Applied Computer Science, vol. 20, no. 4, pp. 77–99
doi: https://doi.org/10.35784/acs-2024-41

*Submitted: 2024-09-26 | Revised: 2024-10-18 | Accepted: 2024-10-21*

*Keywords: Pupil Diameter (PD), Major Depressive Disorder (MDD), Machine Learning (ML), Hilbert–Huang Transform (HHT), Cross-Validation (CV)*

*Islam MOHAMED* [0009-0001-4408-7190]*,
*Mohamed EL-WAKAD* [0000-0003-2637-1048]**, *Khaled ABBAS* [0009-0002-0913-4163]***,
*Mohamed ABOAMER* [0000-0002-4433-776X]****,
*Nader A. Rahman MOHAMED* [0000-0001-7680-306X]*****

# PUPIL DIAMETER AND MACHINE LEARNING FOR DEPRESSION DETECTION: A COMPARATIVE STUDY WITH DEEP LEARNING MODELS

**Abstract**

*According to the World Health Organization, the Global Mental Health Report estimated that between 251 and 310 million individuals worldwide experienced depression during the first year of the COVID-19 pandemic. Most methods for detecting depression rely on clinical diagnoses and surveys. However, the American Psychiatric Association reports that over 50% of patients do not receive appropriate treatment. This study aims to utilize machine learning and pupil diameter features to identify depression and evaluate the accuracy of these classifiers in comparison to our previous deep learning model. While limited research has explored the use of pupillary diameter as a classification tool for distinguishing between individuals with and without depression, several studies have focused on EEG signals for this purpose. The study involved 58 participants, with 29 classified as depressed and 29 as healthy. The classification was based on statistical features extracted from the Hilbert-Huang Transform. Results showed a significant improvement in average accuracy compared to the authors' prior work, with the current study achieving 77.72% accuracy, compared to 64.78% in their previous research. Machine learning methods, particularly Bagging, outperformed deep learning models such as AlexNet when classifying data from the left and right eyes individually (90.91% vs. 78.57% for the left eye; 90.91% vs. 71.43% for the right eye). However, when combining data from both eyes, deep learning using AlexNet demonstrated superior performance (98.28% accuracy compared to 93.75% using Bagging with statistical features from both eyes). Despite the higher accuracy of deep learning, machine learning is recommended for its faster execution times.*

* Helwan University, Faculty of Engineering, Biomedical Engineering Department, Cairo, Egypt; Higher Technological Institute, Biomedical Engineering Department, 10th of Ramadan City, Egypt
** Future University, Faculty of Engineering and Technology, Biomedical Engineering Department, New Cairo, Egypt
*** Higher Technological Institute, Electronics and Communication Department, 10th Ramadan City, Egypt
**** Majmaah University, College of Applied Medical Sciences, Medical Equipment Technology Department, Majmaah 11952, Saudi Arabia
***** Misr University for Science and Technology, Faculty of Engineering, Biomedical Engineering Department, Giza, Egypt, nader.shaaban@must.edu.eg

## 1. INTRODUCTION

The Global Mental Health Report indicated that between 251 and 310 million persons worldwide experienced depression during the first year of the COVID-19 pandemic ( World Health Organization, 2022). Most depression detection methods use clinical diagnoses and subjective structured scales, which are subjective, time-consuming, and resource-intensive. As a result, traditional methods may delay diagnosis and treatment in many cases, even in severe cases. The American Psychiatric Association reports that over 50% of patients do not get appropriate therapy (Skowron et al., 2022).

There has been a growing focus in recent years on investigating objective physiological markers for the diagnosis of depression. Out of these markers, pupil diameter (PD) has shown as a very promising measure for differentiating between those with depression and those without. The path of this research, which directed author's attention towards PD, represents a wider pattern in the integration of psychology and artificial intelligence with the objective of creating more precise and effective algorithms to diagnose the depression.

Recent developments in both fields have emphasized the capacity of several physiological indicators in detecting depressed conditions. For example, studies using response time measured by (Li et al., 2014), functional magnetic resonance imaging (fMRI) by Drysdale et al. (2017) and electroencephalography (EEG) by Newson & Thiagarajan (2019) have shown their effectiveness in identifying depression. While each of these methods provides distinct perspectives on the physiological basis of depression, they frequently need complicated and resource-intensive procedures.

Eye movement measurements have therefore become a feasible option in this context. Previous research showed eye movements as a highly important behavioral indicator for diagnosing depression according to (Suslow et al., 2020; Zhu et al., 2020). Particularly, the measurement of pupil diameter, which is a measurable component of eye movement, has been shown to be a dependable and may be accurately measured automatically (Zhong et al., 2022). Moreover, the study done by (Zhao et al., 2019) evaluated several eye movement characteristics for the purpose of identifying human emotions. The researchers found that pupil diameter had a higher level of discriminative ability in categorizing emotions compared to other indicators of eye movement.

The study (Siegle et al., 2011) provided more evidence for the importance of PD in depression research, where, according to (Siegle et al., 2001), by showing statistically significant differences in the pupillary responses of depressed patients compared to healthy controls. The results of their study revealed that those with depression had more prominent and prolonged dilatation of the pupils in reaction to emotional stimuli, compared to those without depression. In addition, Jones et al. (2015) investigated the correlation between motivational states and affective processes, providing evidence that the depressed people had more prominent pupillary responses. The study conducted by (Wang et al., 2014) showed a significant difference in the reactions of depressed individuals to light stimuli compared to normal controls. This finding supports the idea that pupil width might be a reliable indicator of depression. (Siegle et al., 2003) conducted a more precise examination of the initial differences in PD between those with depression and those without and discovered a significant difference. The results of their investigation indicated that the average initial PD for those with depression was 3.5 mm, whereas for those without depression it was 4.0 mm. This difference had a statistically significant effect size (d = 0.8,

$p < 0.01$). Such a significant difference again and again highlights the capacity of baseline PD as a biomarker for depression.

In brief, our study evaluates the importance of pupil diameter as a diagnostic instrument for depression, expanding upon an increasing amount of information that substantiates its effectiveness. By prioritizing this physiological indication, the goal is to make a valuable contribution to the development of more unbiased and easily understandable techniques for detecting depression, thereby improving our capacity to identify and treat this widespread mental health disorder.

Building on the growing number of research investigating physiological markers as indicators of depression, several studies have effectively employed pupil data to distinguish between healthy individuals and those with depression.

For instance, (Ding et al., 2019) used EEG data to classify healthy controls and depressed patients. Participants observed neutral stimuli and emotional responses while low-cost, portable instruments captured their eye tracking data, EEG, and galvanic skin responses. The binary classification model was trained using ML classifiers, where Logistic Regression (LR) achieved the highest classification F1-score of 80.70%. Similarly, (Schultebraucks et al., 2022) used a deep neural network to categorize people with MDD using movement parameters (e.g., pupil dilation), speech prosody, facial features, and natural language content. This technique achieved an area under the curve (AUC) of 0.86. These studies highlight pupil data as a valuable, efficient tool for diagnosing depression.

Based on the insights from the authors' literature review of this paper, the proposed approach for this study was designed to develop a unimodal and cost-effective approach to diagnosing depression using pupil diameter readings only. To ensure robust results, the authors implemented well-established preprocessing techniques, proven effective in previous studies for similar data types. These techniques included Z-Score Normalization, Median Filtering, Signal Detrending, and the Hilbert-Huang Transform (HHT).

For instance, (Zhu et al., 2016) who used Z-Score normalization for examining internet activity time-frequency analysis for early depression identification. They used categorization and prediction models to identify the mental states of individuals. For accurate comparisons, they used z-score normalization to standardize internet behavior features. They used the Naive Bayes algorithm and attained 76.8% precision.

Similarly, (Schumann et al., 2015) who used Median Filtering for investigating cardiovascular rhythms, PD fluctuations, and respiratory relationships. They found pupil unrest lateralization in the PDs of 29 individuals. They used a median filter with a 600-ms time window and 200-ms temporal smoothing to reduce eye blink impact. The relationship with baroreflex sensitivity and vagal heart rate regulation emphasizes the importance of left pupil changes in determining the status of the autonomic nervous system.

Further supporting the utility of signal processing in physiological research, (Kramarić et al., 2019) who used polynomial detrending with Heart Rate Variability (HRV) data to detect acute stress in infants. The analysis of Receiver Operating Characteristic curve showed HRV indicators diagnostic usefulness as clinical markers. After polynomial detrending eliminated signal trends, the signals were 87.5% accurate. This research shows infant HRV analysis may identify acute stress.

Finally the authors utilized the extracted statistical features from HHT output and used the cross validation technique for machine learning as supported by research conducted by (Aboamer et al., 2014) who employed time, frequency, HH, and a hybrid combination of

those features to classify healthy and paranoid patients. Hybrid features had the greatest classification success rate of 95.24 % in validation and 97.5 % in training using IMF1 and six folds.

Several other researchers have conducted studies that are comparable to ours. Later, we will compare our findings with theirs in the results and discussion sections. Those researchers are:

(Zeng et al., 2019) who proposed a unimodal approach using eye movement data, which is more accessible and cost-effective compared to EEG and functional magnetic resonance imaging (FMRI) data. Their model achieved a 76.04% accuracy result by using SVM classifier with a 10-Fold CV.

(Li et al., 2020) who proposed a multimodal approach integrating eye movement behavioral features and physiological signal features. They also evaluated a unimodal approach using eye movement physiological signals, which achieved 76.84% accuracy by using a KELM classifier with a 10-Fold CV.

(Shen et al., 2021) who introduced a unimodal approach based on psychological features extracted from eye movement data, considering both free viewing and frame tracking stages. Their model achieved a 77% accuracy result by using SVM classifier with a 10-Fold CV.

(Zhu et al., 2023) who developed a multimodal approach (MIBFM) that uses pupil area signals to select EEG electrodes based on mutual information. They also tested the use of pupil area as a unimodal approach which achieved a 72.1% accuracy result through a 10-Fold CV.

(Yang et al., 2023) who proposed a multimodal approach (TSTCCA) for depression recognition, combining facial expression and pupil diameter, and evaluated pupil diameter only as a unimodal approach, which achieved a 64.78% accuracy result by utilizing SVM classifier.

Few studies have investigated PD as a marker of normal and depressed states. In contrast, several studies have focused on EEG signals to discriminate between normal and depressed individuals. Pupillometry may be useful in clinical settings because some studies have consistently shown differences in pupillary responses between non-depressed and depressed individuals. Where the pupil area (Ding et al., 2019) yields the most reliable findings, not surpassing 90%, PD's potential as a diagnostic tool for normal and depressed individuals is intriguing. Additionally, computational tools may improve diagnostic speed and accuracy, which might help us comprehend psychiatric disorders. This article presents a novel method for identifying depression using ML and precise analysis of PD-derived signals.

This research aims to explore the potential of machine learning (ML) and harness the power of the Hilbert-Huang Transform (HHT) technique to extract statistical features. Based on the analysis of signals obtained from pupil diameter (PD, these features will distinguish individuals without depression from those with depression. This study successfully achieved a high level of accuracy in classifying depression using PD data only. The study will examine the viability of PD as a diagnostic tool, with the goal of surpassing the accuracy criteria attained by current approaches. This study aims to make a valuable contribution to the field of depression identification by providing insights into the efficacy of using HHT and PD-related signals to assist physicians in the diagnosis and monitoring of clinical depression.

## 2. METHOD

### 2.1. Proposed approach

This study presents a comprehensive methodology designed to process and analyze the physiological data of individuals (depressed or healthy) PD, emphasizing the implementation of robust techniques for feature extraction and classification. The process begins by uploading raw data pertaining to each PD measurement across individual cases into Python version 3.10.9 interface. Then, these data undergo a series of preprocessing steps aimed at standardizing signal sizes and reducing noise and baseline drift. These steps include data truncation to ensure uniform signal sizes, as well as sliding window median filtering spanning intervals of 200 ms and 600 ms, with alongside polynomial detrending techniques. Following preprocessing, the signals are subjected to decomposition using two distinct methodologies: Empirical Mode Decomposition (EMD) to yield Intrinsic Mode Functions (IMF), and conversion of each resulting IMF into a Helbert Hang Spectrum. This multi-step decomposition process is crucial for extracting meaningful information from physiological data. After decomposition, statistical features are extracted from each signal generated through the decomposition methods.

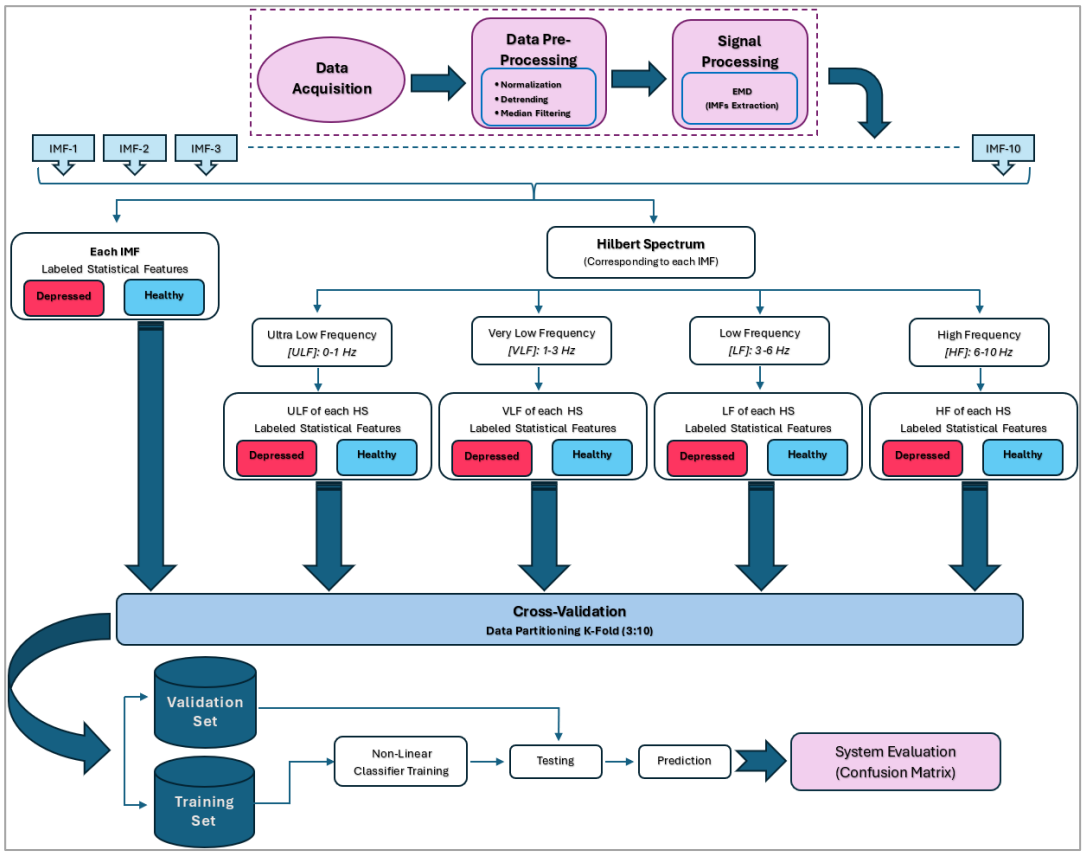


**Fig. 1. Proposed Machine Learning Model**

These features encompass a diverse range, including mean, variance, standard deviation, median, maximum, minimum, first quartile (Q1), third quartile (Q3), skewness, and kurtosis. Such a comprehensive set of features enables a nuanced understanding of the underlying physiological processes within the human eye. Finally, classification tasks are undertaken employing six distinct machine learning algorithms: Support Vector Machine (SVM), Decision Tree, k-Nearest Neighbors (KNN), Bagging, Gradient Boosting, and Random Forest. To ensure the robustness of the findings, cross-validation is employed post data partitioning using a K-fold technique with a ratio of 3:10. This rigorous evaluation framework aims to assess the performance of the classification algorithms across various scenarios, thereby enhancing the reliability and applicability of the proposed methodology. Figure 1 illustrates the proposed Machine Learning model.

### 2.2. Data collection

A dataset from a recent study was used on cardiovascular dysfunction and pupillary fluctuations in major depression (Schumann et al., 2015). The MP150 polygraph from BIOPAC Systems Inc. in Goleta, CA, USA, recorded cardiovascular and pupillometric data on a group of participants. The research included 29 unmedicated persons with significant depression (21 females, 8 males, mean age: 37.8 ± 12.2, mean BMI: 23.8 ± 4.1) and a control group of 29 healthy persons (21 females, 8 males, mean age: 36.9 ± 12.5, mean BMI: 23.5 ± 4.1) (Schumann et al., 2015). Participants were individually hospitalized in inpatient wards under the diagnosis of a staff psychiatrist. All subjects fulfilled DSM-IV major depressive disorder criteria. Importantly, none of these individuals had used antidepressants before the study. The Hamilton Rating Scale for Depression was used to measure depression severity. To reduce clinical deficiencies, Hamilton (1960) recommended conversations with patients and healthy volunteers. All individuals also took Beck's 1961 depression evaluation. No psychiatric, neurological, or clinically significant disorders were present in the healthy control group. All participants gave informed written permission according to the Jena Ethics Committee policy. The studies were conducted between 2:00 pm and 7:00 pm. The exam room was quiet and lit by a low-intensity ambient light. Participants wore headphones to reduce noise. A monitor above the sofa presented a dark grey ellipse for fixation. The ambient temperature was 22°C. The room remained silent throughout the 20-minute exam. A beamer provided constant illumination. To measure eye movement in the pupillometric system region, a 22-inch monitor projected an ellipse across the screen. The first five minutes were omitted from analysis to let participants adjust. Pupil size was measured at 4-millisecond intervals using SensoMotoric Inc.'s RED 250 infrared camera system (Schumann et al., 2015). The requirement to capture quick pupillary responses to emotional processing prompted this frequency of pupil size recording. To capture PD changes with great temporal precision, the 4ms period was used. The PD data was recorded in CSV files and then uploaded to Python version 3.10.9 as a data frame for further usage in subsequent phases.

### 2.3. Data preprocessing

The following steps aim to improve model accuracy. Unwanted noise may obscure model learning. Such noise must be eliminated. Figure 2 displays the unprocessed data signal of a sample that includes one example of the left pupil of both depressed and healthy patients, as

well as the enhancement achieved by implementing the following four steps: Normalization, Median Filtering for 200ms, Detrending, and Median Filtering for 600ms.

### 2.3.1. Normalization (Z-Score)

Data normalization using Z-Score is the first stage in the process of data processing to eliminate biases and variances. This enhances research reliability. Normalization standardizes data to compare variables. This facilitates data integration and statistical interpretation. Z-Score can be calculated as follows: where, x represents a variable, μ stands for the mean, and σ denotes the standard deviation.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Z-score measures the gap size between x and μ, impacted by the standard deviation σ. A negative $z$ value indicates that $x$ is below the mean μ (Zhu et al., 2016).

### 2.3.2. Median filtering (200ms)

Secondly, the normalized data was subjected to median filtering. The median filter is known to reduce random noise, enhance signal quality, and identify key trends and patterns. This study employs the median filter to address eye blinks, which cause sudden PD declines.

The median of n observations $x_i$ , $i = 1,..n$ is denoted by $med(x_i)$ and it is given by:

$$med(x_i) = \begin{cases} x_{(v+1)} & where\ n\ is\ odd,\ and\ \ n = 2v + 1 \\ \frac{1}{2}\left(x_{(v)} + x_{(v+1)}\right) & where\ n\ is\ even,\ and \quad n = 2v \end{cases} \tag{2}$$

where: $x_{(i)}$ – denotes the $i$-th order statistic.

A one-dimensional median filter of size $n = 2v + 1$ is defined through the input-output relation:

$$y_i = med(x_{i-v},..,x_i,..,x_{i+v)} \qquad i \in Z \tag{3}$$

The input is the sequence $x_i$ where i ∈ Z, and the output is the sequence $y_i$ where i ∈ Z. The running median, sometimes known as the moving median (Kowalski & Smyk, 2018), is important in data processing.

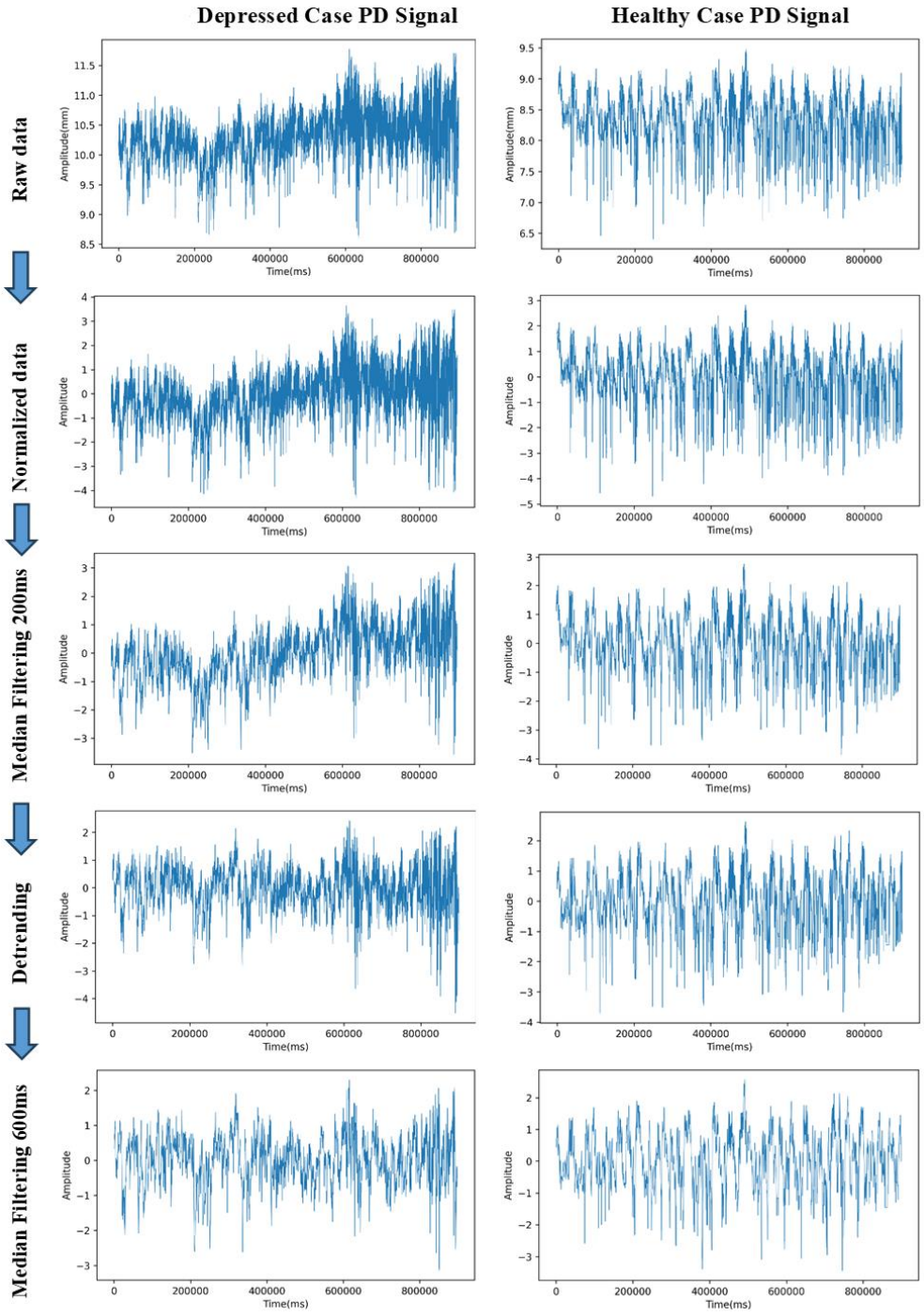


**Fig. 2. Preprocessing steps implementation to PD signal of depressed and healthy individuals**

### 2.3.3. Nonlinear detrending (Sixth-Degree Polynomial)

A sixth-degree polynomial detrending method was then used to remove trends from the signals obtained from the median filter 200ms step. The signal's fluctuations over time must be detrended to concentrate analysis and filtering on the fluctuations of interest. The sixth-degree polynomial detrending prepares the data for subsequent processing, addressing trends before the median filter 600ms is used. Nonlinear detrending was performed using “polyfit” and “polyval” functions. The "polyfit" function computes the polynomial coefficients $p(x)$ of degree $n$ that optimally approximate the data in $y$ (Harris et al., 2020).

$$p(x) = p_1 x^n + p_2 x^{n-1} + \cdots + p_n x + p_{n+1} \quad (4)$$

The coefficients in $p$ are arranged in a descending order of powers, and the length of $p$ is equal to $n + 1$. The vector $x$ represents the query points that correspond to the fitted function values in $y$. The outcome $y$ represents the estimated values at the specific sites of interest specified in $x$. The parameter $n$ represents the polynomial fit degree, which determines the power of the coefficient on the leftmost side of $p$. Subsequently, the “polyval” is used to calculate the value of the polynomial $p$ at each point in $x$. The resulting polynomial curve is then subtracted from the original signal to eliminate trends (Harris et al., 2020).

### 2.3.4. Median filtering (600ms)

In the last step, a 600ms median filter was applied to eliminate sudden PD drops.

## 2.4. Signal processing

After a normalized time-based signal, free from both trend and noise, is obtained for each participant's PD data, it is then subjected to the following signal processing:

HHT, a mixture of EMD and Hilbert Spectral Analysis (HSA), has been recommended recently in this field’s contributions (Shen et al., 2005). Which efficiently obtains information in the time and frequency domains from the data directly. In HHT, EMD extracts characteristic scales from the data to breakdown the signal into oscillation modes. Through IMF component representation, EMD may be used for time–frequency filtering. Fig. 3 shows the two proposed steps for signal processing (IMF and HHS). The first raw includes one depressed case’s PD signal and one healthy case’s PD signal after applying EMD to produce many IMFs, while the second raw includes the HHS of one of the extracted IMFs for each case.

### 2.4.1. Empirical Mode Decomposition (EMD)

EMD is a data-driven signal-decomposing approach without a priori basis functions (Huang et al., 1998; Junsheng et al., 2006). Decomposing the signal into IMFs is the EMD's goal. An IMF function meets two conditions: 1) The number of extrema and the number of zero crossings must be equal or differ by one across the data set; and 2) The mean value of the local maxima and minima envelops must be zero at any point. A Fourier analysis simple harmonic function is compared to an IMF, which reflects the oscillatory mode in the data (Anas et al., 2010).

Given the detrended and filtered signal $x(t)$, the starting point of EMD involves identifying all the local maxima and minima. The local maxima are connected by a cubic spline curve to form the upper envelope (Gregory, 1985), while the local minima are similarly connected to form the lower envelope.

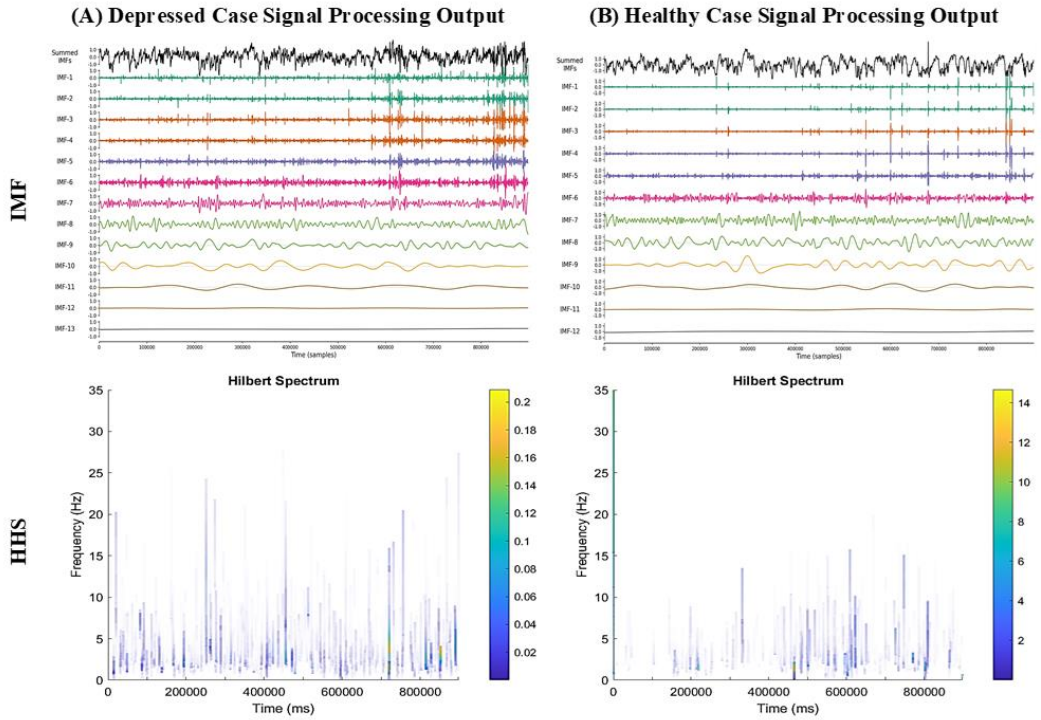


**Fig. 3. Processing steps implementation to pre-processed PD signals of depressed and healthy individuals**

The mean of these two envelopes is then calculated and subtracted from the original signal to obtain the first component $h_1(t)$. This process of extracting the IMF is known as the sifting process (Huang & Shen, 2005). Ideally, $h_1(t)$ should be an IMF, as its construction aims to meet all IMF requirements. However, since $h_1(t)$ still contains multiple extrema between zero crossings, the sifting process is repeated on $h_1(t)$. After repeated sifting, if the resulting difference satisfies the IMF properties, it is considered the first IMF component, denoted as $c_1(t)$. A common criterion to determine the sufficiency of the sifting process is the standard deviation (SD) between two consecutive siftings. When the SD falls below a certain threshold, the first IMF $c_1(t)$ is obtained. This first IMF is then separated from the remaining data, producing a residue $r_1(t)$. It is important to note that this residue still contains valuable information. Therefore, the residue is treated as a new signal, and the sifting process is applied again to extract additional IMFs. This iterative process continues, with each new residue being used to extract the next IMF.

The whole procedure terminates when either the component $c_q(t)$ or the residue $r_q(t)$ becomes very small or when the residue $r_q(t)$ becomes a monotonic function. So, the EMD of the original signal is given by equation:

$$x(t) = \sum_{i=1}^{q} c_i(t) + r_q(t) \quad (5)$$

where: $c_1(t)$, $c_2(t)$,…, $c_q(t)$ – are all the IMFs included in the signal.
$r_q(t)$ – is a negligible residue (Huang & Attoh-Okine, 2005).

The decomposition yields $q$-intrinsic modes and a residue. In the presented model, the EMD method extracts a variable number of IMF components based on each signal intensity. For both the healthy and depressed groups, the first ten IMF components are utilized for analysis (From IMF-1 to IMF-10). So, PD signal is decomposed into ($q$) IMF components with various time scales. Since the first component has the lowest time scale, it has the quickest signal time fluctuation. The mode mean frequency drops as the time scale grows during decomposition. The EMD has been used to show the relative energy distribution of individual IMFs to detrended PD signals, even though it lacks a precise mathematical definition. Thus, Hilbert transform overcomes this problem by generating a complex signal with instantaneous energy, phase, and frequency fluctuations regardless of whether the signal is stationary or nonstationary and produced by linear or nonlinear processes.

After decomposition, we got 10 IMF signals for each of the 29 healthy persons' left and right eye PDs, 29 depressed individuals' left and right eye PDs. Then statistical features will be extracted as follows.

### 2.4.2. IMF statistical features extraction

Statistical features, such as mean, variance, standard deviation, median, maximum, minimum, quartiles (Q1 and Q3), skewness, and kurtosis, are then computed for each set of IMFs. Which might be a better tool than using the huge amounts of data have been collected. After processing all subjects, the individual feature data frames for the right and left eyes are concatenated to create combined data frames for each condition. The features and labels are separated into different data frames, which are then exported to Excel files for further analysis. The exports include separate files for right eye features, left eye features, both eye combined features, and their corresponding labels.

### 2.4.3. Hilbert Huang Transform (HHT)

For individual IMFs, the Hilbert Huang transform has been calculated using the Cauchy principal value integral (Huang & Attoh-Okine, 2005). As a result, we may compute the instantaneous frequencies, instantaneous phase, and instantaneous energy magnitude. The HT can show how the power and frequency change over time. Converting the actual signal to its complex analytic version is beneficial since using the real signal will result in cross-terms due to the existence of positive and negative frequencies. Furthermore, it permits the calculation of the signal's phase and magnitude without requiring an estimate that the signal is stationary or the result of a linear process. Additionally, when employing a genuine signal, an analytical signal will provide an accurate description of the average frequency and remove the requirement to sample at double the Nyquist rate. The so-called Hilbert spectrum (HS) is obtained if the instantaneous frequency and the temporal variation of energy (envelope) are connected at the same moment.

### 2.4.4. HS statistical features extraction

Similarly to the approach used in extracting statistical features from IMF in the IMF statistical features extraction method, after all Hilbert Spectra (HS) have been obtained for each individual IMF component of each preprocessed PD signal, Statistical features, including mean, variance, standard deviation, median, maximum, minimum, quartiles (Q1 and Q3), skewness, and kurtosis, were obtained from the power spectrum of the detrended PD signal. These features represent the energy for the following frequency bands: ultra-low frequency ([ULF]: 0–1 Hz), very low frequency ([VLF]: 1–3 Hz), low frequency ([LF]: 3–6 Hz), and high frequency ([HF]: 6–10 Hz). Again, the individual feature data frames for the right and left eyes are concatenated to create combined data frames for each condition. The features and labels are separated into different data frames, which are then exported to Excel files for further analysis. The exports include separate files for right eye features, left eye features, both eye combined features, and their corresponding labels.

## 2.5. K-Fold Cross-Validation

K-fold Cross-Validation divides the extracted statistical features of both IMF and HS into $k$ equal-sized folds. After that, $k$ training and validation iterations are performed, with each iteration holding out a different data fold for validation and using the remaining $k-1$ folds for learning (Lendasse et al., 2003). Stratifying data before splitting into $k$ folds is frequent. Stratification reorganizes data to make each fold indicative of the whole. This method's sequence is:

1. Split the features extracted data sets $X$ of $n$ into $K$ nearly equal-sized sets. The validation set $X_{val}$ contains the $k$th set. Other sets constitute $X_{learn}$, a learning dataset. (The model was evaluated for $K$ one- to ten-folds.)
2. $X_{learn}$ used to train model g, and the error $E_k(\mathrm{g})$ is determined as:

$$xE_k(\mathrm{g}) = \frac{\sum_{i=1}^{\frac{n}{K}} (\mathrm{g}(x_i^{val} - y_i^{val})^2}{\frac{n}{K}} \tag{6}$$

where: $x_i^{val}$, $y_i^{val}$ – are the elements of $X_{val}$.
$\mathrm{g}(x_i^{val})$ – the approximation of $y_i^{val}$ by model g.

Steps 1 and 2 are repeated for $K$ varying from 1 to $K$. The average error $\hat{E}_k(\mathrm{g})$ is calculated according to:

$$\hat{E}_k(\mathrm{g}) = \frac{\sum_{k=1}^{K} E_k(\mathrm{g})}{K} \tag{7}$$

## 2.6. Classification

Many classification and clustering methods can predict initial psychiatric disorders (Kamel & Selim, 1994; Mao & Jain, 1996; Selim & Alsultan, 1991; Yu & Yuan, 1995). They can also help psychiatry decide on medication and treatment cycles. Discriminant functions like KNN (Benvenuto et al., 2002), and SVM (Zhang et al., 2011) can be used to classify psychiatric disorders. In this research, six different classification methods and machine learning algorithms were applied: SVM, Decision Tree, KNN, Bagging, Gradient

Boosting, and Random Forest have been applied on the first ten IMFs extracted statistical features.

### 2.7. System evaluation

The Confusion Matrix (CM) is often used in system assessment to assess the effectiveness of a proposed machine learning model. It includes many parameters that effectively measure the model's performance. The values for these parameters are determined by The Equations shown below (Sokolova & Lapalme, 2009):

$$Accuracy = \frac{Tp+Tn}{Tp+Fp+Tn+Fn} \times 100 \quad (8)$$

$$Precision = \frac{Tp}{Tp+Fp} \times 100 \quad (9)$$

$$Recall\ (Sensitivity) = \frac{Tp}{Tp+Fn} \times 100 \quad (10)$$

$$Specificity = \frac{Tn}{Tn+Fp} \times 100 \quad (11)$$

$$F_1 - Score = \frac{2\,Tp}{2Tp+Fp+Fn} \quad (12)$$

where: "True Positive" ($Tp$) – case count when model predicted positive class.
"False Positive" ($Fp$) – shows model mispredictions of positive class.
"True Negative" ($Tn$) – case count when model predicted negative class.
"False Negative" ($Fn$) – shows model mispredictions of negative class.

The classification was performed by classifying depressed individuals as "positive" and healthy individuals as "negative". The data analysis will focus on accuracy, precision, specificity, sensitivity (recall), and F1-score. The accuracy represents the model's ability to differentiate between depressed and normal people. The precision determines the model's ability to correctly detect relevant cases without misclassifying too many irrelevant cases as positive; the sensitivity determines the ratio of depressed patients correctly classified; and the F1-score considers the model's recall and precision.

## 3. RESULTS AND DISCUSSION

All results have been analyzed using Power BI to get the best model for classification between healthy and depressed cases. Features may be classified with respect to various perspectives. The conducted work was built upon the following selected categorization.

### 3.1. IMF features

The extracted IMF's statistical features contribute significantly to achieving high classification accuracy results. Based on the following insights, it is evident that the proposed model can effortlessly achieve high-accuracy results, where:

– “IMF-6” obtained the highest classification accuracy, achieving 93.8% accuracy, 88.9% precision, 87.5% specificity, 100% sensitivity, and 94.1% F1-score. That was achieved by applying sevenfold cross-validation by utilizing the statistical features of the combined Both Pupils signal and employing Bagging as a classification method.
– As shown in Fig. 4, "IMF-1" emerged 113 out of 773 times to produce high classification accuracy results over 80% across all CV folds, demonstrating the power of IMF-1 as a signal that can enhance the proposed model's classification accuracy.
– Figure 5 shows that “Random Forest” stands out with 193 out of 773 contribution times in achieving high classification accuracy results above 80% compared with other classifiers across all CV folds.
– Figure 6 shows that about half of the pupil data contributions resulted in high classification accuracy over 80% achieved by using the combined "Both Pupil" data. As a result, it is evident that the inclusion of "Both Pupil" data is critical for improving the model's accuracy.

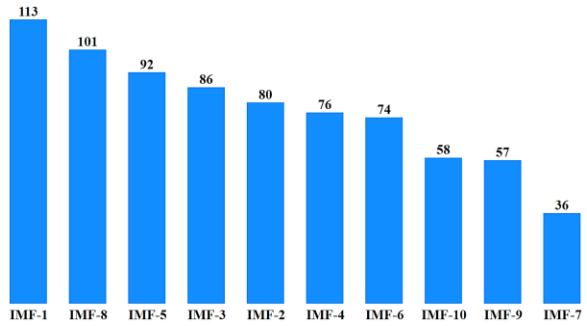


**Fig. 4. IMFs’ contribution number of times for achieving classification accuracy above 80% by using IMF features**

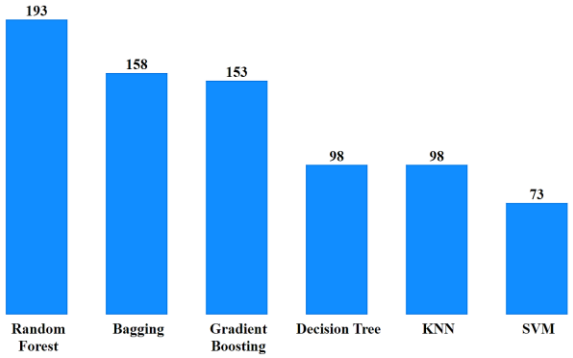


**Fig. 5. Classifiers’ contribution number of times for achieving classification accuracy above 80% by using IMF features**

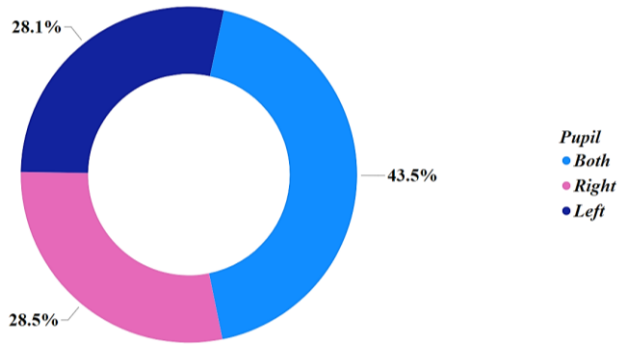


**Fig. 6. Pupil diameter data contribution ratios in achieving classification accuracy above 80% by using IMF features**

Following the analysis of the previous results, the authors optimized the model by utilizing the most powerful factors that contributed to achieving high accuracy results. This was achieved by applying the statistical features extracted from the IMF-1 of the combined Both Pupil data and using Random Forest as a classification method. Using Four-folds cross-validation, the model was able to achieve 93.8% accuracy, 88.9% precision, 87.5% specificity, 100% sensitivity, and 94.1% F1-score.

### 3.2. Hilbert–Huang Spectrum Features

Hilbert–Huang features produced the highest success rate and were better than the use of IMF features in classification between depressed and healthy cases, where CV and subsequent classification by means of the same six classifiers have been applied to all Hilbert–Huang statistical features that have been extracted from the different frequency bands of HS of the first ten IMFs. The main observations and results are summarized below:

- The LF band statistical features of the HS of "IMF-1" obtained the highest classification accuracy result, where it was able to achieve 93.3% accuracy, 87.5% precision, 87.5% specificity, 100% sensitivity, and a 93.3% F1-score. That was achieved by applying Eight-fold CV, utilizing the statistical features of Both Pupil's signals, and employing Bagging as a classification method.
- "IMF-1" emerged 441 out of 1372 times to produce high classification accuracy results exceeding 80% across all CV folds, as shown in Fig. 7.
- Figure 8 shows that "Random Forest" stands out with 279 out of 1372 contribution times (20.3%) in achieving high classification accuracy results above 80% compared with other classifiers across all CV folds.
- "Both Pupil" data significantly contribute 41.3% to achieving high accuracy results, as shown in Fig. 9.
- Configurations with the "ULF" band were able to achieve high accuracy more frequently than the other bands, so it became the most frequent band achieving high classification accuracy results regardless of the classifier used, as shown in Fig. 10,11.

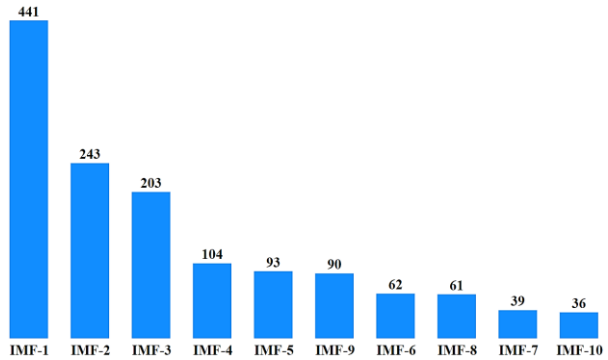


**Fig. 7. IMFs' contribution number of times for achieving classification accuracy above 80% by using HS features across all frequency bands**

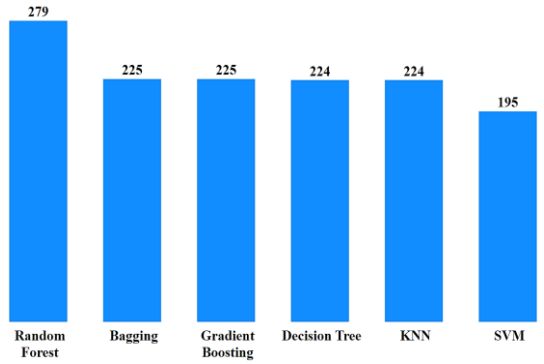


**Fig. 8. Classifiers contribution number of times for achieving classification accuracy above 80% by using HS features across all frequency bands**

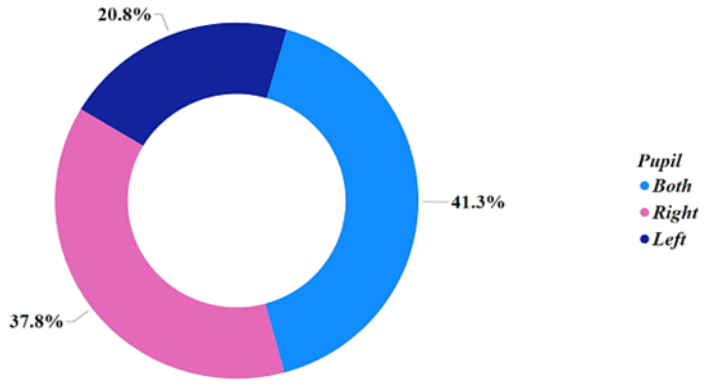


**Fig. 9. Pupil diameter data contribution ratios in achieving classification accuracy above 80% by using HS features across all frequency bands**

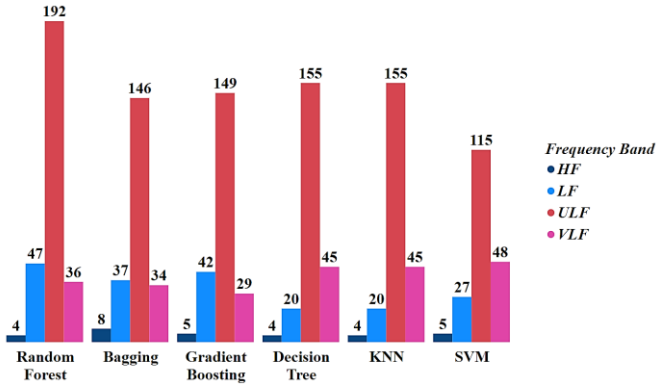


**Fig. 10. A comparative analysis of all classifiers and its most frequent Frequency Bands utilized in attaining above 80% accuracy by using HS features**


**Fig. 11. Frequency bands contribution number of times for achieving classification accuracy above 80% by using HS features across all classifiers**

### 3.3. Summary of results

The research compared the classification accuracy of IMF features and each frequency band of HS features for distinguishing between healthy and depressed cases. Here are the key findings for each technique used:

IMF Features:

IMF-6 achieved the highest classification accuracy of 93.75% using Bagging as a classifier and statistical features from the Both Pupils signal. Despite the fact that IMF-6 achieved the highest classification accuracy, IMF-1 consistently produced high accuracy across multiple cross-validation folds, emphasizing its importance in enhancing classification accuracy. Also, despite Bagging achieving the highest classification accuracy, Random Forest was notably effective, contributing significantly to achieving high accuracy results in various scenarios. The inclusion of Both Pupil data was significant for enhancing model accuracy by leveraging combined pupil signals.

Hilbert–Huang Spectrum Features:

LF band features of HS from IMF-1 achieved the highest accuracy of 93.33%, employing Bagging with Both Pupil's signals. Similar to IMF features, despite Bagging achieving the highest classification accuracy, Random Forest demonstrated strong performance in achieving high accuracy. ULF band configurations were particularly successful, achieving the highest accuracy results more frequently across different classifiers.

The following Tab. 1 presents the highest attained results across all signals that had been processed, where different classifiers were applied to PD signals for the classification of depressed and healthy individuals.

**Tab. 1. The highest attained results obtained from our proposed approach**

| | Classifier | IMF | PD Signal | CV | Accuracy | Precision | Specificity | Sensitivity | F1-Score |
|---|---|---|---|---|---|---|---|---|---|
| IMF | Bagging | IMF-6 | Both | 7-Fold | 93.75% | 88.89% | 87.50% | 100.00% | 94.12% |
| HS – ULF band | Bagging | IMF-9 | Both | 6-Fold | 92.86% | 100.00% | 100.00% | 85.71% | 92.31% |
| HS – VLF band | Decision Tree | IMF-3 | Both | 9-Fold | 91.67% | 100.00% | 100.00% | 83.33% | 90.91% |
| HS – LF band | Bagging | IMF-1 | Both | 8-Fold | 93.33% | 87.50% | 87.50% | 100.00% | 93.33% |
| HS – HF band | Gradient Boosting | IMF-1 | Left | 7-Fold | 87.50% | 100.00% | 100.00% | 75.00% | 85.71% |

## 3.4. Comparison with the authors' previous research results

The following comparative analysis shown in Tab. 2 and Fig. 12 between the authors' recent proposed Machine Learning model and their previous contribution in the same field utilizing the same modality of Pupil Diameter signals using Deep Learning model (Ismail et al., 2024) reveals that the combined data from both pupils significantly enhances classification accuracy in distinguishing between depressed and healthy individuals. Specifically:

The accuracy of 90.91% from the Bagging method in the left eye significantly surpasses the 78.57% accuracy of AlexNet. This indicates that machine learning provides a more effective classification for the left eye dataset.

Similarly, the accuracy of 90.91% from machine learning approaches outperforms GoogLeNet's 71.43% accuracy for the right eye. This further highlights the superior performance of machine learning methods in handling the right eye data.

In comparing the accuracy of machine learning and deep learning, the machine learning approach with IMF-6 achieved the highest classification accuracy of 93.75% by using Bagging as a classifier and statistical features from both eyes' pupils. In contrast, the Deep Learning method utilizing AlexNet reached an accuracy of 98.28% by combining data from both eyes' pupils, significantly surpassing the machine learning results.

In conclusion, while both machine learning and deep learning methods demonstrate strong accuracy, machine learning is typically preferred for its faster execution times.

**Tab. 2. Comparative analysis between Deep Learning and Machine Learning classification accuracy results**

| Method | Deep Learning | | | Machine Learning | | | |
|---|---|---|---|---|---|---|---|
| PD Dataset | Left | Right | Combined | Left | Right | | Combined |
| Model | AlexNet | GoogLeNet | AlexNet | Bagging, Gradient Boosting, or Random Forest | SVM | Gradient Boosting, or Random Forest | Bagging |
| Accuracy | 78.57% | 71.43% | 98.28% | 90.91% (IMF) | 90.91% (ULF, VLF) | 90.91% (LF) | 93.75% (IMF) |

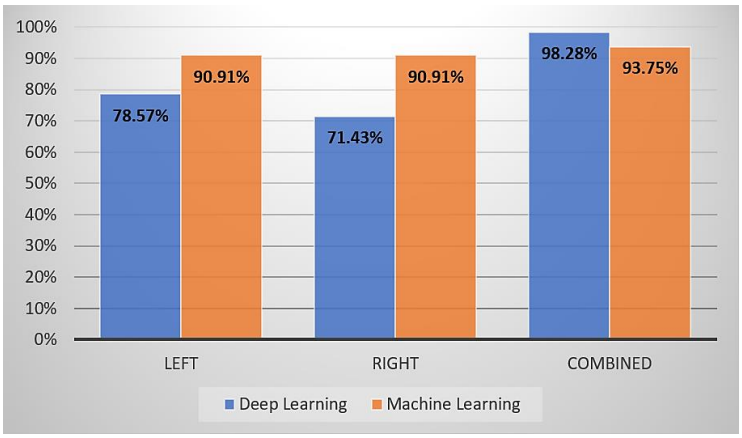


**Fig. 12. Comparative analysis between Deep Learning and Machine Learning classification accuracy results**

### 3.5. Comparison with other researchers results

The following Tab. 3 presents a comparative analysis of the highest classification average accuracy results for our model against those achieved by other researchers over the past five years in the field of depression recognition. The results indicate that our proposed classification method outperforms other methods using a unimodal approach with only 3-fold CV which uses a lower processing time.

**Tab. 3. Comparative analysis of different depression recognition methods**

| Reference | Modality | Classifier | CV | Accuracy | Precision | Specificity | Sensitivity | F1-Score |
|---|---|---|---|---|---|---|---|---|
| Ours | Pupil Diameter | Random Forest | 3-Fold | 77.72% | 83.33% | 82.76% | 71.85% | 76.09% |
| Yang et al., 2023 | Pupil Diameter | SVM | *NI* | 64.78% | *NI* | 67.16% | 63.88% | 63% |
| Zhu et al., 2023 | Pupil Area | SVM | 10-Fold | 72.1% | *NI* | *NI* | *NI* | *NI* |
| Shen et al., 2021 | Eye movement | SVM | 10-Fold | 77% | *NI* | 70.3% | 69% | *NI* |
| Li et al., 2020 | Eye movement | KELM | 10-Fold | 76.84% | *NI* | 79.34% | 75.95% | 76.66% |
| Zeng et al., 2019 | Eye movement | SVM | 10-Fold | 76.04% | *NI* | *NI* | *NI* | *NI* |

*NI*: Refers that the result for that item was "Not Included".

## 4. CONCLUSION

The conducted research demonstrates that employing advanced machine learning classifiers on pupil diameter signals significantly enhances the accuracy of depression recognition systems. Among the various classifiers tested, the Random Forest and Bagging classifiers consistently produced the highest performance metrics across different IMFs and frequency bands of the HS.

The results of this work indicate that the use of the extracted statistical features from the IMFs of the individual PD signals, in conjunction with the easy and fast processing machine

learning model proposed by the authors, can significantly boost the classification accuracy of depression recognition systems. This is further corroborated by the comparative analysis with other researchers' results over the past five years, which highlights the superior performance of our proposed model, especially in terms of accuracy, precision, and specificity.

Overall, this study underscores the potential of leveraging machine learning and signal processing techniques in analyzing pupil diameter signals for the effective classification of depressed and healthy individuals. The high classification accuracy achieved using our approach suggests that it could be a valuable tool in clinical settings for early and accurate depression detection. In the following points you can find the advantages of our proposed model:

1. It attained higher classification accuracy results compared with the other unimodal approaches.
2. Its CV outperforms other methods using only 3-fold which costs less processing time.
3. It is a cost-effective approach where utilizing only pupil diameter readings which can be measured with relatively low-cost equipment, makes the model practical and accessible for widespread clinical use, unlike other utilized data like EEG or FMRI.
4. The model's reliance on well-known machine learning algorithms and statistical feature extraction techniques makes it straightforward to implement and deploy. This ease of use can facilitate quicker adoption in clinical and research settings.
5. In recent years, the eye tracking and measuring of PD have become more and more easily accessible than ever. In the future, with the development of cameras and sensors in mobile phones, we can even make an app for depression detection using PD measurement; it can even be used directly and give a diagnosis very fast with high accuracy.

So, by combining these advantages, the proposed model not only sets a new benchmark in the field of depression recognition but also offers practical benefits that enhance its viability for real-world applications.

## 5. FUTURE WORK

In the future, the authors will primarily focus on enhancing the depression detection algorithm and employing alternative experiments not only to leverage the ability of PD data to classify between depressed and healthy individuals but also to measure depression severity. This will result in more dependable classification accuracy for real-time applications, thereby establishing eye tracking as a more convenient, cost-effective, and widely adopted method for detecting depression.


## Funding

*The authors declare that they have no competing interests.*


## Conflicts of Interest

*The authors declare that this study was not funded.*

## Acknowledgements

*The authors express their sincere gratitude to Prof. Karl Jürgen Bär and Prof. Andy Schumann (Jena University Hospital, Department of Psychiatry and Psychotherapy, Jena, Germany) for sharing the dataset which were used in this study and their helpful support.*

*Also, authors would like to thank (Biomedical Engineering Department, Faculty of Engineering, Misr University for Science and Technology, Egypt) for any advice or discussion that improved the study, as well as providing guidance throughout this work.*

### REFERENCES

Aboamer, M. A., Azar, A. T., Mohamed, A. S. A., Bär, K.-J., Berger, S., & Wahba, K. (2014). Nonlinear features of heart rate variability in paranoid schizophrenic. *Neural Computing and Applications*, *25*(7), 1535-1555. https://doi.org/10.1007/s00521-014-1621-1

Anas, E. M. A., Lee, S. Y., & Hasan, M. K. (2010). Sequential algorithm for life threatening cardiac pathologies detection based on mean signal strength and EMD functions. *BioMedical Engineering OnLine*, *9*, 43. https://doi.org/10.1186/1475-925X-9-43

Benvenuto, J., Jin, Y., Casale, M., Lynch, G., & Granger, R. (2002). Identification of diagnostic evoked response potential segments in Alzheimer's disease. *Experimental Neurology*, *176*(2), 269-276. https://doi.org/10.1006/exnr.2002.7930

Ding, X., Yue, X., Zheng, R., Bi, C., Li, D., & Yao, G. (2019). Classifying major depression patients and healthy controls using EEG, eye tracking and galvanic skin response data. *Journal of Affective Disorders*, *251*, 156-161. https://doi.org/10.1016/j.jad.2019.03.058

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., Schatzberg, A. F., Sudheimer, K., Keller, J., Mayberg, H. S., Gunning, F. M., Alexopoulos, G. S., Fox, M. D., Pascual-Leone, A., Voss, H. U., … Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*, 28-38. https://doi.org/10.1038/nm.4246

Gregory, J. A. (1985). Shape Preserving Spline Interpolation. *NASA. Langley Research Center Computational Geometry and Computer-Aided Design.*

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357-362. https://doi.org/10.1038/s41586-020-2649-2

Huang, N. E., & Attoh-Okine, N. O. (Eds.). (2005). *The Hilbert-Huang Transform in Engineering*. CRC Press. https://doi.org/10.1201/9781420027532

Huang, N. E., & Shen, S. S. P. (2005). *Hilbert-huang Transform And Its Applications*. World Scientific. https://doi.org/10.1142/5862

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *454*(1971), 903-995. https://doi.org/10.1098/rspa.1998.0193

Ismail, I., El-Wakad, M. T., Shafie, K. A., Aboamer, M. A., & Mohamed, N. A. R. (2024). Major depressive disorder: Early detection using deep learning and pupil diameter. *Indonesian Journal of Electrical Engineering and Computer Science*, *35*(2), 916-932. https://doi.org/10.11591/ijeecs.v35.i2.pp916-932

Jones, N. P., Siegle, G. J., & Mandell, D. (2015). Motivational and emotional influences on cognitive control in depression: A pupillometry study. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 263-275. https://doi.org/10.3758/s13415-014-0323-6

Junsheng, C., Dejie, Y., & Yu, Y. (2006). Research on the intrinsic mode function (IMF) criterion in EMD method. *Mechanical Systems and Signal Processing*, *20*(4), 817-824. https://doi.org/10.1016/j.ymssp.2005.09.011

Kamel, M. S., & Selim, S. Z. (1994). A relaxation approach to the fuzzy clustering problem. *Fuzzy Sets and Systems*, *61*(2), 177-188. https://doi.org/10.1016/0165-0114(94)90232-1

Kowalski, P., & Smyk, R. (2018). Review and comparison of smoothing algorithms for one-dimensional data noise reduction. *2018 International Interdisciplinary PhD Workshop (IIPhDW)* (pp. 277-281). IEEE. https://doi.org/10.1109/IIPHDW.2018.8388373

Kramarić, K., Šapina, M., Garcin, M., Milas, K., Pirić, M., Brdarić, D., Lukić, G., Milas, V., & Pušeljić, S. (2019). Heart rate asymmetry as a new marker for neonatal stress. *Biomedical Signal Processing and Control*, *47*, 219-223. https://doi.org/10.1016/j.bspc.2018.08.027

Lendasse, A., Wertz, V., & Verleysen, M. (2003). Model selection with cross-validations and bootstraps-application to time series prediction with RBFN models. In O. Kaynak, E. Alpaydin, E. Oja, & L. Xu (Eds.), *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003* (pp. 573-580). Springer. https://doi.org/10.1007/3-540-44989-2_68

Li, M., Cao, L., Zhai, Q., Li, P., Liu, S., Li, R., Feng, L., Wang, G., Hu, B., & Lu, S. (2020). Method of depression classification based on behavioral and physiological signals of eye movement. *Wiley Online Library*, *2020*(1), 4174857. https://doi.org/10.1155/2020/4174857

Li, W., Ma, H., Wang, X., & Shi, D. (2014). *Features Derived from Behavioral Experiments to Distinguish Mental Healthy People from Depressed People*. Biomedical Engineering / 817: Robotics Applications. https://doi.org/10.2316/P.2014.818-021

Mao, J., & Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (HEC). *IEEE Transactions on Neural Networks*, *7*(1), 16-29. https://doi.org/10.1109/72.478389

Newson, J. J., & Thiagarajan, T. C. (2019). EEG frequency bands in psychiatric disorders: A review of resting state studies. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00521

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., & Galatzer-Levy, I. R. (2022). Deep Learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine*, *52*(5), 957-967. https://doi.org/10.1017/S0033291720002718

Schumann, A., Kralisch, C., & Bär, K.-J. (2015). Spectral decomposition of pupillary unrest using wavelet entropy. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6154-6157). IEEE. https://doi.org/10.1109/EMBC.2015.7319797

Selim, S. Z., & Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, *24*(10), 1003-1008. https://doi.org/10.1016/0031-3203(91)90097-O

Shen, R., Zhan, Q., Wang, Y., & Ma, H. (2021). Depression detection by analysing eye movements on emotional images. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7973-7977). IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414663

Shen, S. S. P., Shu, T., Huang, N. E., Wu, Z., North, G. R., Karl, T. R., & Easterling, D. R. (2005). Hht analysis of the nonlinear and non-stationary annual cycle of daily surface air temperature data. *Hilbert-Huang Transform and Its Applications*, *5*, 187-209. https://doi.org/10.1142/9789812703347_0009

Siegle, G. J., Granholm, E., Ingram, R. E., & Matt, G. E. (2001). Pupillary and reaction time measures of sustained processing of negative information in depression. *Biological Psychiatry*, *49*(7), 624-636. https://doi.org/10.1016/S0006-3223(00)01024-6

Siegle, G. J., Steinhauer, S. R., Friedman, E. S., Thompson, W. S., & Thase, M. E. (2011). Remission prognosis for cognitive therapy for recurrent depression using the pupil: Utility and neural correlates. *Biological Psychiatry*, *69*(8), 726-733. https://doi.org/10.1016/j.biopsych.2010.12.041

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage*, *20*(1), 114-124. https://doi.org/10.1016/S1053-8119(03)00298-2

Skowron, K., Budzyńska, A., Wiktorczyk-Kapischke, N., Chomacka, K., Grudlewska-Buda, K., Wilk, M., Wałecka-Zacharska, E., Andrzejewska, M., & Gospodarek-Komkowska, E. (2022). The role of psychobiotics in supporting the treatment of disturbances in the functioning of the nervous system - A systematic review. *International Journal of Molecular Sciences*, *23*(14), 7820. https://doi.org/10.3390/ijms23147820

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Suslow, T., Hußlack, A., Kersting, A., & Bodenschatz, C. M. (2020). Attentional biases to emotional information in clinical depression: A systematic and meta-analytic review of eye tracking findings. *Journal of Affective Disorders*, *274*, 632-642. https://doi.org/10.1016/j.jad.2020.05.140

Wang, J., Fan, Y., Zhao, X., & Chen, N. (2014). Pupillometry in chinese female patients with depression: A pilot study. *International Journal of Environmental Research and Public Health*, *11*(2), 2236-2243. https://doi.org/10.3390/ijerph110202236

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. https://www.who.int/publications/i/item/9789240049338

Yang, M., Wu, Y., Tao, Y., Hu, X., & Hu, B. (2023). Trial selection tensor canonical correlation analysis (TSTCCA) for depression recognition with facial expression and pupil diameter. *IEEE Journal of Biomedical and Health Informatics*, 1-12. https://doi.org/10.1109/JBHI.2023.3322271

Yu, B., & Yuan, B. (1995). A global optimum clustering algorithm. *Engineering Applications of Artificial Intelligence*, *8*(2), 223-227. https://doi.org/10.1016/0952-1976(94)00067-W

Zeng, S., Niu, J., Zhu, J., & Li, X. (2019). A study on depression detection using eye tracking. In Y. Tang, Q. Zu, & J. G. Rodríguez García (Eds.), *Human Centered Computing* (pp. 516–523). Springer International Publishing. https://doi.org/10.1007/978-3-030-15127-0_52

Zhang, W.-R., Pandurangi, A. K., Peace, K. E., Zhang, Y.-Q., & Zhao, Z. (2011). MentalSquares: A generic bipolar Support Vector Machine for psychiatric disorder classification, diagnostic analysis and neurobiological data mining. *International Journal of Data Mining and Bioinformatics*, *5*(5), 532-557. https://doi.org/10.1504/IJDMB.2011.043034

Zhao, L.-M., Li, R., Zheng, W.-L., & Lu, B.-L. (2019). Classification of five emotions from EEG and eye movement signals: Complementary representation properties. *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 611-614). IEEE. https://doi.org/10.1109/NER.2019.8717055

Zhong, J., Wang, D., Wu, H., Wang, P., Yang, M., Peng, H., & Hu, B. (2022). Filterable sample consensus based on angle variance for pupil segmentation. *Digital Signal Processing*, *130*, 103695. https://doi.org/10.1016/j.dsp.2022.103695

Zhu, C., Li, B., Li, A., & Zhu, T. (2016). Predicting depression from internet behaviors by time-frequency features. *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 383-390). IEEE. https://doi.org/10.1109/WI.2016.0060

Zhu, J., Wang, Z., Gong, T., Zeng, S., Li, X., & Hu, B. (2020). An improved classification model for depression detection using EEG and eye tracking data. *IEEE Transactions on NanoBioscience*, *19*(3), 527-537. https://doi.org/10.1109/TNB.2020.2990690

Zhu, J., Yang, C., Xie, X., Wei, S., Li, Y., Li, X., & Hu, B. (2023). Mutual information based fusion model (MIBFM): Mild depression recognition using EEG and pupil area signals. *IEEE Transactions on Affective Computing*, *14*(3), 2102–2115. https://doi.org/10.1109/TAFFC.2022.3171782