A Systematic Evaluation of Traditional Machine Learning Algorithms and Handcrafted Features Across Multiple Plant Disease Datasets

Abdulla Al Mahin Khan
CSSA
Multimedia University
Cyberjaya, Malaysia
abdulla.al.mahin@student.mmu.my

***Abstract*—**

This research paper is a systematic review of classical machine learning methods employed in the classification of plant diseases across datasets of varying complexity, aimed at comparing these methods to assess their suitability as a control group for the early detection of disease in Vanilla planifolia. Three publicly accessible datasets pertaining to plant diseases were employed: a comprehensive multi-class dataset of tomato leaves reflecting authentic field settings, an imbalanced dataset of potato leaves, and a controlled binary dataset of chili plants. Each dataset was utilized in a singular experimental pipeline comprising image preprocessing, feature extraction utilizing Haralick texture descriptors and HOG, and classification through SVM, RF, and K-NN. Five-fold stratified cross-validation, learning curve analysis, per-class recall, and confusion matrices were employed to evaluate performance. Experimental observations indicate that classical machine learning models exhibit great accuracy on simple, well-separated datasets but experience significant performance degradation on complicated multi-class datasets. The learning curves indicate early performance saturation with increased training data, suggesting a limitation in feature representation rather than in data availability. Furthermore, the examination of classes reveals consistent misclassification of visually similar disease categories, which is not reflected in overall accuracy metrics. The results demonstrate that traditional machine learning algorithms employing manually crafted features lack the representational capacity and predictive power necessary for accurately forecasting early plant disease in practical agricultural settings. The results advocate for the implementation of scalable and precise methods for learning disease detection features with deep learning.

Keywords*— *Deep Learning, Machine Learning, Transfer Learning, Feature Extraction, Cross Validation, Accuracy Metrics

I. Introduction

Timely and proper identification of plant diseases is important to the agricultural productivity, food security, and economical sustainability, especially in areas where crop production is highly dependent [1]. Manual inspection is the most common method of disease diagnostics conducted by the agricultural experts and is labor and cost intensive and not scalable as well as it is subject to human error [2]. The recent developments in computer vision and machine learning have made it possible to implement automated plant disease detection systems that consider visual symptoms manifested on the plant leaves with the help of digital pictures [3]. This task has been assumed with classical machine learning algorithms coupled with image processing and manual feature extraction because of its relatively cheap computational cost and simple to implement [2]. Many studies show high classification rates with the use of these methods to controlled data, taken in the same light and on the same background [5],[6]. Nonetheless, the reported performance on different datasets and in the experimental conditions also differs significantly, and it is unclear how well such solutions can be robust and able to generalize real-world agricultural conditions [4]. The current classification frameworks of plant diseases are usually based on the pipeline image acquisition, preprocessing, segmentation, feature extraction, and classification [2]. In this context, visual patterns that are caused due to a disease, such as the morphology of the lesions and surface deformities, are represented by handcrafted characteristics, such as texture, gradient, shape, and color descriptors [1], [14]. However, such handcrafted depictions are not flexible to visual changes and nuanced symptom development and thus cannot be used to identify disease at an early stage [1]. Although classical machine learning methods have been widely used, the majority of the available research considers overall accuracy in small or controlled datasets as a primary measure of performance [1],[2],[4]. The assessment practices used in such a way confound class-level failure modes and do not give adequate information as to the reliability of the models with changing data complexity and class distributions. This is especially problematic to the case of plant disease detection in its early stages, where a false negative can be subject to much downstream effects. This study aims to evaluate the applicability of classical machine learning methods to plant disease detection in the early stages carefully through systematic evaluation of model behavior using datasets that are more realistic and structurally complex. In this direction, the various handcrafted combinations of feature classifiers are assessed in terms of class-level performance, such as per-class recall and confusion matrix analysis, to reveal misclassification behaviour not reflected by the aggregate performance metrics. Section II outlines the datasets used. Section III is all about the methodologies handpicked from all different methods used previously which are best suited for this particular case. Section IV shows the results of the experiment and analysis. Section V accounts for the limitations and implications of classical machine learning approaches, as observed. Lastly, Section VI summarizes the paper and proposes future research directions according to the latest deep learning and explainable artificial intelligence methodologies.

II. Dataset Overview

A. Tomato Plant

The primary dataset is comprised of a multi-class multi-scale tomato leaf disease dataset of large scale data that includes several visually related categories of diseases (10 classes in total) as well as healthy samples. These multi-class datasets are popularly utilized to test the scaleability of machine learning classifiers since when the disease classes are fine-grained, there is high inter-class feature overlap and difficulty in classification is high [1], [6].

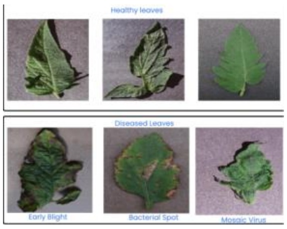

Fig 1. Different classes of unhealthy tomato leaves

B. Potato Plant

The second data set has concentrated on potato leaf disease and has a very high degree of class imbalance, whereby there are very few healthy samples in comparison to the diseased classes.

C. Chili Plants

Dataset three is a dichotomous chili plant data, which is a studied data, where the data were taken under controlled settings, with only healthy and unhealthy plants classifying these two categories. Binary datasets that have a small intra-class variability are usually treated as baseline benchmarks in which classical machine learning models tend to show high apparent performance because the decision boundaries are simplified [2], [5].

III. Methodology

The general pipeline adheres to conventional measures in disease classification using images and focused on controlled comparison across datasets and measures of assessment [2], [10]. The following describes an experimental pipeline that investigates the relationships between performance measurement and project management.

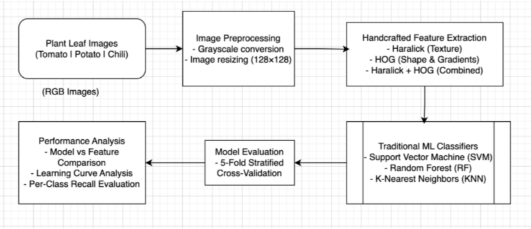
Plant Leaf Images
(Tomato I Potato I Chili)
(RGB Images)
Image Preprocessing
- Grayscale conversion
- Image resizing (128×128)
Handcrafted Feature Extraction
- Haralick (Texture)
- HOG (Shape & Gradients)
- Haralick + HOG (Combined)
Traditional ML Classifiers
- Support Vector Machine (SVM)
- Random Forest (RF)
- K-Nearest Neighbors (KNN)
Model Evaluation
- 5-Fold Stratified
Cross-Validation
Performance Analysis
- Model vs Feature
Comparison
- Learning Curve Analysis
- Per-Class Recall Evaluation

Fig 2. Flowchart of general pipeline in desease classification mechanism

A. Experimental Pipeline Overview

The methodology proposed includes image processing, manual feature exploration, training of a model with classical machine learning algorithms, and the evaluation of performance with the help of effective validation methods [2]. Comparable preprocessing procedures, feature extraction methods, and evaluation procedures were done in all datasets to provide a fair comparison [10].

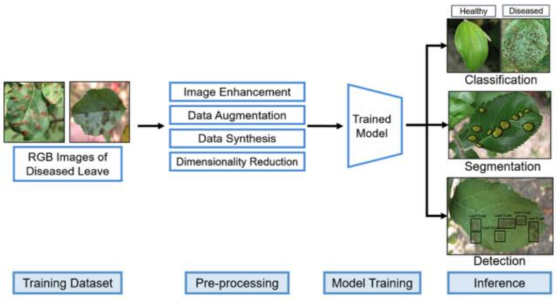

Fig 3. Generalized overview of the whole pipeline

B. Image Preprocessing

Preprocessing of images involved elimination of noise, normalization of visual features and enhancement of consistency of feature extraction across datasets [2]. Some of the preprocessing tasks that are usually performed are resizing, grayscale conversion and normalization that would address the variation due to the conditions of taking the image including lighting and background clutter [14]. It has been demonstrated through such preprocessing methods that stability of handcrafted feature descriptors is enhanced due to reduction of irrelevant variability and maintenance of disease-related visual patterns [5], [12].

C. Feature Extraction

1)Handcrafted Feature Selection

To turn raw images into structured numerical representations, which can be fed into classical machine learning classifiers, handcrafted feature extraction was used to transform raw images [14].

2) Haralick Texture Features

Gray-Level Co-occurrence Matrices (GLCM) were used to obtain texture features of Haralick in an effort to capture the spatial relationships of pixel intensities [12]. These properties have proven to be useful in the classification tasks based on texture especially in medical imaging and biometric recognition where structured texture patterns are salient [12], [13]. The haralick features are chosen in this study as they are used to measure the capacity of the features to capture the texture changes that occur in plant leaves due to diseases in various levels of dataset complexity [1].

3) Histogram of Oriented Gradients (HOG)

HOG has been used to extract the information of edge and gradient orientation of the leaf images [15]. HOG has been demonstrated to do well in the tasks that involve structured shapes and regular object boundaries, especially in conjunction with SVM classifiers [15], [16]. This is made possible by incorporation of HOG where comparative analysis of both texture based and gradient based handcrafted descriptors can be achieved within the same experimental framework [14].

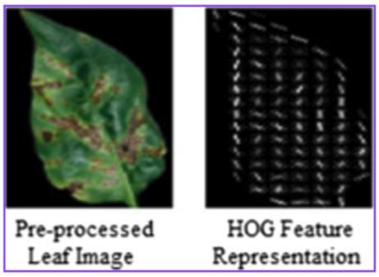

Fig 4. HOG Feature extraction

D. *Machine Learning Classifiers*

1) Support Vector Machine (SVM)

Support Vector Machine (SVM) has been chosen because it has a high generalization property in high-dimensional feature spaces, and it is widely used in the literature of plant diseases classification [8]. Non-linear class boundaries can be modeled with kernel-based SVM, and it is therefore an appropriate model that is capable of complex visual feature distributions [9].

2) Random Forest (RF)

Random Forest was used as one of the ensemble learning methods to decrease overfitting and enhance resistance to noise in feature representations [7]. It is great in the ability of aggregating choices across numerous trees and thus it is useful in datasets of heterogeneous feature distributions [6].

3) K-Nearest Neighbors (KNN)

K-Nearest Neighbors was also used as a distance-based baseline classifier to assess how sensitive handcrafted features are to feature-space proximity and class imbalance [10]. Although KNN is rather simple, it is still widely applied in the plant disease classification research and serves as a good point of reference compared to more advanced classifiers [7].

E. *Model Training and Validation*

In order to have good confidence in estimating the performance of model generalization, k-fold cross-validation was used in both training and evaluation [10], [17]. K-fold cross-validation minimizes the variance when compared to single holdout techniques and allows researchers to use scarce data effectively [17]. The adoption of unified validation approaches in all experiments will help to assume that the observed performance variations are due to the properties of the datasets and behavioral patterns of the models and not evaluation bias [11].

F. *Performance Evaluation Metrics*

Accuracy, per-class recall, and the analysis of the confusion matrix were used to determine model performance [18]. Although accuracy gives a high-level overview of the classification performance, it can hide systematic misclassification patterns especially in imbalanced or multi-class data [18].

As Juba and Le [21] have shown, overall accuracy is theoretically immune to class imbalance, and could be high even when a classifier repeatedly fails to detect minority or rare classes, giving false conclusions of good performance.

The authors demonstrate that the imbalance in the classes fundamentally influences the precision-recall-based measurements and give a more significant signal of the behavior of the classifier in imbalanced dataset, where errors are classified as positive are particularly important [21]. This is of particular concern to the detection of plant diseases at an early stage, where a false negative detection of a diseased sample can have far reaching consequences in the downstream decision-making, even though high accuracy is reported.

IV. Results and Analysis

A. *Classification Performance*

In tomato data, the accuracy also differed significantly in various feature-model combinations (between about 0.54 and 0.78) with the reported performance in mean and standard deviation since repeated cross-validation was used, and it was possible to evaluate the stability of the performance. Comparatively, the potato data provided strong accuracy values (0.82-0.91) even with a strong class imbalance whereas the binary chili dataset had all uniformly high accuracy (0.80-1.00) values, indicating the stable acquisition conditions and clear separation of classes in this case.

TABLE I. CLASSICAL PERFORMANCE – TOMATO DATASET

Tomato Plant Dataset			
Feature	SVM	Random Forest	KNN
Haralick	0.65±0.0	0.66±0.0	0.59±0.01
HOG	0.75±0.0	0.61±0.0	0.54±0.00
Haralick+HOG	0.78±0.0	0.72±0.0	0.56±0.00

TABLE II. CLASSICAL PERFORMANCE – POTATO DATASET

Potato Plant Dataset			
Feature	SVM	Random Forest	KNN
Haralick	0.86	0.86	0.85
HOG	0.88	0.82	0.82
Haralick+HOG	0.91	0.87	0.83

TABLE III. CLASSICAL PERFORMANCE – CHILLI DATASET

Chili Plant Dataset			
Feature	SVM	Random Forest	KNN
Haralick	1.00	1.00	1.00
HOG	0.90	0.90	0.81
Haralick+HOG	0.90	0.92	0.80

B. Model vs Feature Analysis

Several classifiers were also tested on the same handcrafted feature sets to understand that performance constraints were due to the choice of classifier or representation of features. The experiments were all performed with constant preprocessing and validation conditions to remove the influence of feature representation. Disease classes with low recall using a single classifier always had low recall in other classifiers, especially in the multi-class tomato dataset. Conversely, varying performance change between switching between feature types in which hand crafting was used was greater than switching between classifiers. Nevertheless, the best-performing feature-classifier combinations did not work in complex data sets, stabilizing at the class-level recall.These results suggest that the issue of misclassification is mainly caused by constraints on handcrafted feature separability as opposed to choice of classifier.

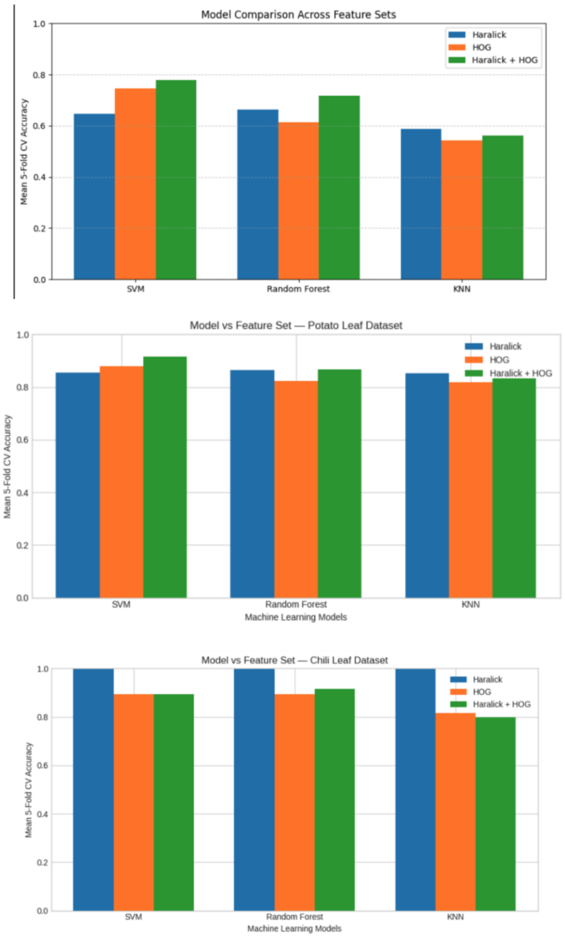

Fig 5. Charts illustrating model performance against features for all datasets

C. Learning Curve Analysis

To investigate the correlation between the size of training data and the performance of models, learning curves have been studied. In the case of the chili dataset, the learning curves of all the classifiers were also found to converge quickly, and the saturation of the performance was reached after fairly small training set sizes. Learning curves in the potato data set, however, showed no more than moderate improvements as more data was used; but early saturation was observed, especially on the minority classes. In the tomato case, learning curves were not increasing with additional training data, but rather they began to stabilize and then improved insignificantly with increased training data. These tendencies indicate that the performance constraints cannot be explained only by the lack of sufficient data but can be also determined by the feature representation capacity.

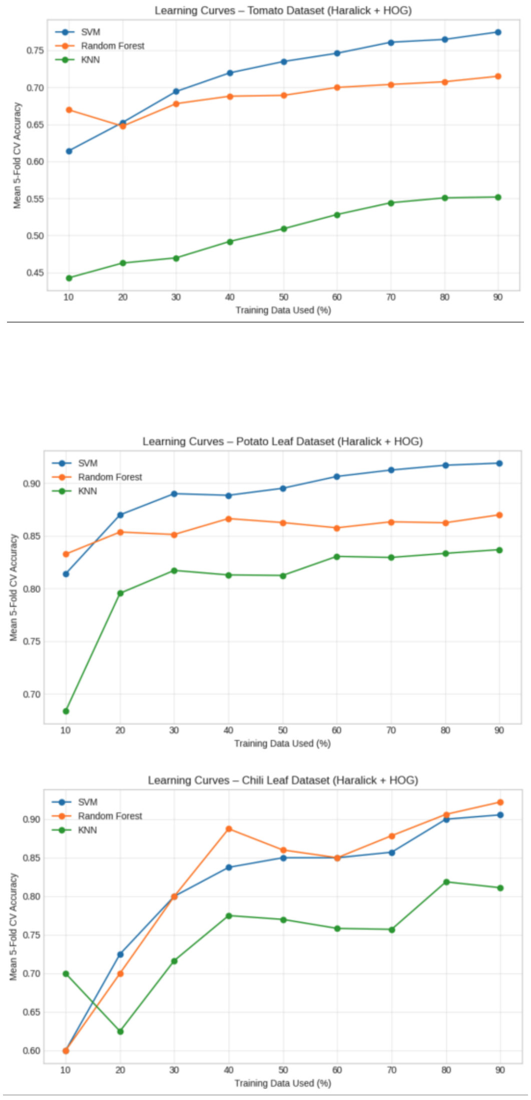

Fig 6. Line charts showing learning curves against increasing dataset sizes

D. Per-Class Recall Analysis

The behavior of the classifier in the dataset and classifier-wise prediction was checked by per-class recall heatmaps. In the case of tomato data, recall values differ significantly in accordance to the disease classes with the lowest recall of about 0.16 and the highest recall of about 0.96 based on disease class and model. A number of classes tend to have low recall in Support Vector Machine, Random Forest, and K-Nearest Neighbors, such as 0.16-0.42 in K-Nearest Neighbors and 0.38-0.66 in the Random Forest. Although there are variations in absolute values, the patterns of relative recall among the different classifiers are similar and proves that the main source of misclassification is the limitations in the representation of features by hand, rather than a choice of classifier. This is an indication that some of the classes of tomato diseases have similar visual appearances that cannot be easily distinguished using the classical features.

In general, these results indicate that aggregate accuracy obfuscates systematic failures on a class level with multi-class and imbalanced configurations being the most prominent examples. Therefore, confusion matrix analysis is needed to determine certain inter-class misclassification patterns and to describe the perceived drop in recall in complex and realistic data.

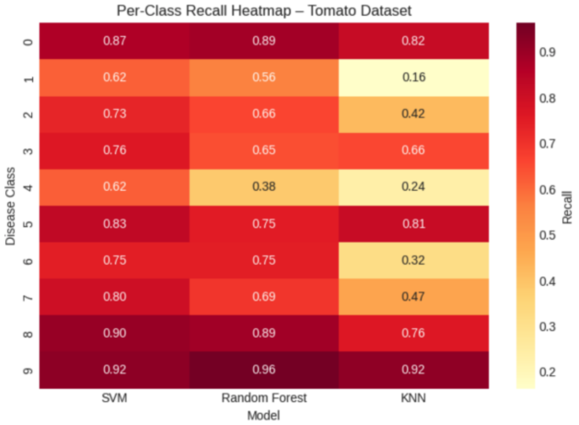

Fig 7. Per Class Recall Heatmap for Tomato Dataset

E. *Confusion Matrix Analysis*

To examine further the patterns of misclassification at the class level, confusion matrices were examined. In the tomato dataset, confusion matrices showed a wide area of confusion between various sets of diseases that had similar visual traits.

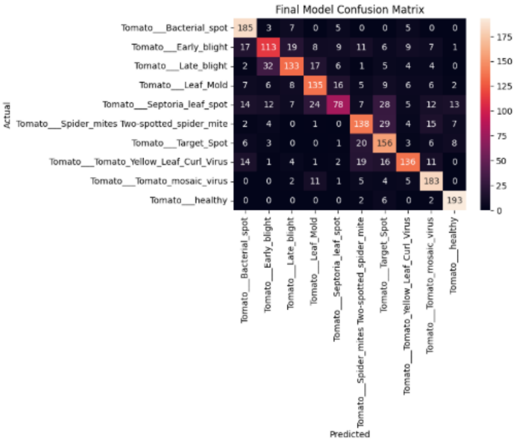

Fig 8. Confusion Matrix – Tomato Dataset

There was also strong evidence of misclassification between diseases that had similar texture and appearance of lesions, which could indicate a low level of discriminative ability of handcrafted features in systems with multiple classes.

V. Conclusion and Future Work

This paper critically reviewed the classical machine learning methods in the plant disease detection task through datasets of sequentially increasing structural complexity. Experimental findings prove that though handcrafted feature-based models have high apparent performance in controlled binary environments, their performance cannot be repeated in multi-class and imbalanced environments when judged by class-level metrics. Recall and confusion matrix analyses of each class show systematic misclassification trends which would not appear when using aggregate accuracy measures. These results show that the constraint on the performance of most of the feature representations is mainly due to the lack of capacity to represent the features, but not the choice of a classifier, which presents inherent limitations to conventional machine learning pipelines in the field of early detection of plant diseases.

The future working process will be oriented on enhancing the robustness of classification based on the current deep learning approach. To take advantage of the hierarchical feature representations that are learned on massive image datasets, transfer learning with pretrained convolutional neural networks will be considered. The strategies to be used to address the issue of class imbalance and enhance model generalization in limited and skewed data will be data augmentation strategies. Moreover, their explainable AI methods will be implemented to improve transparency of the model and give visual reasons behind the prediction of diseases, contributing to trust and interpretability of agricultural decision-making systems.

References

[1] P. Yerunkar, A. Ganer, and K. Kalbande, “Comprehensive review on machine learning algorithms for plant disease detection,” Proc. 2024 Parul Int. Conf. Eng. Technol. (PICET), Nagpur, India, 2024.

[2] F. Z. Touati, L. Loukil, and B. Amrane, “Plant diseases detection using machine learning and image processing: Review,” Proc. 4th Int. Conf. Embedded & Distributed Systems (EDiS), 2024.

[3] L. Viveka K. et al., “Machine learning-driven object detection techniques for automated identification and spatial localization of plant diseases,” Proc. 3rd Int. Conf. Augmented Intelligence and Sustainable Systems (ICAISS), 2025.

[4] V.S Babu and R. Satheesh, “A comparative study on disease detection of plants using machine learning techniques,” Int. Conf. Advanced Computing and communication systems (ICACCS), 2021.

[5] R. U. Rahamathunnisa et al., “Vegetable disease detection using K-means clustering and SVM,” Proc. 6th Int. Conf. Advanced Computing & Communication Systems (ICACCS), 2020.

[6] P. Mekha and N. Teeyasuksaet, “Image classification of rice leaf diseases using random forest algorithm,” Proc. Int. Conf. Digital Arts, Media and Technology (DAMT), 2021.

[7] K. A. Hemanthkumar and P. S. Bharathi, “Improved accuracy of plant leaf classification using random forest classifier over K-nearest neighbours,” Proc. Int. Conf. Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022.

[8] Ö. Karal, “Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation,” IEEE, 2020.

[9] Kai Chen, Heng Yeo, “Arithmetic optimization algorithm to optimize support vector machine for chip defect identification,” M2VIP, 2022.

[10] Kaushika Pal, Dr. Biraj “Data classification with k-fold cross-validation and holdout accuracy estimation methods with five different machine learning techniques,”Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020.

[11] Fengming Chang, H.C Cheng, Hsian Chuang Liu “Double k-fold cross-validation in support vector machines,” 9th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2015.

[12] S. Ribaric and M. Lopar, “Palmprint recognition based on local Haralick features,” IEEE Trans. Image Process, 2012.

[13] S. Vijayarani and M. Vinupriya, “Automated diagnosis of glaucoma using Haralick texture features,” Proc. Int. Conf. Signal Processing, 2016.

[14] Sideshwar Routray, Arun Kumar Ray, C. Mishra, “Analysis of various image feature extraction methods against noisy images: SIFT, SURF and HOG,” Int. J. Comput. Vision, IEEE, 2017.

[15] A. Rahman et al., “Handwritten Bengali numeral recognition using HOG based feature extraction algorithm,” Proc. IEEE Int. Conf. Informatics, Electronics & Vision, 2018.

[16] Siwar Rekik, A. Al Otaibi, Sarah A., “Face identification using HOG-PCA feature extraction and SVM classifier,” Seventh International Women in Data Science Conference at Prince Sultan University (WiDS PSU), 2024

[17] M. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proc. Int. Joint Conf. Artificial Intelligence (IJCAI), 1995.

[18] D. M. Powers, “Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation,” J. Mach. Learn. Technol., vol. 2, no. 1, pp. 37–63, 2011.

[19] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 6, 2020.

[20] T. Saito and M. Rehmsmeier, “The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” PLoS ONE, vol. 10, no. 3, 2015.

[21] B. Juba and H. S. Le, “Precision–recall versus accuracy and the role of large data sets,” *Proc. 33rd AAAI Conf. Artificial Intelligence (AAAI)*, Honolulu, HI, USA, pp. 4039–4046, 2019.