Applied Machine Learning and Decision Combination for Identifying the Lazy Eye Vision Disorder

Patrick G. Clark¹, Christopher M. Gifford², Jonathan Van Eenwyk¹, Arvin Agah¹, and Gerhard W. Cibis³

¹Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045 USA
²National Security Technology Department, The Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723 USA
³Cibis Eye Care, Kansas City, MO, 64112 USA

Abstract—Amblyopia is a neurological vision disorder that studies show affects two to five percent of the population. Several early screening procedures are aimed at finding the condition while the patient is a child, including an automated vision screening system developed by Cibis, Wang, and Van Eenwyk. The system uses artificial intelligence software algorithms to achieve a 77% accuracy in identifying patients who are at risk for developing the amblyopic condition and should be referred to a specialist. Based on later work, two additional feature sets are also extracted from the same data captured with the AVVDA system, both achieving a 67% accuracy. This work explores the application of a multi-classifier collaborative learning architecture on the three data sets. The architecture has been shown to be more successful in problems that exhibit difficulty for single pass classifiers.

Keywords: Multi-Classifier, AI in Medicine, Applied AI, Lazy Eye, Amblyopia.

1. Introduction

Eye trouble of organic origin must be diagnosed and treated before the onset of an irreversible amblyopia or what is commonly referred to as lazy eye. This condition is a developmental disorder of the visual system caused by ocular abnormalities early in life. While surgery or optical correction of refractive errors can often address the initial cause of amblyopia, once amblyopia has developed, such interventions cannot restore visual function since amblyopia itself is a cortical deficit, a neurological disorder and not a physical one [1]. Amblyopia has two primary causes, namely, strabismus and anisometropia. Strabismus is a misalignment between the two eyes. Anisometropia is when the refractive error between the two eyes is different [2]. The reason that these two conditions can lead to amblyopia is because they cause the brain to begin ignoring the signals from the weaker or blurrier eye. Corrective action after amblyopia has developed becomes problematic, as the brain will not be able to regenerate the neural pathways. Thus, early detection is essential for the patient to have a healthy visual outcome.

Fortunately, amblyopia can be successfully treated if identified when the patient’s brain is still in the developmental stages, generally when the patient is fewer than six years old, with the non-controversial methods of glasses and patching therapy achieving 80 to 90% effectiveness [3].

The optimal solution for vision diagnosis would be a self-contained, low-cost, completely automated system that could accurately identify disorders with minimum operator training and low patient cooperation. The solution would need to enable widespread use with small children, including infants. One approach for this sort of solution is based on [4], [3], pioneering the science of analyzing images for identifying features that may indicate the development of amblyopia, and using artificial intelligence techniques to automate the process [5].

1.1 Problem Statement

In this work, we utilize a multi-classifier machine learning architecture consisting of a heterogeneous mixture of classification techniques in a team of intelligent agents. This architecture is applied to the problem of accurately identifying the amblyopic condition using three existing data and feature sets. The effort looks to make use of both ensemble and single classifiers on the data sets in order to ascertain the best features for identifying and referring patients to a specialist.

The previous work involved reviewing each individual data set on its own merits with a subset of classifiers, but here the work looks to take a larger view of the features by measuring them in relation to a number of classifiers. Results between the approaches are compared and conclusions drawn about the efficacy of the feature sets, applied to a variety of classifiers, as well as a multi-classifier decision fusion methodology in the vision diagnosis domain.

2. Background and Related Work

2.1 Vision Disorders

Amblyopia is the primary vision disorder this research attempts to accurately identify. It has two primary physical
causes, anisometropia and strabismus. Anisometropia is a condition where the refractive error in one eye is significantly different than the other. The difference in the refractive errors is difficult to overcome for a developing visual cortex, primarily due to the very different images being presented. Studies have shown that anisometropia is the predisposing condition that leads to amblyopia 50% of the time, and that an undiagnosed anisometropia will lead to strabismus [2]. Strabismus can be identified as a misalignment of the focal point between the two eyes, but the condition is not always identifiable with the naked eye and may require a thorough screening. Strabismus typically involves a lack of coordination between the two eyes and the extra-ocular muscles where the patient is unable to bring both eyes into focus on the same point in space, thus preventing proper binocular vision.

Both of these physical abnormalities in the eye have the potential to cause the development of a patient’s visual function to be impaired and cause the image from the amblyopic eye to be disregarded by the visual cortex. If the condition is allowed to persist, the neural pathways become permanently formed and the use of the amblyopic eye is diminished. The degree to which it is diminished varies based on how early the condition developed in the patient’s life and if any remediation treatment was used [2].

2.2 Automated Photo-Refractive Screening

Photo-refractive screening is based on a system to interpret the images of the eyes. It does not directly identify amblyopia, but it looks for defects in the eyes that may lead to amblyopia. One such system for identifying strabismus and amblyopia using video images, called Video Vision Development Assessment (VVDA), has been developed [3]. The method involves a consumer-grade video camera with a light source attached to the base of the camera. The patient sits approximately 52 inches from the camera and looks at the light source while approximately two minutes of video is recorded. The video is then digitized and analyzed by specialists or trained technicians to determine if the patient should be referred to a specialist. This processing is generally divided into frame selection and feature extraction.

More recently, research has been performed in order to automate the analysis of the frames using artificial intelligence techniques [6], [4], [5]. These works focused on implementing the image processing and Case-Based Reasoning [7] algorithms that constituted the first version of the Automated Video Vision Development Assessment (AVVDA) system. In a completely automated fashion, they were able to identify the key frames of the video, isolate the pupils, and locate the Hirschberg point [6], [4]. Figure 1 shows the image output from the AVVDA system, where the Hirschberg reflex and iris diameter are highlighted. The automated photo-refractive screening system works by having an operator take a short video of the patient, which is then analyzed automatically by the software in the following manner. Initially, the software identifies the frames where both eyes are open and looking at the light source. These frames are identified as key frames. Next, the software isolates the location of the eyes and pupils in the key frames. Finally, the software uses various techniques to extract the distinguishing features that may be indicators for amblyopia.

In order to further enhance AVVDA, researchers have investigated using the same feature set with a different set of classifiers, with the overall goal of the AVVDA system being to allow an unskilled technician to operate the system and accurately obtain a decision about patient referral to an optometrist or ophthalmologist [5]. In the current form, AVVDA uses Case-Based Reasoning [7], C4.5 Decision Tree [8], and Artificial Neural Network [9] classifiers to assist in making the decision. 54 features are extracted from the images to train the classifiers.

At this point, various artificial intelligence techniques have been utilized to automatically produce the referral to a specialist [3], [5]. The results reported in this work expands upon the automated photo-refractive screening method, with the goal to evaluate other classifiers, and the use of multi-classifier ensembles, as a means to produce more accurate results from the same feature set.
2.3 Combining Decisions from Multiple Classifiers

Classifier combination is an area in machine learning that has offered advances in classification accuracy for complex data sets. It has been termed differently in the literature, namely, classifier fusion, mixture of experts, committees, ensembles, teams, pools, collective recognition, composite systems, etc. When predictions from multiple classifiers are combined, they are said to form an ensemble that is then used to classify new instances. Several methods have been developed to combine classifiers, the most popular of which are voting, boosting, bagging, and stacking.

One of the primary questions in this area of study is whether combining classifiers is better than selecting the best classifier. Several works support that classifier combination provides an improvement in most cases, assuming that the classifiers exhibit reasonable individual accuracy. One such study utilized stacking with model trees to combine multiple heterogeneous learners [10]. Each learner utilized the full data set to produce a base-level model, the output of which is then combined with other base-level models using a meta-level classifier. Their results also indicated that the number of base-level classifiers did not significantly affect the results. Similarly, [11] found that when using voting and entropy methods as the heterogeneous classifier combination mechanism for word-sense disambiguation, increasing the number of less accurate classifiers adversely affected those with higher accuracy. Use of more classifiers, even if heterogeneous, does not always translate to better results. Depending on the combination method, the relative impact of adding classifiers can diminish as team size increases. Our work attempts to address this by implementing a framework where team configurations can be analyzed to determine what factors make a team of classifiers successful given the application and domain.

Although selection of the best individual classifier is easier and occasionally effective, combination techniques scale better to larger and more complex learning problems. Even combining all classifiers in an ensemble can be improved upon by selecting for combination only those that perform significantly better than others, termed Selective Fusion [12]. Using this technique, together with simple voting methods, enables the fine-tuning of diverse ensembles for specific data sets. It also offers performance comparable to other heterogeneous classifier combination methods such as stacking, without the additional computation and meta-learning costs. Researchers have studied the effectiveness of switching between selection (occurring in regions of the feature space where single classifiers are dominant) and fusion (occurring in every region not dominated by a single classifier) [13].

Other efforts have studied the use of training multiple classifiers on different feature subsets prior to their combination [14]. Focusing on differing and potentially overlapping feature subsets creates additional diversity, which could lead to an improved combined model. Some researchers also found that performance degradation occurs as the percentage of training batches (sets of training examples) overlap [15]. Multi-classifier systems offer a medium for additional study on how knowledge from classifiers with different, potentially overlapping feature and data subsets can be combined. Interaction between the classifiers during the training process may prove to increase learning efficiency, robustness, and accuracy.

3. Research Methodology

This section covers the research methodology followed to analyze and measure the utility of single- and multi-classifier configurations on a set of established features derived from 723 unique patients.

3.1 Features

We study three distinct feature sets in this work. Each originates from the same raw data, but represent three separate concepts on where the necessary data holding the predictive information is to be found. Previous experiments were performed on these data sets using a small set of individual classifiers. Therefore, the goal is to compare the results of various classifier configurations with the original published results.

3.2 The Multi-Classifier Collaborative Learning Architecture

The WEKA machine-learning suite is utilized as a base for the implementation of the proposed Multi-Classifier Collaborative Learning Architecture [16]. The primary uses of WEKA in this architecture are to prepare the data for experimentation and provide implementations of various classification algorithms. Additional Java and script/batch file implementations act as a wrapper for machine learning experiments involving single or multiple classifiers (i.e., for teams of any size), homogeneous or heterogeneous team composition, independent or collaborative learning (with the ability to vary the number of collaboration events during learning), and combining the decisions of the classifiers using a variety of accuracy- and vote-based combination techniques. This offers a robust and flexible architecture for machine learning studies involving multiple, collaborative learning agents.

The Multi-Classifier Collaborative Learning Architecture was utilized to study single- and multi-classifier machine learning configurations for the patient vision data set. By default, each classifier is provided a randomized version (different order of instances for each classifier) of the full training data set. Model creation takes place by training each classifier using the training data set. At this time, each classifier tests its model on the data on which it was trained. Once training is complete, the testing data set is passed
through each classifier’s model and the corresponding class probabilities (predictions) are recorded. The final individual testing predictions are then used for combining decisions from multiple classifiers. This acts as the final collaboration step, which fuses the knowledge from multiple classifiers to a single team classifier via accuracy- and vote-based mechanisms.

Decision combination utilizes each classifiers decision/label for each testing instance to arrive at a single team classification per combination method. A classifiers decision consists of a probability that the testing instance belongs to each of the possible classes. The highest probability represents the predicted class for each classifier. The Multi-Classifier Collaborative Learning Architecture calculates team classification accuracy for each of the 12 implemented vote-based combination methods. The combination method resulting in the best classification accuracy is selected, reflecting the overall performance of the team. For a detailed discussion of the 12 vote-based combination methods used for evaluation, consult [17].

While the original architecture offers a number of vote-based methods to be used for combining classifier results for evaluation [17], this research endeavor utilized only the Average method. The Average method averages all predictions for each class from all the classifiers, selecting the class with the highest average value. The class selected by the classifier with the highest accuracy on the testing data breaks a tie.

4. Experiments for Three Feature Sets

4.1 Experimental Setup

The experimental studies followed the process of a series of rounds based on classifier performance. As long as the results of some classifier in the round exhibited reasonable improvement over the previous work, then team configurations exhibiting the highest testing accuracy move on to the next experiment, where larger teams are constructed from the best teams from the previous experiment.

In all experiments the classifiers are provided with the full training portion for model creation, and no collaboration events occurred. The reasoning behind this decision is to keep the experiments focused on measuring machine learning algorithm and decision combination performance without reviewing the effect of collaboration and independent learning. Further work could be done to measure the effect of those additional configuration options of the architecture.

The first step for a specific data set is to coarsely test a variety of individual machine learning algorithms, each with a few settings and initialization seeds, to determine which algorithms perform well in general for the data set. The best settings and corresponding 10-fold cross-validation testing accuracy for each algorithm are recorded, and the top four algorithms are selected to advance to the next experiment.

The next step is to evaluate a team four classifiers in four different configurations. This inherently includes homogeneous and heterogeneous team compositions. The same process takes place for teams of size eight, where the top teams of size four are selected to advance to the experiment of size eight teams. The same process is used with the results of the size eight teams to construct the final sixteen-team experiment. Once the experiment of size sixteen teams is complete, all results are compiled and analyzed.

It is important to note that because this work is primarily a comparison of feature sets, an exhaustive multi-classifier experiment set is not performed. Rather, the methodology looks at a shallow set of initial experiments to determine if further classifier fusion is worth pursuing based on an improvement in overall accuracy. This phased experimentation process results in running fewer round one experiments for Experimental Data Set 1, and more round one experiments Experimental Data Set 2 and Experimental Data Set 3. This is because positive results are discovered early in the Experimental Data Set 1.

In addition to differences in round one, the methodology further results in all three feature sets not following to the final sixteen-team experiment. Since this work investigates the efficacy of the three feature sets for identifying the amblyopic condition, when the results in the early rounds do not advance the state of the previous work, all the rounds are ended and the results are presented.

4.2 Experimental Results

This section presents the experimental results from a variety of machine learning configurations for the three patient vision data sets. The notation used in the subsequent figures and analyses are abbreviations of machine learning classification algorithms. The abbreviations for the ten utilized classification algorithms are: Decision Table (DTB) [18], Decision Tree (DT) [8], Instance-Based KNN (IBK) [19], Logistic Regression (LGR) [20], Naïve Bayes (NB) [21], PART Rule Learner (PRT) [22], Random Forest (RFT) [23], Radial Basis Function Network (RBF) [18], RIPPER Rule Learner (JRP) [24], and Support Vector Machines (SVM) [25]. The reader is referred to the referenced papers for the algorithms and details on their underlying structure and theory.

When listing team compositions, the number of classifiers for each algorithm of the team is listed, separated by +. For example, 3IBK+2PRT+2DT+1LGR represents a team of size eight composed of three IBK, two PRT, two DT, and one LGR classifiers.

A total of 39 machine learning experiments were performed, including the variation of team size and composition. Experiments are broken down as follows:

- 29 size one
- 4 size four
- 3 size 8
4.3 Experimental Data Set 1: Color Density and Hirschberg Reflex

The first experimental data set finds its genesis in previous work from Wang and the original AVVDA system. A total of 54 key features of patients’ eyes (27 features per eye) were extracted from the video feeds and key frames. The two primary features are the pupil radius and the degree of fixation based on the Hirschberg point. The remaining 25 features for each eye are calculated using the color of pixels within 80% of the radius of the pupil. This includes, for example, the average red, green, and blue values throughout the pupil. The reader is referred to [4] for additional details on how these values are calculated. These 54 features for each of the 723 patients and all of the identified key frames were used as input into the multi-classifier architecture.

As described in the previous section, many classifiers where tested on the data set, both single and combined. As a result of the experimental process, the best outcome was distilled to 16 Random Forest classifiers. We can investigate the initial outcome of the four rounds of experiments by reviewing table 1. The results show increasing improvement as the rounds advance (i.e., team size increases), and ultimately show the best results were achieved with a homogeneous, 16-classifier configuration.

In contrast, the results of the AVVDA system represent the original research and cover single instance Case-Based [7], Artificial Neural Network [9], and Decision Tree [8] classifiers. Table 2 is a summarization of the results for comparison, and shows that the use of a multi-classifier decision combination approach improved accuracy and specificity at the cost of sensitivity.

4.4 Experimental Data Set 2: Iris and Pupil Color Slope with Middle Stack Key Frame Selection

The second experimental data set is based on the work of [26]. It utilizes the same 723 patient videos but instead extracts a different set of frame and feature data. The difference begins with the key frame concept. In the previous experiment, a patient will have typically produced many key frames. As described in the Related Work section, the software identifies a key frame based on two primary criteria. First, the patient must have both eyes open. Second, the patient must be looking at the light source so that a reflection off of the retina is clear. From the key frame stack identified for a patient, a single frame is then selected. In the second experiment, the research takes a very simple approach to identify the single frame that will be used for feature extraction. It selects the frame from the middle of the stack of frames. If there are an odd number of frames, it selects the frame that is a whole number division of the entire frame count by two. The goal is to avoid the fringe frames (the first or last), and analysis of a subset of the data shows that the middle frame has the greatest chance of being one of the best images [26].

Now that a frame is selected, the color information is extracted from the image starting at the edges of the iris for both the left and right eyes. The red, blue, and green information is extracted from the eye image in a left to right pattern, representing the rate of change of color across they eye, or color slope. This is illustrated in figure 2. The key frame for each patient is then further processed to produce 552 color features.

Experimental results using the iris and pupil color slope data set are summarized in table 3. The results do not show...
an improvement over the previous work when the data was applied to a number of classifiers in the Multi-Classifier Collaborative Learning Architecture. The best single classifier from the architecture was the Support Vector Machine with improved specificity at the cost of both sensitivity and accuracy.

In contrast, the results of the work presented by [26] represent the original research and cover single team Decision Tree, Neural Network, and Random Forest classifiers. Table 4 exhibits the summarized observed results using the iris and pupil color slope data set.

Notably, when reviewing the preliminary single-classifier results from the Multi-Classifier Collaborative Learning Architecture on data set two, the results did not warrant further classifier ensembles. The reasoning behind this decision is that the results were not coming reasonably close to the accuracies achieved with data set one. Since this work is to investigate the best features for identifying the amblyopic condition, the results did not advance data set two and did not improve on previous work.

4.5 Experimental Data Set 3: Pupil Color Slope with Level Key Frame Selection

The third experimental data set is also based on the work of [26]. Utilizing the same initial set of 723 patient videos, the goal is to find the best patient and frame sets for selecting a frame. The set of key frames is still identified as described in the Related Work section. However, the frame selection process was altered so that a method different from selecting the middle frame is utilized. In this case, the frame selected out of all the key frames is that where the patient’s head is most level. Since the entire classification problem

Table 3: Results from the experimental study performed using the Multi-Classifier Collaborative Learning Architecture on the initial color slope data set.

<table>
<thead>
<tr>
<th>Team</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Size: 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTB</td>
<td>59.34</td>
<td>72.60</td>
<td>38.95</td>
</tr>
<tr>
<td>DT (Pruned)</td>
<td>56.29</td>
<td>64.16</td>
<td>44.21</td>
</tr>
<tr>
<td>DT (Unpruned)</td>
<td>56.43</td>
<td>64.38</td>
<td>44.21</td>
</tr>
<tr>
<td>IBK</td>
<td>63.62</td>
<td>75.57</td>
<td>45.26</td>
</tr>
<tr>
<td>JRP</td>
<td>58.65</td>
<td>73.97</td>
<td>35.09</td>
</tr>
<tr>
<td>LGR</td>
<td>56.43</td>
<td>60.73</td>
<td>49.82</td>
</tr>
<tr>
<td>NB</td>
<td>58.92</td>
<td>64.61</td>
<td>50.18</td>
</tr>
<tr>
<td>PRT</td>
<td>56.57</td>
<td>66.90</td>
<td>40.70</td>
</tr>
<tr>
<td>RFT</td>
<td>62.38</td>
<td>69.86</td>
<td>50.88</td>
</tr>
<tr>
<td>RBF</td>
<td>60.86</td>
<td>83.11</td>
<td>26.67</td>
</tr>
<tr>
<td>SVM</td>
<td>64.73</td>
<td>73.97</td>
<td>50.53</td>
</tr>
</tbody>
</table>

Table 4: Comparison of results achieved by the Multi-Classifier Collaborative Learning Architecture on the initial color slope data set.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Classifier Collaborative Learning Architecture</td>
<td>64.73</td>
<td>73.97</td>
<td>50.53</td>
</tr>
<tr>
<td>AVVDA Pupil and Iris Color Slope</td>
<td>67.91</td>
<td>82.42</td>
<td>45.61</td>
</tr>
</tbody>
</table>

is centered on the reflection of light off the lens and the retina, keeping the angle of reflection the same for both eyes would, theoretically, further enhance any differences between a healthy eye and the eye of a patient who should be referred to a specialist. The important features extracted around the crescent reflection and the Hirschberg point that is measured from that crescent further support this concept. The frame that is most level is determined by the pixel location of the center of the pupil at the Y-axis. The value is compared between the two eyes for each key frame, and the frame corresponding to the closest value is selected. When no key frame could be found where the difference in the height of the eyes was less than five pixels, then the patient data was discarded. The data was discarded so that the classifiers would not be trained with imperfect data and in hopes of further isolating the features that will yield the best results. This process reduced the number of patients in the entire sample from 723 to 499. The key frame is then further processed to extract color data from the pupil only. This process produced 180 features for each patient. Figure 3 illustrates the feature vectors extracted.

Experimental results using the pupil color slope data set are summarized in table 5. The results exhibit similar results
as the experimental data set two, an improved specificity at the cost both accuracy and specificity. The best overall result from the Multi-Classifier Collaborative Learning Architecture was from a support vector machine.

In contrast, the results of the work presented by [26] represent the original research and cover single team Decision Tree, and Random Forest classifiers. Table 6 exhibits a summary of the observed results using the pupil color slope data set.

Notably, data set three also does not exhibit reasonable improvement over the data set one to warrant further classifier combination. The reasoning behind this decision stands the same as the experiment two results, lack of reasonable improvement over data set one, and previous work.

Table 5: Results from the experimental study performed using the Multi-Classifier Collaborative Learning Architecture on the pupil only color slope data set.

<table>
<thead>
<tr>
<th>Team</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Size: 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTB</td>
<td>62.93</td>
<td>77.45</td>
<td>38.90</td>
</tr>
<tr>
<td>DT (Pruned)</td>
<td>58.52</td>
<td>67.32</td>
<td>44.04</td>
</tr>
<tr>
<td>DT (Unpruned)</td>
<td>58.32</td>
<td>67.32</td>
<td>44.04</td>
</tr>
<tr>
<td>IBK</td>
<td>62.32</td>
<td>80.07</td>
<td>34.20</td>
</tr>
<tr>
<td>JRP</td>
<td>62.12</td>
<td>78.76</td>
<td>35.75</td>
</tr>
<tr>
<td>LGR</td>
<td>58.92</td>
<td>67.32</td>
<td>45.60</td>
</tr>
<tr>
<td>NB</td>
<td>58.92</td>
<td>59.80</td>
<td>57.51</td>
</tr>
<tr>
<td>PRT</td>
<td>57.11</td>
<td>68.63</td>
<td>38.86</td>
</tr>
<tr>
<td>RFT</td>
<td>62.53</td>
<td>68.30</td>
<td>53.37</td>
</tr>
<tr>
<td>RBF</td>
<td>58.92</td>
<td>87.58</td>
<td>13.47</td>
</tr>
<tr>
<td>SVM</td>
<td>63.93</td>
<td>81.70</td>
<td>35.75</td>
</tr>
</tbody>
</table>

Table 6: Comparison of results achieved by the Multi-Classifier Collaborative Learning Architecture on the pupil color slope data set.

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Classifer Collaborative Learning Architecture</td>
<td>63.93</td>
<td>81.70</td>
<td>35.75</td>
</tr>
<tr>
<td>AVVDA Pupil Color Slope</td>
<td>67.94</td>
<td>37.31</td>
<td>87.25</td>
</tr>
</tbody>
</table>

5. Conclusion

The primary contribution of this work is to take an existing problem, analyze three feature sets produced and find the most predictive set for accurately classifying the markers for the amblyopic condition. The feature sets were processed through a Multi-Classifier Collaborative Learning Architecture system, producing results for many classifiers, some single and some ensemble. As observed in the Experimental Results section, the original AVVDA features were improved upon as the experimental rounds progressed, thus showing promise for the ensemble classification method with certain data sets.

In contrast, the color slope experimental data sets performed worse using the Multi-Classifier Collaborative Learning Architecture. This suggests that the features are not only difficult to use for classification, but do not hold much predictive ability when presented with new patient instances. Therefore, the results further the understanding of the predictive ability for the given data sets. Notably, of equal contribution to the vision assessment work is to illustrate that there may have been a limit reached in the classification ability of the selected color density, Hirschberg reflex, and color slope features. This conclusion is further supported by the success seen in previous work with the Multi-Classifier Collaborative Learning Architecture against challenging classification problems, and its ability to improve results over previous work on the difficult data sets [17], [16].

References


