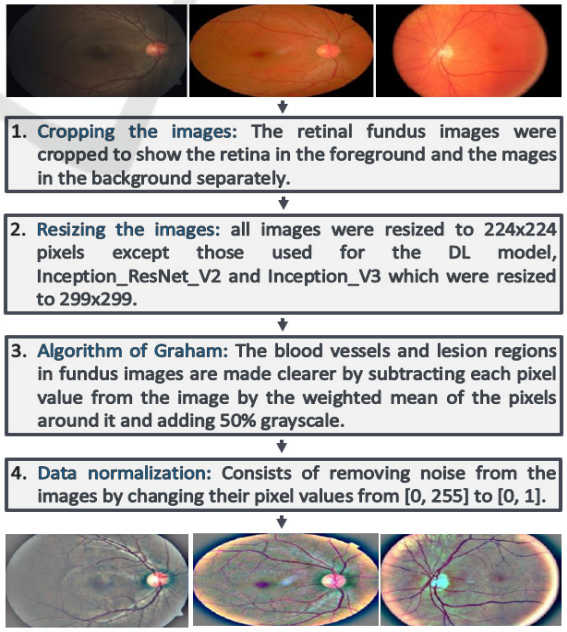

Figure 1: Data preparation process.

3. For each dataset, sort the hybrid architectures that will be applied as base learners using the Borda Count voting method. The number of base learners is 21 hybrid architectures (7 DL for feature extraction * 1 dimensionality reduction technique * 3 datasets).
4. For each dataset, design the homogeneous ensembles from Top 2 to Top 7 of the ranked hybrid architectures of step 3, using the two combination rules hard and weighted voting. The number of the designed ensembles will be 12 for each dataset (6 ensembles * 2 combination rules) and 36 for all the experiment (6 ensembles * 2 combination rules * 3 datasets).
5. For each dataset, apply SK test based on accuracy on the best 7 deep hybrid architectures of step 3 and the 12 homogeneous ensembles constructed in step 4.
6. For each dataset, apply Borda Count voting method, in terms of four-performance measures (accuracy, precision, recall and F1-score), on the best SK cluster of the previous step.

4.2 Experimental Setup

To build the proposed ensembles, the following experiment configurations were used:

- For feature extraction, we used transfer learning to extract the features with the seven DL models. Then, for each dataset, we created a new dataset of feature vectors, which will be used as input for the dimensionality reduction techniques.
- For the dimensionality reduction techniques, the PCA, SFM, and Chi2 are used with the default configuration of the scikit learn library. For the LVF, the default variance threshold is equal to zero to remove all features with zero variance but it did not remove any feature from the vectors, therefore, we needed to use a threshold value equal to 1 to remove the quasi-constant features. As for the FA, the default number of components is set to the number of features, therefore, we needed to select a number of components, and after several experiment, we obtained the best results using 100 components. And for the RFECV, the default number of steps or the number of features to remove at each iteration is equal to 1, and because of the large dimensions of the vectors of feature, we needed to select a greater number, after several experiments, we obtained the best results, using a number of steps equal to 0.2 which corresponds to the percentage of features to remove at each iteration.
- For the classification, the KNN model is used with the default configuration of the scikit learn library. Note that we stored the predictions of each base learner to create the ensembles.
- For the ensembles, the two-combination methods weighted and hard voting are used to combine the predictions of the base learners. For the weighted voting, we assigned weights for each base learner based on the ranking obtained in step 3 of the empirical design. While no weighs are used with the hard voting.

4.3 Abbreviations

The names of the base learners are shortened as follows: K for KNN. And for the feature extractors: RES for ResNet50, V16 for VGG16, DEN for DenseNet201, INR for InceptionResNetV2, IN for InceptionV3, V19 for VGG19 and MOB for MobileNetV2. For example, KDEN stands for to the base learner created using KNN classifier and DenseNet201 feature extractor.

As for the names of the proposed ensemble, they were abbreviated as follows: E for ensemble, K for KNN, the number of base learners, and the combination method with HV for hard voting and WV for weighted voting. For example, EK7WV refers to the ensemble developed utilizing seven base learners (hybrid architectures) and weighted voting (WV).

5 RESULTS AND DISCUSSION

The results of the empirical evaluations conducted over the three datasets are presented and discussed in this section. The Keras and Tensorflow deep learning frameworks were used to implement the empirical assessments of the base learners and the proposed ensembles in Python using Google's Colab Notebook based on a TPU processor with 8 cores, 35 GB of RAM, and a Linux-based operating system.

5.1 Ranking of the Dimensionality Reduction Techniques

In this step of the experiment, we evaluated the six dimensionality reduction techniques (PCA, FA, LVF, Chi2, SFM, and RFECV), to identify the techniques that will be used with the architectures used as base learners to create the proposed ensembles (RQ1). We started by applying the six dimensionality reduction

techniques to reduce the size of features of each feature extraction technique over the three datasets. Then, to identify the dimensionality reduction techniques that have the same effect on the classification performance, we used the SK test based on accuracy. Finally, according to the four-performance metrics, we used the Borda count to rank the techniques belonging to best SK clusters. The objective is to determine which dimensionality reduction technique has significantly impacted the classification performance of each hybrid architecture. We found that:

- Using the DenseNet201, we obtained 2 SK clusters when using the APTOS dataset where the best one includes the RFECV, SFM, PCA, LVF and FA. As for the Kaggle DR, we found 3 clusters where the best one includes the SFM, RFECV, LVF and PCA. Finally, for the Messidor-2, we found 3 clusters where the best one includes SFM, PCA, RFECV and LVF.
- Using the MobileNetV2, we found 3 SK clusters when using the APTOS dataset where the best one includes only the SFM. As for the Kaggle DR and Messidor-2 datasets, we found 2 clusters where the best one includes all the dimensionality reduction techniques except the Chi2.
- Using the VGG16, we found 2 SK clusters when using the APTOS dataset where the best one includes the SFM, RFECV, PCA, LVF and FA. As for the Kaggle DR, we found 3 clusters where the best one includes the SFM, FA, PCA, RFECV. Finally, for the Messidor-2, we found 2 clusters where the best one includes the SFM, PCA, RFECV and FA.
- Using the VGG19, we found 3 SK clusters when using the APTOS dataset where the best one includes the SFM and RFECV. As for the Kaggle DR, we found 3 clusters where the best one includes the SFM, PCA, RFECV and FA. Finally, for the Messidor-2, we found 2 clusters where the best one includes the RFECV, SFM, LVF and PCA.
- Using the InceptionV3, we found 2 SK clusters when using the APTOS dataset where the best one includes the SFM, RFECV, LVF and PCA. As for the Kaggle DR, we found 2 clusters where the best one includes the SFM, FA, PCA, RFECV. Finally, for the Messidor-2, we found 2 clusters where the best one includes the SFM, PCA, RFECV and LVF.
- Using the ResNet50, we found only 1 SK cluster when using the APTOS dataset. As for the Kaggle DR and Messidor-2 datasets, we found 2 clusters where the best one includes all the dimensionality reduction techniques except the Chi2.
- Using the InceptionResNetV2, we found 2 SK clusters when using the APTOS and Kaggle DR datasets where the best one includes all the dimensionality reduction techniques except the Chi2. Finally, for the Messidor-2, we found 3 clusters where the best one includes SFM, RFECV, PCA, LVF.

Thereafter, depending on the four-performance metrics, we used the Borda count to rank the techniques belonging to best SK clusters. We found that the best dimensionality reduction technique over the three datasets, is the SFM regardless of the feature extractor, except when using the DenseNet201 with the APTOS dataset, the ResNet50 and VGG19 with the Messidor-2 dataset, the best dimensionality reduction technique is the RFECV.

Finally, to select the best dimensionality reduction technique regardless of the dataset, we denoted the number of appearances of each technique in the best SK clusters. In the case of equality, we use the Borda count ranking. We found that the SFM outperformed all the other dimensionality reduction techniques since it appeared in the best SK clusters 21 times and it was ranked first 18 times. Followed by the RFECV, PCA, LVF and FA respectively. Finally, the Chi2 underperformed the other dimensionality reduction techniques.

5.2 Ranking of the Base Learners

In this step, we ranked the architectures constructed using the first ranked dimensionality reduction techniques selected in the previous step using the Borda count method. Therefore, we obtain seven ranked hybrid architectures over each dataset. Note these architectures will be used as base learners to create the proposed ensembles. We found that:

- Using APTOS dataset, KDEN was ranked first, followed by KV16, KV19, KINR, KMOB, KRES and KIN.
- Using Kaggle DR dataset, KDEN was ranked first, followed by KV19, KV16, KMOB, KRES, KIN and KINR.
- Using Messidor-2 dataset, KMOB was ranked first, followed by KDEN, KV16, KV19, KRES, KINR and KIN.

5.3 Evaluation of the Proposed Ensembles and Their Base Learners

In this step, we created the proposed ensembles utilizing the top 2 to 7 architectures according to the ranking we obtained in the previous step, and then, we combined them using weighted and hard voting methods. Note that when using weighted voting, the base learners were assigned weights depending on their rankings. Also, in the event of a tie for the pair combinations in the hard voting, we refer to the ranking of the base learners. As result, we obtained 12 homogeneous ensembles. The ensembles were assessed with regard to three factors: accuracy, the impact of increasing the number of base learners and the combination method. First, we started by using the SK test to compare the ensembles and their base learners in terms of accuracy (RQ2). As illustrated in Figure 2, the SK test detected 4 clusters for the Messidor-2 and APTOS datasets, and 3 clusters for the Kaggle DR dataset. We found that:

- For the three datasets, the base learners belong to the last SK clusters.
- For the three datasets, only the proposed ensembles belong to the best SK clusters.

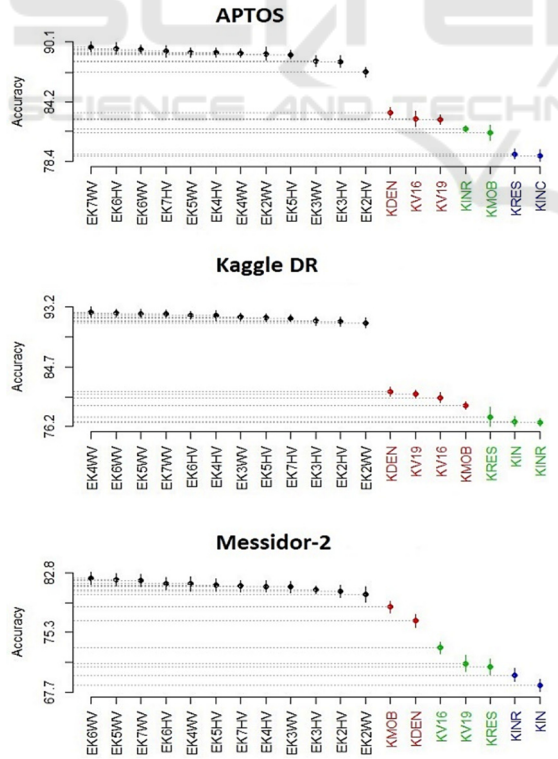

Figure 2: SK test's results over the proposed ensembles and their base learners.

Therefore, we conclude that the proposed ensembles differ significantly from their base learners. As for the number of base learners used to design the ensembles belonging to the best SK clusters, there is no conclusive evidence of the optimal number of base learners.

The Borda Count voting method was then used, based on the results of the four-performance metrics, to choose the best number of base learners (RQ3) as well as the best combination rule to create the ensembles (RQ4). The Borda Count's ranking of the ensembles belonging to the best SK clusters is presented in Table 1, we found that:

- The EK6WV ensemble is ranked first for the Messidor-2 dataset, second for the Kaggle DR dataset and third for the APTOS dataset.
- The EK7WV ensemble is ranked first for the APTOS dataset, fourth for the Kaggle DR dataset, and sixth for the Messidor-2 dataset.
- The EK4WV ensemble is ranked first for the Kaggle DR dataset, sixth for the Messidor-2 dataset, and seventh for the APTOS dataset.
- The EK2WV ensemble is ranked in the last place for the Messidor-2 and Kaggle DR datasets when the EK3HV is ranked in the last place for the APTOS dataset.

Regardless of the dataset, we refer to the Borda Count ranking to identify the best-performing ensembles. We found that the EK6WV ensemble, which occurs one time first, once second, and once third, is the best-performing ensemble, followed by the ensemble EK7WV (1 time first, once fourth and once sixth) and EK4WV (1 time first, once sixth and once seventh). Tables 2-4 show the mean values of the four-performance metrics of the best ensembles over the three datasets.

As a summary, the evaluation's principal findings indicate that the ensembles outperformed their base learners over the three datasets, particularly the ensemble EK6WV followed by EK7WV and EK4WV. Also, we found that the classification performance of the proposed ensembles can be influenced by the number of base learners since, in the three datasets, the ensembles with 2 and 3 base learners are ranked in the last place. Lastly, we found that the ensembles produced by the weighted voting gave the best performance, as shown by the fact that the ensembles produced by this combination rule are ranked in the first place across the three datasets.

Table 1: Borda Count ranking of the ensembles of the best SK cluster over the three datasets. The three first ranked ensembles are identified with three colors: Green for EK7WV, Red for EK4WV, and Orange EK6WV.

Rank	APTOS	Kaggle DR	Rank	Messidor-2
1	EK7WV	EK4WV	1	EK6WV
2	EK6HV	EK6WV	2	EK5WV
3	EK6WV	EK5WV	3	EK4HV
4	EK7HV	EK7WV	4	EK5HV
5	EK4HV	EK3WV	5	EK2HV
6	EK5WV	EK4HV	6	EK4WV
7	EK4WV	EK6HV	6	EK7WV
8	EK2WV	EK5HV	7	EK6HV
9	EK5HV	EK7HV	8	EK3HV
10	EK2HV	EK3HV	9	EK7HV
11	EK3WV	EK2HV	10	EK3WV
12	EK3HV	EK2WV	11	EK2WV

Table 2: Performance metrics values of the best performing ensembles over the APTOS dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EK4WV	88.98	89.00	89.14	89.06
EK6WV	89.36	89.26	89.66	89.45
EK7WV	89.59	89.95	89.31	89.62

Table 3: Performance metrics values of the best performing ensembles over the Kaggle DR dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EK4WV	92.47	99.32	85.52	91.88
EK6WV	92.37	99.32	85.32	91.77
EK7WV	92.25	99.49	84.93	91.61

Table 4: Performance metrics values of the best performing ensembles over the Messidor-2 dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
EK4WV	81.36	94.30	67.21	78.49
EK6WV	82.03	95.64	67.22	78.93
EK7WV	81.79	96.32	64.96	77.56

6 THREATS OF VALIDITY

In this study, a 5-fold cross-validation was used to evaluate the proposed ensembles as an internal threat. The use of ensemble learning with two combination rules (hard and weighted voting) to create the ensembles is another internal threat to this experiment. Further, this work trained and evaluated the proposed ensembles using three publicly available fundus images datasets: APTOS, Messidor-2, and Kaggle DR, for external validity. To validate or refute the results of this work, it will be interesting to repeat the study using different ensemble learning methods, different deep learning techniques for feature extraction, different dimensionality reduction techniques and different machine learning classifiers with various public or private datasets. For the reliability of the classification performance, we evaluated the proposed ensembles using four performance metrics (accuracy, precision, recall, and F1-score), these metrics were chosen because they were frequently used to assess DR classification performance (Islam et al., 2020). Moreover, for the conclusion, SK statistical test and Borda count voting method were used based on the four-performance metrics with equal weights to avoid favoring one performance metric over the others. This strategy was used to determine the best-performing ensembles based on statistical tests.

7 CONCLUSION AND FUTURE WORK

In this study, we discussed the importance of referable DR classification using homogeneous ensemble learning. For that, we created 12 homogeneous ensembles for each dataset using the KNN classifier, 7 deep feature extractors, dimensionality reduction techniques to reduce the size of features, and 2 combination rules. The key findings of each RQ in this investigation are:

(RQ1): What is the overall performance of the dimensionality reduction techniques in DR detection?

The SFM is the best performing dimensionality reduction technique since it was used alongside 18 out of 21 base learners, followed by the RFECV since it was used alongside 3 out of 21 base learners. Note that none of the remaining dimensionality reduction techniques was used with the base learners since they were never ranked in the first place.

(RQ2): Does the proposed ensembles perform better than their singles?

The results across the three datasets demonstrate that the proposed ensembles significantly outperformed their base learners.

(RQ3): Does increasing the number of base learners affect the classification performance of the proposed ensembles?

The results indicate that the classification performance is significantly influenced by the number of base learners utilized to build the ensembles. The ensembles created using 2 or 3 base learners were actually ranked last for the three datasets, in contrast to the ensembles created using 7 or 6 base learners. As a result, the classification performance is improved by increasing the number of base learners utilized to create the ensembles.

(RQ4): Out of the two combination rules, which one is the best performing?

The results show that the combination rule used to create the ensembles has an impact on the classification performance, since the ensembles created using the weighted voting are ranked first over the three datasets.

Ongoing works intend to develop new approaches for detecting DR by combining deep learning with different ensemble learning strategies.

REFERENCES

Bellemo, V., Lim, Z. W., Lim, G., Nguyen, Q. D., Xie, Y., Yip, M. Y. T., … Ting, D. S. W. (2019). Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health*, *1*(1), e35–e44. https://doi.org/10.1016/S2589-7500(19)30004-4

García-Lapresta, J. L., & Martínez-Panero, M. (2002). Borda count versus approval voting: A fuzzy approach. *Public Choice*, *112*(1), 167–184. https://doi.org/10.1023/A:1015609200117

Gurcan, O. F., Beyca, O. F., & Dogan, O. (2021). A comprehensive study of machine learning methods on diabetic retinopathy classification. *International Journal of Computational Intelligence Systems*, *14*(1), 1132–1141. https://doi.org/10.2991/IJCIS.D.210316.001

Han, J. H. (2022). Artificial Intelligence in Eye Disease: Recent Developments, Applications, and Surveys. Diagnostics, 12(8), 12–15. https://doi.org/10.3390/diagnostics12081927

Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. Computer Methods and Programs in Biomedicine, 177, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019

Islam, M. M., Yang, H. C., Poly, T. N., Jian, W. S., & (Jack) Li, Y. C. (2020). Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, *191*, 105320. https://doi.org/10.1016/j.cmpb.2020.105320

Jinfeng, G., Qummar, S., Junming, Z., Ruxian, Y., & Khan, F. G. (2020). Ensemble framework of deep CNNs for diabetic retinopathy detection. *Computational Intelligence and Neuroscience*, *2020*. https://doi.org/10.1155/2020/8864698

Lahmar, C., & Idri, A. (2021). On the value of deep learning for diagnosing diabetic retinopathy. Health and Technology 2021, 1–17. https://doi.org/10.1007/S12553-021-00606-X

Lahmar, C., & Idri, A. (2022). Classifying Diabetic Retinopathy using CNN and Machine Learning. *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING*, 52–62. ISBN 978-989-758-552-4; ISSN 2184-4305. https://doi.org/10.5220/0010851500003123

Lahmar, C., & Idri, A. (2023). Deep hybrid architectures for diabetic retinopathy classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging \& Visualization*, *11*(2), 166–184. https://doi.org/10.1080/21681163.2022.2060864

Litjens, G., Ciompi, F., Wolterink, J. M., de Vos, B. D., Leiner, T., Teuwen, J., & Išgum, I. (2019). State-of-the-Art Deep Learning in Cardiovascular Image Analysis. *JACC: Cardiovascular Imaging*, *12*(8P1), 1549–1565. https://doi.org/10.1016/j.jcmg.2019.06.009

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. *Computers in Biology and Medicine*, *112*(February), 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/WIDM.1249

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. *Diabetes Research and Clinical Practice*, *87*(1), 4–14. https://doi.org/10.1016/j.diabres.2009.10.007

Wong, T. Y., & Sabanayagam, C. (2020). Strategies to Tackle the Global Burden of Diabetic Retinopathy: From Epidemiology to Artificial Intelligence. *Ophthalmologica*, *243*(1), 9–20. https://doi.org/10.1159/000502387

Worsley, A. K. J. (2010). *A Non-Parametric Extension of a Cluster Analysis Method by Scott and Knott Published by : International Biometric Society Stable* URL : http://www.jstor.org/stable/2529369. 33(3), 532–535.

Zhang, W., Zhong, J., Yang, S., Gao, Z., Hu, J., Chen, Y., & Yi, Z. (2019). Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowledge-Based Systems*, *175*, 12–25. https://doi.org/10.1016/j.knosys.2019.03.016