Feature Selection Pipeline based on Hybrid Optimization Approach with Aggregated Medical Data

Palwinder Kaur[1]
Department of Computer Science
IKG Punjab Technical University, Jalandhar, Punjab, India

Rajesh Kumar Singh[2]
Department of Computer Applications
SUS Institute, Tangori, Punjab, India

***Abstract*—For quite some time, the usage of many sources of data (data fusion) and the aggregation of that data have been underappreciated. For the purposes of this study, trials using several medical datasets were conducted, with the results serving as a single aggregated source for identifying eye illnesses. It is proposed in this paper that a diagnostic system that can detect diabetic retinopathy, glaucoma, and cataract can be built as an alternative to current methods. The data fusion and data aggregation techniques used to create this multi-model system made it conceivable. As the name implies, it is a way of compiling data from a large number of legitimate sources. The development of a pipeline of algorithms was accomplished through iterative trials and hyper parameter tweaking. CLAHE (Contrast Level Adaptive Histogram Equalization) approaches, which increase the gradient between picture edges, improve segmentation by raising the contrast between picture edges. The Gabor filter has been shown to be the most effective method of selecting features. The Gabor filter was selected using a hybrid optimization method (LION + Cuckoo), which was developed by the author. For automation, the Support Vector Machine (SVM) radial is the most effective method since it delivers excellent stability and accuracy in terms of accuracy and recall, as well as precision and recall. The discoveries and approaches detailed here provide a more solid foundation for future image-based diagnostics researchers to build on in the future. Eventually, the findings of this study will help to improve healthcare workflows and practices.**

Keywords—Content-based image retrieval system; CLAHE; Gabor filter; Cuckoo search; LION optimization; support vector machine

I. Introduction

The covid-19 pandemic has forced rethinking about the effectiveness of clinical workflows and practices [1]. The current advances in image processing and machine learning algorithms make it imperative that new algorithms and methods be incorporated into health care technologies and systems[2]. Current research trends and evidence point out that creating multi-model systems requires using multiple protocols, stacks of technologies, and an array of algorithms and multiple data sources[3]. Choosing a specific technology stack and pool of algorithms for building a reliable system has become tedious work and confusing. The primary reason is the availability and choice of ready-to-use frameworks, APIs, libraries, and technological stacks. Hence, finding and appropriating existing algorithms requires exhaustive experimentation and optimization e.g., at the data processing level: importing, validating, cleaning, converting, normalizing, and pre-processing the data requires a lot of experience and intuitiveness for selecting the suitable method, which would yield the best possible outcomes [4]. Application of methods such as data fusion , aggregation and argumentation also need to be explored, especially when the number of data instances of particular class are less and there is an imbalance in the dataset.

The term "pipeline" in computer science has broader connotations in the current context [5]. It is referred to as multi instructions performed as a unit. The computer unit may be some software module or hardware such as Graphics Cards. HTTP pipeline is a sequence of steps taken to handle HTTP traffic and tasks in the context of the research work. "Algorithm Pipelining" is more appropriate as this research work involves constructing algorithms, frameworks, and systems sequences that can perform tasks such as prediction of corona virus with the help of experiments with high reliability [6]. A pipeline is another method of defining an experiment [7]. An experiment whose objectives are known and defined and the outcome helps construct a fully functional system. In this research work, an attempt will be made to identify an accurate workflow of the methods (image processing and machine learning) that would yield a high-performing system that can detect at least three types of eye diseases, i.e.; diabetic retinopathy, glaucoma, and cataract. Hence, in the next section, the workflows, approaches, and methods are discussed which are used these days to construct multiple-model eye disease detection systems.

II. Review

Medical imaging technology has significantly progressed, which has helped to reduce the burden of detection of numerous diseases. Machine learning algorithms and their comprehensive frameworks have greatly helped the image segmentation field [8][9]. However, the biggest problem that the researcher faces in this context is building diagnostic systems related to the characteristics of the data set [10]. By analyses of the publically available medical image data set e.g. eye diseases, it can be observed that many of these data sets belong to particular or specific modalities, and at the same time, they are poorly annotated [11][12]. Many data sets are not labelled as per the stage of the disease; in many cases, the dataset is not as per the requirement of machine learning modelling [13]. Due to this challenge, multi-model disease detection systems are hard to realize [14]. In simple words, it means that highly specialised systems of detection can be constructed. However, detection system for a specific domain is hard to construct; for example, the current literature quotes many examples of handling diabetic retinopathy detection systems[15], but few new systems are illustrated in high impact journals that deal with the detection of multiple diseases such as glaucoma, cataract, and diabetic retinopathy at the same time [16][17][18]. This industry faces two kinds of problems: the first is limited annotated data sets, and the second one is weak or incomplete annotations in the data set as per medical grade system. Contemporary literature cites different solutions to overcome the problem depending upon the problem.

The most frequently used technique is data aggregation [19], data augmentation [20][21] and data fusion [22]. In data augmentation, the existing data set size is increased by adding more synthetic data to it, or learning from the existing annotated data is done for constructing a more significant size data set; this way, multiple diseases and modalities can be covered. Such methods also help overcome the imbalance in the data set and help leverage active learning. The most significant advantage is that multi-disease detection systems can now be constructed using multiple data sets or data fusion techniques. Data fusion algorithms leverage multiple disease data sets for constructing image processing functions that work on heterogeneous disjoint sets that can support multi-disease detection systems [23]. Some authors refer to such procedures as data adaptation also. Data adaptation is a process by which a data set is constructed, which helps the learning component of the detection system to discriminate between the various disease modalities and come out with an effective solution. Data augmentation, data fusion, and data adaptation help immensely overcome the challenge of building generic systems of detection [24]. However, the current literature also points out that there are limitations in using single algorithms for building multi-disease detection systems. Research in this context also shows that building classification systems relying on specific features and single classification methods may not yield a stable numerical system. There is always a need to use various methods and solutions for detecting multiple diseases. Hence, many researchers have concluded that the usage of hybrid techniques and combination approaches is far better than training a specific machine learning model. This way, a robust model for constructing multiple disease detection systems can be realised and implemented.

From the current literature it is amply clear that as a strategy for building medical detection systems, three possible path ways can be used for constructing systems of disease detection [25]. The first uses purely statistical methods, the second uses optimization methods, and the third uses machine learning algorithms or deep learning models. It should be however be noted that Image segmentation is a precursor for using these three approaches because extracting the object of interest from the medical images is a fundamental step in building disease systems. An important gap that is generally visible in the current literature is that few scholars are building systems can multiple diseases detection in medical domain. Generally, the focus of research paper is to work with a single medical modality with specific dataset. However, the need to hour is to construct models that can automatically detect multiple ailments in a comprehensive way. It became critical in context of detecting eye problems due to covid-19 pandemic norms.

In short, it can also be observed from the current research works in context of most relevant approaches are statistics, machine and deep models. Statistical methods such as descriptive statistics, correlation, f-test, t-test, etc., are generally used to understand the nature of the data and identify the suitability of the data for machine learning model [26]. The optimisation algorithms [27][28] such as Genetic Algorithm, ant-colony-optimisation [29], differential-evolution,cuckoo-search [30], particle-swarm-optimisation, firefly, metaheuristic swarm-optimisation, Harris-hawks-optimisation, bat-algorithm, lion-optimiser, grey-wolf-optimiser, moth-flame-optimisation, flower-pollination-algorithm whale-optimisation-algorithm, etc. are used for constructing feature engineering hypothesis for attaining the best possible solutions [31][32][33]. The automation of diagnose comes with implementing machine learning and deep learning algorithms. Current literature gives ample evidence that authors are primarily citing hybrid methods for producing high-precision systems of medical disease detection [34][35]. Large amounts of citations can be found that are showing the most frequently used machine learning algorithms for detecting eye problems include K-Means, K-nearest neighbour (KNN), support vector machine (SVM), ANN or neural networks , decision trees, logistic regression.

This research attempts to analyse three pipelines that would yield a numerically stable multi-disease detection system. The selection of methods used in each type of pipeline is based on the previous research works done by contemporary technical people. Secondly, it is a sincere attempt to find a novel pipeline of methods that can offer consistently repeatable performance detecting multiple eye diseases.

III. Materials and Methods

In this section, the steps that make up the workflow of this research are given. It explains techniques, procedures, and algorithms used for building a system designed for eye diagnostics. The research flow block diagram Fig. 1 may be referred to for better understanding. The dataset used in this research work is publically available (https://github.com/palavibhangu/retina_dataset.)The dataset has 300 images of each of three types (diabetic retinopathy, cataract, glaucoma) of eye diseases and healthy eye images.

The aggregated size of the dataset of 1200 images was realized with the help of operation referred as data fusion. As mentioned earlier, in data fusion, multiple datasets are stacked and organised to act as single source of dataset. This has been done to overcome the challenge of low availability of particular class of instances of medical data.

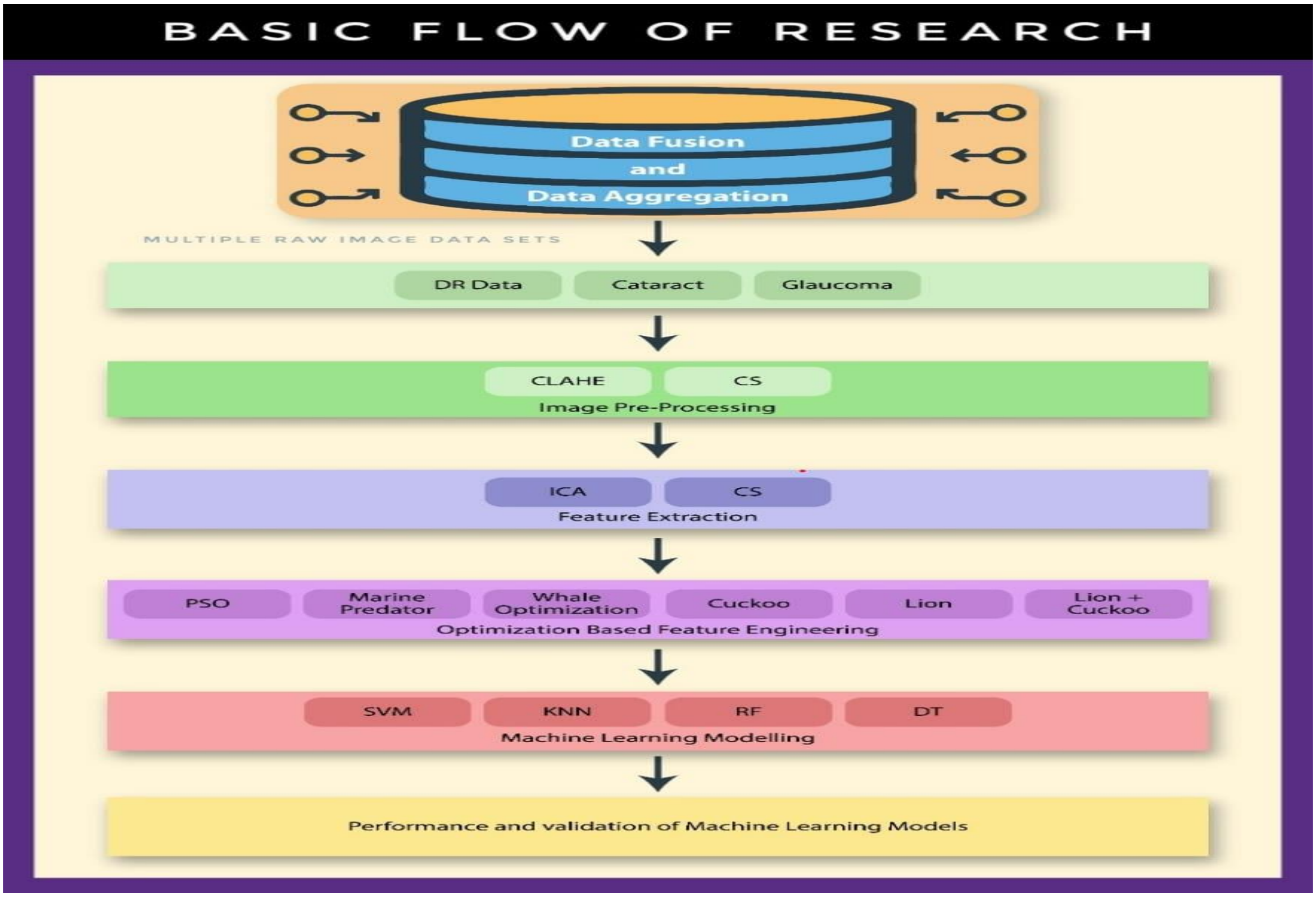


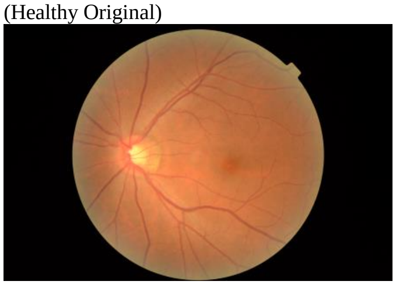
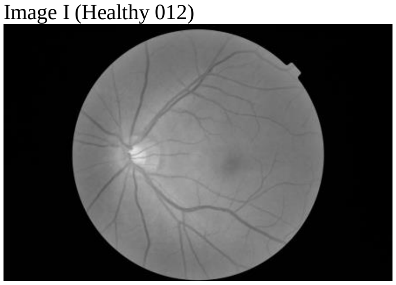
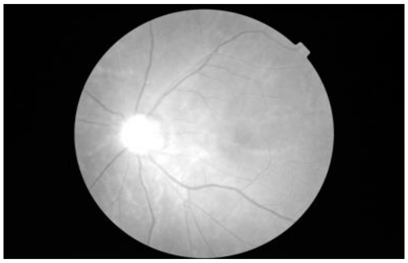
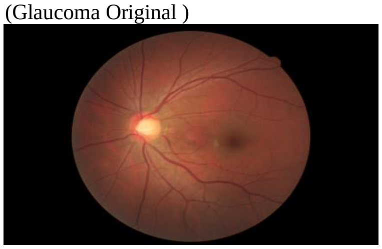
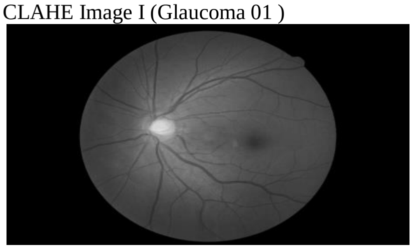
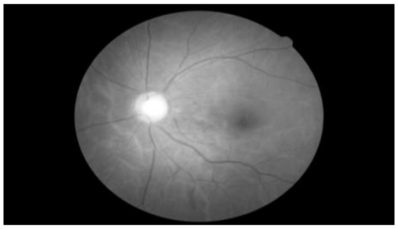
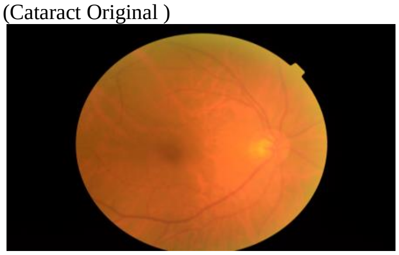
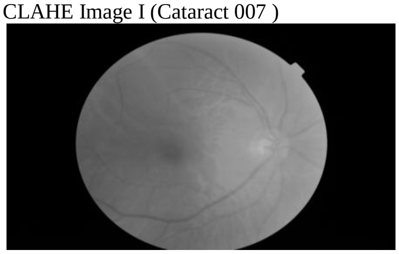
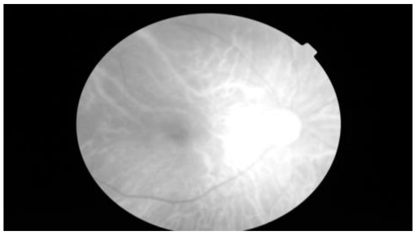
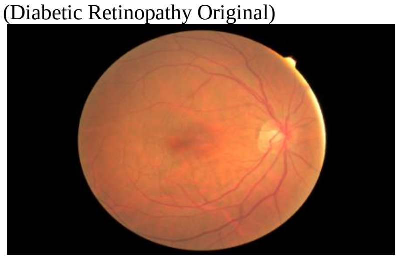
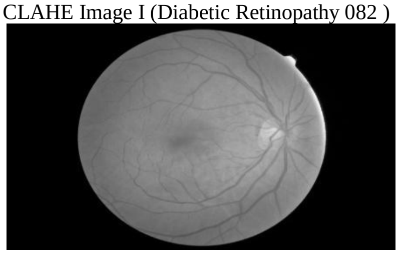
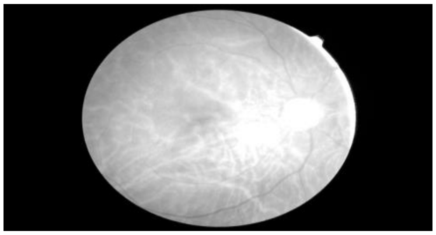
Fig. 1. Basic Flow of Research.

A. Image Processing

In this section, demonstration of the pre-processing steps that include selection of appropriate contrast methods and performance assessment of this step is discussed. It is apparent that for building a generic system of eye problem detection, the images were subjected to pre-processing operations such as contrast enhancement. The purpose of the contrast enhancement is to increase the differential between the various segments of the images. Increasing the differential between the object' pixels that have higher values will attain higher intensity levels and the similarity of the lower-intensity pixels will acquire lower levels of intensity. This process is quite helpful when the segmentation process has to be done as an essential step. Therefore, technically it refers to any technique that uses a function to exaggerate the apparent difference between adjacent structures created during image processing. This helps characterize objects of interest, i.e., a characteristic that can hint at eye problems. Multiple algorithms are available in the image processing domain to improve the images' medical quality. These include histogram-based methods such as adaptive histogram equalisation and contrast stretching (CS) methods such as min-max stretching. In the context of the problem undertaken, after a lot of experimentation and quality grading, it was found that the CLAHE method[36] is most suitable for the said purposes. The CLAHE method has step operations with which it increases the contrast. The first step divides the image into tiny regions and creates a local histogram for each region. A map of the local histogram is constructed. After this, a clipping point of the histogram is identified for each region. As the contrast process iterates, every region's noise is also reduced with the help of the subtraction method. The result is redistribution of the intensity values of the image. Table I gives the output of the CLAHE.

It can be observed from the comparison that the CLAHE algorithm produces better levels of differentiation between the various entities embodied in the fundus eye images. The selection of the CLAHE algorithm is based on the assessment of output given in the next section. It should also be noted that classical histogram equalization method was initially evaluated and it was found that dynamic or adaptive methods always perform better.

TABLE I. CONTRAST OUTPUT TABLE

Original Image(s)	CLAHE	1) *Min-Max Contrast Stretching*
(Healthy Original) 	Image I (Healthy 012) 	
(Glaucoma Original) 	CLAHE Image I (Glaucoma 01) 	
(Cataract Original) 	CLAHE Image I (Cataract 007) 	
(Diabetic Retinopathy Original) 	CLAHE Image I (Diabetic Retinopathy 082) 	

It is difficult to objectively evaluate the quality of contrast that an algorithm may provide in photographs from a technical standpoint. The use of a subjective judgment of the image is more appropriate in these situations. The advantage is that domain experts will make decisions in accordance with the medical grade standard of care. Thus, two judges (Judge 1 & Judge 2) were assigned the task of evaluating four factors related to the quality of the images changed by contrast algorithms: distortion in the image, artefacts, noise and information gain that can valued after the contrast enhancement operation. A questionnaire was developed, and judges assigned scores between 1 and 3 on a scale of 1 to 3. The number 3 indicates that there is no introduction of noise, distortion, or artefacts as a result of contrast. The number 1 represents the presence of 100 percent noise, distortion, or artefacts in the freshly produced images. If a number 2 is assigned, the value signifies a 50 percent chance of noise, distortion, and artefacts occurring. The same is true for the factor information gain: one indicates that there is no information gain when the contrast algorithm is performed, and 3 indicates that there is a 100 percent gain in the information, indicating that the contrast transformation will be beneficial in better segmentation. There were two experts participated in the evaluation process, and the inter-rater agreement (using average score) between them was computed, since, there four medical modalities the results are shown in Tables II, III, IV and V, respectively.

TABLE II. HEALTHY IMAGES (RANDOM SAMPLE = 25)

Healthy Samples CLAHE vs CS				
	Factors	Judge 1Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.68	2.6	2.64
	Distortion	2.68	2.6	2.64
	Artefacts	2.60	2.68	2.64
	Information Gain	2.84	3	2.92
CS	Noise	2	2	2
	Distortion	2	2	2
	Artefacts	2	2	2
	Information Gain	2	2	2

TABLE III. GLAUCOMA IMAGES

Glaucoma Samples CLAHE vs CS				
	Factors	Judge 1Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.64	2.62	2.63
	Distortion	2.68	2.64	2.66
	Artefacts	2.92	2.92	2.92
	Information Gain	2.92	2.92	2.92
CS	Noise	2.1	2.1	2.1
	Distortion	2	2	2
	Artefacts	2	2.3	2.1
	Information Gain	2.4	2.4	2.4

TABLE IV. CATARACT IMAGES

Cataract Samples CLAHE vs CS				
	Factors	Judge 1Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.68	2.64	2.64
	Distortion	2.92	2.68	2.8
	Artefacts	2.8	2.8	2.8
	Information Gain	2.76	2.8	2.78
CS	Noise	2.2	2.2	2.2
	Distortion	2	2	2
	Artefacts	2.2	2	2.1
	Information Gain	2	2.2	2.1

TABLE V. DIABETIC RETINOPATHY IMAGES

Diabetic Retinopathy Samples CLAHE vs CS				
	Factors	Judge 1Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.8	2.68	2.78
	Distortion	2.8	2.8	2.8
	Artefacts	2.76	2.68	2.72
	Information Gain	2.88	2.88	2.88
CS	Noise	2	2	2
	Distortion	2.2	2	2.1
	Artefacts	2	2.4	2.2
	Information Gain	2.1	2.1	2.1

Observations from Table II to V demonstrate that CLAHE method is more effective than contrast stretching. In all healthy images, Glaucoma, Diabetic Retinopathy and Cataract**,** the evaluation shows the CLAHE method is the most stable and reliable algorithm for the said purpose. This may be attributed to the fact; the correct parameters were selected for taking maximum advantage of the CLAHE algorithm. The parameters; Windows size = 8, Clip limit =0.4, Bin size 255) of CLAHE and use of Rayleigh (alpha value =0.35) based distribution for construction of the histogram yield a better output. Min-Max Contrast Stretching is intensity normalization; this is a typical well established pre-processing step taken by many researchers; nevertheless, in the current context, it is performing not well as compared to the CLAHE method. In the case of Min-Max Contrast stretching intensity increases but loss of information/pixels is also happening. It is now time for extracting features from these quality enhanced image dataset. The coming section discuss the process of extracting and selecting appropriate optimization algorithm that produces highest possible accuracy of the detection system that is based on machine learning model.

B. Optimization based Feature Engineering

It is possible to take full advantage of machine learning when one looks for recognisable patterns in large quantities of data. With conventional statistics, data consolidation and reduction is the key, and the quality of diversity of the data is given a lower mark. However, machine learning depends on extensive data and high levels of detail (think variety) (think columns or attributes). *Feature engineering* is used to get manual and automated analyses to speed up by adding more features/attributes and providing more details on existing data[37]. Feature analysis can help developers exploit and investigate data with more profound patterns. They are helpful for many machine learning procedures and vital to spot trends that can give real-time hints on diseases in our context. There are two main ways to expand features: ingesting more data after pre-defined features are created or training data to increase available features. The feature selection process includes selecting combinations of variables with large discriminative values to support the detection of various types of classes in the dataset.

As indicated in recent publically available literature assessments, critical variables for diagnosing eye abnormalities include an examination of the eyes' colour, texture, and form. The Gabor Filter was used to analyse the texture [38][39] and it was compared to a independent component analysis (ICA) method [40] for determining the most acceptable features in the image dataset. ICA assists in the discovery of a reduced projection picture or sub space of the original image with decreased dimensions. This reduces overhead while extracting the best feasible statistically independent information from each image. Additionally, ICA method encompasses a wide range of kurtosis and skewness values. The fixed objective function is determined by the differential equation (1).

f'(x) = (x-a) f(x) / (b0 + b1 x + b2 x2) (1)

Where a, b0, b1, and b2 are distribution parameters. When the source distributions are known (as they are in this scenario), the score functions are the ideal choice for the objective function. The Pearson ICA system's scoring function is defined as (x) = - f '(x) / f (x) = (x-a) / (b0 + b1 x + b2 x2). The parameters a, b0, b1, and b2 are estimated using the moments approach.

The Gabor filter helps to extract features of images by computing features at different frequencies and by changing the theta angle. This way features from multiple orientations and directions are extracted. Mathematically, it is computed using equation (2).

$$g(x, y, \lambda, \theta, \varphi, \sigma, \gamma) = \exp\left(\frac{-x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) * \cos(2\pi \frac{x'}{\lambda} + \varphi) \quad (2)$$

Where

$$x' = x\cos\theta + ysin\theta\ , y' = -x\sin\theta + y\cos\theta,$$

After applying the Gabor filter and extraction of Gabor vector and ICA components, the dataset was transformed into a feature matrix of 1000x100 with the help of the reshape function of Matlab. This was done so that uniform sets of features are used from each image for machine learning. It is expected that as a result of application of these feature selection methods, the chosen features will have a smaller classification error and a higher degree of generalisation when machine models will be constructed.

The analysis for finding the best optimisation algorithm for feature selection depends on four performance factors. The first one is the coverage percentage: it constantly desired that it should be 100% so that no feasible area during the optimisation process is left uncovered to find the optimal solution to the problem. For example, the PSO is not a global optimisation algorithm [41]; hence, it cannot guarantee convergence to a local optimum. Due to this fact, the stability of the solution may be questionable. The excellent level of coverage is reflected in the accuracy parameter. The second performance analysis is about the computational time the optimisation solution takes to reach the most feasible solution. It is better to have feasible solutions fast. Then, the number of features that the algorithm finds useful is critical. It is generally expected that the lower the number of features, the lower is the overhead for the machine learning algorithms for building an appropriate solution.

All these optimization algorithms were executed using standard parameters values such as the number of iterations = 10 and solutions=100. Each algorithm has specific parameters that need to be configured before these algorithms can be executed. It can be observed from Table VI that cuckoo and LION algorithms [42] are most competent in terms of accuracy and number of selected features. However, it is better to use a hybrid approach and combination of the LION and cuckoo as it further reduces the overhead and keeps the accuracy levels a bit higher than individually using the LION or Cuckoo algorithm. The optimization algorithms' performance analysis shows that applying a hybrid algorithm helped obtain the best possible solution in terms of the number of features. The coverage of the hybrid algorithm is excellent, which lead to the selection of a feature matrix that has the lowest number of features (18).

The PSO, Whale Optimization and Marine Predator algorithm give a good level of accuracy (above 90%), but the number of features are more than the Hybrid approach. In the next section, however, an examination of the machine learning models will be done to ascertain the performance of classifiers using these selected features.

TABLE VI. FEATURE SELECTION USING OPTIMIZATION

S.No	Optimization Algorithm	Accuracy	Number of selected features
1	Cuckoo's	92.5	40
2	LION	87. 14	28
3	Particle Swarm Optimization	95.7	64
4	Marine Predators	92.1	22
5	Whale Optimization Algorithm	90.9	20
6	**Cuckoo+LION (Hybrid)**	**98.9.0**	**18**

IV. RESULT AND DISCUSSION

The accuracy of machine learning entirely depends on the quality of data it processes for learning patterns of data. This section explains the procedure followed for finally automating detecting four medical conditions of the eyes. For automation, four classifiers (KNN, SVM (Radial), DT, RF) were chosen based on the previous work done by other researchers and organisations.

The feature matrix of eighteen numerical features was selected using a hybrid feature selection algorithm (Cuckoo and LION), and it was subjected to all the four classifier models. However, The rigorous experimentation showed that the accuracy of the SVM radial after full hyper parameter search and tuning give 95% accuracy as shown in Table VII. Correspondingly, the recall and precision values are also high. The better performance of the SVM radial algorithm can be attributed to the fact that the feature engineering process is paying off here. Secondly, to evaluate the consistency and validation of all classifier models, the ten-fold validation process was followed, and the standard deviation of each metric was noted. From Table VII and Table VIII, it can be observed that SVM radial have the lowest standard deviation for almost all the metrics, including recall and precision.

TABLE VII. PERFORMANCE ANALYSIS OF MACHINE LEARNING MODELS

Metric | Algorithm	KNN	SVM (R)	DT	RF
Accuracy Cuckoo	0.84	0.88	0.84	0.80
Accuracy Lion	0.85	0.88	0.84	0.80
Accuracy Hybrid	**0. 89**	**0.95**	**0.83**	**0.83**
F_ScoreCuckoo	0.87	0.89	0.83	0.82
F_ScoreLion	0.87	0.89	0.84	0.82
F_ScoreHybrid	**0.87**	**0.91**	**0.83**	**0.82**
PrecisionCuckoo	0.87	0.83	0.84	0.82
PrecisionLion	0.87	0.89	0.84	0.82
Precision Hybrid	**0.87**	**0.91**	**0.83**	**0.83**
Recall Cuckoo	0.80	0.84	0.83	0.82
Recall Lion	0.80	0.89	0.81	0.81
Recall Hybrid	**0.80**	**0.91**	**0.84**	**0.85**

TABLE VIII. STANDARD DEVIATION VALUES OF PERFORMANCE METRICS OF MACHINE MODELS

Metric | Algorithm	KNN	SVM (R)	DT	RF
Accuracy Cuckoo	0.047589	0.041732	0.055656	0.055328
Accuracy Lion	0.044991	0.040143	0.053754	0.056553
Accuracy Hybrid	0.047383	0.021967	0.055517	0.054785
F_Score Cuckoo	0.036791	0.041807	0.054889	0.057068
F_Score Lion	0.037383	0.041967	0.03517	0.054785
F_Score Hybrid	0.037589	0.031732	0.035656	0.055328
Precision Cuckoo	0.034991	0.030143	0.033754	0.056553
Precision Lion	0.037383	0.031967	0.03517	0.034785
Precision Hybrid	0.036306	0.010105	0.03458	0.032732
Recall Cuckoo	0.037185	0.032468	0.034845	0.035144
Recall Lion	0.037383	0.031967	0.03517	0.034785
Recall Hybrid	0.037383	0.021967	0.03517	0.034785

*KNN=k-nearset Neighbours, SVM= Support Vector Machine, DT= Decision Tree, RF= Random Forest.

It can further be noted that the KNN algorithm is second best in terms of accuracy, and its performance metrics have higher levels of deviations compared to the SVM radial. Similar observations can be made for Decision tree and random forest algorithms. Both these algorithms have performed in a range of eighties per cent with higher levels of deviations in their results when evaluated for validations and reliability using the ten-fold method. It should be emphasised that the selection of these machine learning algorithms was made after conducting a bibliographic examination of the relevant literature. The methods that are most frequently employed to handle the challenges of classification and limited datasets have been incorporated into this book in their most basic forms. Because the dataset used in this study is an aggregate of various datasets, this research report includes a comparison of the dataset utilised in this study with the current dataset.

V. Conclusion and Future Scope

There have only been a few studies in which numerous medical eye problems have been investigated using a single method. The same can be said for identifying relevant picture features using a combinational technique. It was completed through a rigorous procedure that included a great deal of experimenting. The process of developing a generic pipeline of algorithms to facilitate feature selection and automation of the classification process has been followed to completion. An in-depth investigation of the optimization process was carried out in order to identify the most appropriate features and methods that could be used for the construction of the feature matrix, and further investigation resulted in the development of a numerically stable pipeline of the algorithms. It should also be mentioned that the KNN method is the second most accurate algorithm in terms of accuracy, and that its performance metrics have larger levels of deviations when compared to the SVM radial algorithm. Observations similar to these can be made about the decision tree and random forest algorithms. When examined for validation and reliability using the ten-fold approach, both of these algorithms performed within an eighty percent confidence interval, with larger degrees of variances in their results in their results than when evaluated for accuracy.

Following a review of the literature on three keywords, the researchers chose the machine learning models for this work. The keywords were data fusion, eye illness classifiers, and image processing of the eyes. The procedure of picking the most accurate and stable classifier among the candidates was carried out with the assistance of a ten-fold algorithm, which was used to narrow down the field of candidates. This guaranteed that no time was wasted later on while assessing different machine learning models in the field. When hybrid algorithms (Cuckoo and LION) are used for feature engineering and dimension reduction, it has been discovered that there are extra benefits, and that this results in the generation of matrices with decreased features but complete coverage. This research was conducted under the guidance of an exploratory experimentation regime, and it has been discovered that the SVM radial algorithm is the most suited machine-learning model for the development of a multi-modality system that can detect eye abnormalities. In addition, it was discovered that some degree of hyper-parameter adjustment was required in some cases. After conducting an extensive grid search based on hyper-parameter tuning and feature engineering, it was discovered that the optimization strategy resulted in a higher accuracy level (0.95) for SVM than the previous approach.

In this study project, we attempted to develop a multi-disease detection system that would be capable of detecting three different forms of eye diseases: diabetic retinopathy, glaucoma, and cataract, among others. It is recommended that other diseases be added to the scope in the future, and that the detection range be broadened as well. This way, the scalability and generality of the model can be further strengthened.

Acknowledgment

Authors are highly thankful to the RIC department of IKG Punjab Technical University, Kapurthala, Punjab, India for providing the resources and opportunity to conduct this research work.

Declarations Funding

The authors received no specific funding for this work.

Conflict of Interest

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

Availability of Data and Material

The data that support the findings of this study is publically available on https://github.com/palavibhangu/retina_dataset

Availability of Code

The code will be provided once the paper has been conditionally accepted.

References

[1] W. Y. Ng et al., “Blockchain applications in health care for COVID-19 and beyond: a systematic review,” Lancet Digit. Heal., 2021, doi: 10.1016/s2589-7500(21)00210-7.

[2] A. Gupta and R. Katarya, “Social media based surveillance systems for healthcare using machine learning: A systematic review,” Journal of Biomedical Informatics, vol. 108. 2020, doi: 10.1016/j.jbi.2020.103500.

[3] T. F. Ursuleanu et al., “Deep learning application for analyzing of constituents and their correlations in the interpretations of medical images,” Diagnostics, vol. 11, no. 8, 2021, doi: 10.3390/diagnostics11081373.

[4] E. Omolara Abiodun et al., “A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities,” Neural Comput. Appl., vol. 33, doi: 10.1007/s00521-021-06406-8.

[5] P. Dhar and M. Z. Abedin, “Bengali News Headline Categorization Using Optimized Machine Learning Pipeline,” Int. J. Inf. Eng. Electron. Bus., vol. 13, no. 1, pp. 15–24, Feb. 2021, doi: 10.5815/IJIEEB.2021.01.02.

[6] S. Wang et al., “A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19),” Eur. Radiol., vol. 31, no. 8, 2021, doi: 10.1007/s00330-021-07715-1.

[7] J. Carp, “On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments,” Front. Neurosci., no. OCT, 2012, doi: 10.3389/fnins.2012.00149.

[8] P. Kartikeyan and G. Shrivastava, “Review on Emerging Trends in Detection of Plant Diseases using Image Processing with Machine Learning,” Int. J. Comput. Appl., vol. 174, no. 11, 2021, doi: 10.5120/ijca2021920990.

[9] H. Seo et al., “Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications,” in Medical Physics, 2020, vol. 47, no. 5, doi: 10.1002/mp.13649.

[10] A. Oniśko, P. Lucas, and M. J. Druzdzel, “Comparison of rule-based and Bayesian network approaches in medical diagnostic systems?,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2001, vol. 2101, doi: 10.1007/3-540-48229-6_40.

[11] L. Oakden-Rayner, “Exploring Large-scale Public Medical Image Datasets,” Acad. Radiol., vol. 27, no. 1, 2020, doi: 10.1016/j.acra.2019.10.006.

[12] Q. Abbas, “Glaucoma-Deep: Detection of Glaucoma Eye Disease on Retinal Fundus Images using Deep Learning,” Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 6, 2017, doi: 10.14569/ijacsa.2017.080606.

[13] C. M. Sanders, S. L. Saltzstein, M. M. Schultzel, D. H. Nguyen, H. S. Stafford, and G. R. Sadler, “Understanding the limits of large datasets,” J. Cancer Educ., vol. 27, no. 4, 2012, doi: 10.1007/s13187-012-0383-7.

[14] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, “A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, vol. 12614 LNCS, doi: 10.1007/978-3-030-71058-3_11.

[15] G. Lim, V. Bellemo, Y. Xie, X. Q. Lee, M. Y. T. Yip, and D. S. W. Ting, “Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review,” Eye Vis., vol. 7, no. 1, 2020, doi: 10.1186/s40662-020-00182-7.

[16] S. G. Francisco et al., “Dietary patterns, carbohydrates, and age-related eye diseases,” Nutrients, vol. 12, no. 9. 2020, doi: 10.3390/nu12092862.

[17] Y. Wang and S. Shan, “Accurate disease detection quantification of iris based retinal images using random implication image classifier technique,” Microprocess. Microsyst., vol. 80, 2021, doi: 10.1016/j.micpro.2020.103350.

[18] G. R. Hemalakshmi, D. Santhi, V. R. S. Mani, A. Geetha, and N. B. Prakash, “Classification of retinal fundus image using MS-DRLBP features and CNN-RBF classifier,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 9, 2021, doi: 10.1007/s12652-020-02647-y.

[19] O. Alfarraj, “A machine learning-assisted data aggregation and offloading system for cloud–IoT communication,” Peer-to-Peer Netw. Appl., 2020, doi: 10.1007/s12083-020-01014-0.

[20] G. C. Ozmen et al., “An Interpretable Experimental Data Augmentation Method to Improve Knee Health Classification Using Joint Acoustic Emissions,” Ann. Biomed. Eng., vol. 49, no. 9, 2021, doi: 10.1007/s10439-021-02788-x.

[21] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” J. Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[22] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, “A survey on machine learning for data fusion,” Inf. Fusion, vol. 57, 2020, doi: 10.1016/j.inffus.2019.12.001.

[23] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” Neural Computation, vol. 32, no. 5. 2020, doi: 10.1162/neco_a_01273.

[24] X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” Neural Computing and Applications, vol. 32, no. 19. 2020, doi: 10.1007/s00521-020-04748-3.

[25] S. Uddin, A. Khan, E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” doi: 10.1186/s12911-019-1004-8.

[26] J. Karthikeyan, S. H. P. Kumar, and K. Thirunavukkarasu, “Statistical techniques and tools for describing and analyzing data in Elt research,” Int. J. Civ. Eng. Technol., vol. 9, no. 11, 2018.

[27] A. H. Halim, I. Ismail, and S. Das, “Performance assessment of the metaheuristic optimization algorithms: an exhaustive review,” Artif. Intell. Rev., vol. 54, no. 3, 2021, doi: 10.1007/s10462-020-09906-6.

[28] A. Darwish, “Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications,” Futur. Comput. Informatics J., vol. 3, no. 2, 2018, doi: 10.1016/j.fcij.2018.06.001.

[29] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, “Ant colony optimization for text feature selection in sentiment analysis,” Intell. Data Anal., vol. 23, no. 1, 2019, doi: 10.3233/IDA-173740.

[30] P. Kaur and R. Kumar Singh, “An Efficient Approach for Content-Based Image Retrieval Using Cuckoo Search Optimization,” Int. J. Model. Optim., vol. 9, no. 2, pp. 77–81, Apr. 2019, doi: 10.7763/ijmo.2019.v9.688.

[31] L. Abualigah and A. Diabat, “A comprehensive survey of the Grasshopper optimization algorithm: results, variants, and applications,” Neural Computing and Applications, vol. 32, no. 19. 2020, doi: 10.1007/s00521-020-04789-8.

[32] S. S. Panicker and P. Gayathri, “Feature Selection Algorithms in Medical Data Classification: A Brief Survey and Experimentation,” in Lecture Notes in Electrical Engineering, 2020, vol. 601, doi: 10.1007/978-981-15-1420-3_90.

[33] K. K. Patro, A. Jaya Prakash, M. Jayamanmadha Rao, and P. Rajesh Kumar, “An Efficient Optimized Feature Selection with Machine Learning Approach for ECG Biometric Recognition,” IETE J. Res., 2020, doi: 10.1080/03772063.2020.1725663.

[34] D. Devarajan, S. M. Ramesh, and B. Gomathy, “A metaheuristic segmentation framework for detection of retinal disorders from fundus images using a hybrid ant colony optimization,” Soft Comput., vol. 24, no. 17, 2020, doi: 10.1007/s00500-020-04753-7.

[35] L. Parisi, N. RaviChandran, and M. L. Manaog, “Feature-driven machine learning to improve early diagnosis of parKinson’s disease,” Expert Syst. Appl., vol. 110, 2018, doi: 10.1016/j.eswa.2018.06.003.

[36] Sonali et al, “An approach for de-noising and contrast enchancement of retinal fundus image using CLAHE,” Optics and Laser technology., vol. 110, 2019, doi: 10.1016/joptlastec.2018.06.061.

[37] B. Sabir, F. Ullah, M. A. Babar, and R. Gaire, “Machine Learning for Detecting Data Exfiltration,” ACM Computing Surveys, vol. 54, no. 3. 2021, doi: 10.1145/3442181.

[38] V. T, S. M, A. Kumaravel, and K. B, “Gabor filter and machine learning based diabetic retinopathy analysis and detection,” Microprocess. Microsyst., 2020, doi: 10.1016/j.micpro.2020.103353.

[39] M. Rai and P. Rivas, “A Review of Convolutional Neural Networks and Gabor Filters in Object Recognition,” in Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020, 2020, doi: 10.1109/CSCI51800.2020.00289.

[40] N. Sompairac et al., “Independent component analysis for unraveling the complexity of cancer omics datasets,” International Journal of Molecular Sciences, vol. 20, no. 18. 2019, doi: 10.3390/ijms20184414.

[41] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, “Integration of multi-objective PSO based feature selection and node centrality for medical datasets,” Genomics, vol. 112, no. 6, 2020, doi: 10.1016/j.ygeno.2020.07.027.

[42] P. Kaur and R. K. Singh, “Content-based image retrieval using machine learning and soft computing techniques,” Int. J. Sci. Technol. Res., vol. 9, no. 1, 2020.