JOURNAL OF
EMERGING INVESTIGATORS

Article

A HOG feature extraction and CNN approach to Parkinson's spiral drawing diagnosis

Ayush Tripathi[1], Priyanka Mishra[2]

[1] Edison Academy Magnet School, Edison, New Jersey

[2] Systems Engineering & Integration, Illumina, San Diego, California

SUMMARY

Parkinson's disease (PD) is the second most widespread neurodegenerative disorder in the United States following Alzheimer's disease. Its inefficient and often inaccessible diagnostic process relies heavily on a healthcare professional's analysis of a patient's performance on neurological and physical exams. We aim to use machine and deep learning to improve the efficiency, accessibility, and accuracy of the diagnosis process by utilizing the effect that rhythmic shaking in the hand, a common PD symptom, has on a patient's ability to draw a spiral. We used images of spirals drawn by both PD and non-PD individuals to train a deep learning model that relied strictly on computer vision and a support vector machine (SVM)-based machine learning model that utilized histogram of oriented gradients feature extraction. We hypothesized that both models would outperform the current clinical diagnosis process by reducing misdiagnosis rates. This hypothesis was unsupported by our study's findings. The models are unable to distinguish between PD and other neurocognitive disorders like multiple sclerosis or essential tremor. The models, therefore, cannot be directly compared to the current clinical diagnosis process. Instead, both models, in general, can be used as an efficient and highly accessible baseline diagnosis tool for neurocognitive disorders, complementing, rather than replacing, clinical diagnosis methods to improve the accessibility and efficiency of an accurate neurocognitive diagnosis.

INTRODUCTION

Parkinson's Disease (PD), only following Alzheimer's disease, is the second most widespread neurodegenerative disorder in the United States (1). Approximately 500,000 patients in the US have been diagnosed with PD, but accounting for individuals that have either been undiagnosed or misdiagnosed, this number is estimated to be over 1 million (1). By the year 2040, as the US population ages, the number of patients diagnosed with PD is expected to double (1).

Parkinson's is a progressive disorder that weakens the nerve cells in a portion of the brain called the substantia nigra, limiting its ability to produce the neurotransmitter dopamine (2). Dopamine helps transfer messages within the brain to ensure muscles produce steady, efficient movements (1). Insufficient dopamine results in irregular nerve firing which consequently impairs motor skills (1). Studies have shown that PD patients suffer from a loss in over 80% of their dopamine-producing cells in the substantia nigra (2).

Concerns regarding limitations in the current diagnosis process continue as PD numbers rise. At the moment, there is no conclusive lab or imaging test for the diagnosis of Parkinson's (3). A neurologist reviews a patient's medical history, current symptoms, and performance on both a neurological and physical exam to make a diagnosis (4). A dopamine transporter (DaT) scan and other imaging tests (MRIs, PET scans, etc.) can be used to validate a diagnosis or rule out similar disorders; they cannot, however, be used for the direct diagnosis of PD (3, 4). Carbidopa-levodopa, generally consumed to mitigate PD symptoms, can be used to aid in the diagnosis of Parkinson's as well (4). Unfortunately, a sufficient dose must be given, and a diagnosis can only be made if significant improvement in symptoms is seen (4). The current clinical PD diagnosis process, therefore, lacks objectivity and efficiency (5). In a 2020 poll for the charity Parkinson's UK, for example, consisting of over 2000 PD patients, 26% of participants reported being initially misdiagnosed (~74% accuracy), and a further 21% saw their general provider 3 or more times before being referred to a specialist, testifying to the inefficiency and general lack of accessibility of an accurate PD diagnosis (5).

Common symptoms for those suffering from PD include rhythmic shaking, bradykinesia (slowed movement), muscle rigidity, and posture impairment (6). We primarily focused on rhythmic shaking, commonly referred to as a tremor, and the effect it has on a patient's penmanship as a diagnostic marker of PD. Tremors, tending to be the first motor symptom of Parkinson's and commonly occurring in the hands, can significantly affect a patient's ability to draw a spiral (7).

Machine and deep learning employ algorithms for the identification of patterns in an often extensive amount of data (8). Supervised learning, a subset of machine learning, trains algorithms using data and their corresponding labels (9). The algorithms form relationships between the data and their labels before being tested on unseen data with the goal of predicting the correct label (9). Current healthcare applications of machine and deep learning include precision medicine and radiomics (10). Applications in precision medicine involve the use of modeling techniques to predict optimal treatment methods for individual patients (10). Radiomics is the general extraction of numerical features from medical images, and often involves the identification of cancerous lesions in image data using features not visibly distinguishable (10). We utilized two supervised learning algorithms, support vector machines and neural networks, to aid in the early diagnosis of PD.

The goal of this study was to create two models that could be used to increase accessibility to an accurate and more efficient baseline PD diagnosis to alleviate undiagnosed

https://doi.org/10.59720/24-048

and misdiagnosed PD patient rates. We utilized the effect a tremor has on a patient's ability to draw as a means of diagnosing Parkinson's. We hypothesized that both models would outperform the current clinical PD diagnosis process by reducing the misdiagnosis rate. We used an accuracy of 74% for the approximate accuracy of the current clinical diagnosis process, as per a 2020 poll conducted by the Parkinson's UK (5). Our hypothesis was unsupported, however, because of both models' inability to distinguish between neurocognitive disorders in their diagnosis. Rather than outperform and effectively replace the current clinical PD diagnosis process, our models can work with the process as a baseline diagnosis tool for neurocognitive disorders, thus furthering the efficiency and accessibility of an accurate neurocognitive diagnosis.

RESULTS

We posed to investigate deep and machine learning alternatives for the current clinical Parkinson's diagnosis process through the analysis of spirals drawn by patients. We considered the use of two supervised learning algorithms, support vector machines and convolutional neural networks, to aid in this analysis. The dataset used to train both models contained 102, 256 by 256 pixel images of spirals drawn by 51 individuals with PD and 51 without PD (11). The data originated from a study that investigated the correlation between disease severity and a PD patient's speed and pen-pressure while sketching a spiral (12). The dataset was split into both training and testing sets. The convolutional neural network (CNN)-based model was trained on a 72 image training set (36 images from each class, PD and non-PD) and a 30 image testing set (15 images from each class). An 80-20 (80% training, 20% testing) split was used for the SVM training and testing sets.

Five different evaluation metrics were used to evaluate both models: accuracy, recall, precision, f1-score, and the area under the receiver operating characteristics curve (AUC). Accuracy is the total number of correct predictions (true positives and true negatives) divided by the total number of predictions (13). Recall, in the context of this application, is the proportion of PD diagnoses our model correctly made of the number of PD patients (14). Precision is the proportion of Parkinson's diagnoses made by our model that were correct (14). The macro-average of these two values is the average of the recall and precision for each class (13). The explanations provided above are only for the PD class; a similar process would need to be conducted to find the precision and recall for the non-PD, or "healthy," class.

A well-performing model ideally has both a high recall and precision. Achieving both can be difficult, however, because of the inherent trade-off between the two (14). For example, if the classification threshold (divide between the two classes) were increased, the number of false positives would decrease but the number of false negatives would increase, resulting in a decrease in recall met with a corresponding increase in precision (14). An f1-score factors in both values in order to account for this tradeoff (15).

The last metric used to evaluate both models is the area under the receiver operating characteristic (ROC) curve. An ROC curve plots the true positive rate (TPR), which is simply recall, and the false positive rate (FPR), which is a measure of how many non-PD patients were incorrectly diagnosed with PD, at different classification thresholds (16). As the classification threshold gets lower, more patients will be both correctly (higher TPR) and incorrectly (higher FPR) diagnosed with PD (16). The AUC is the definite integral of the ROC curve from 0 to 1 (16). A model merely guessing between two options will have an AUC of 0.5; anything higher than this value is indicative of the ability to differentiate between the two classes (in this case, PD and non-PD) (17).

The CNN deep learning model had an accuracy, f1-score, and macro-average recall of 0.80, and a macro-average precision of 0.82 (**Table 1**). Its AUC was 0.83 (**Figure 1**). The HOG feature extraction-based model performed stronger in every one of the five metrics. Its macro-average recall, f1-score, and accuracy were all 0.86, while its macro-average precision was 0.85 and its AUC was 0.94 (**Figure 1, Table 2**).

DISCUSSION

We aimed to address the inefficiency, inaccessibility, and misdiagnosis rate of the clinical PD diagnosis procedure using modeling algorithms for the analysis of hand-drawn spirals. The feature extraction-based approach achieved an accuracy of 0.86 and an AUC score of 0.94, acting as a baseline diagnosis tool for neurocognitive disorders.

At a surface level, the feature extraction-based model would improve the 74% clinical misdiagnosis rate by 10-15%. The accuracies of our model and the current clinical diagnosis process, however, were not directly comparable. The model has only been trained to differentiate between drawings made by PD and non-PD patients by utilizing the effect a tremor has on a patient's penmanship; therefore, it has no ability to differentiate between PD and other disorders that cause tremors. Corticobasal syndrome (CBS), for example, is an uncommon, progressive neurodegenerative disorder that can cause its patients to lose direct control over a limb, making tasks that require fine motor skills (e.g. drawing a spiral) more difficult (18). Our model has not been trained to differentiate between drawings made by individuals with PD and CBS; thus, it would likely diagnose a shaky, wavering spiral drawn by a CBS patient with Parkinson's. These misdiagnoses are not accounted for in the 86% accuracy stated above.

We initially hypothesized that our models would outperform, and effectively act as a replacement for the current diagnosis process in terms of misdiagnosis rate. This hypothesis was unsupported by the study's findings because our models' inability to distinguish between disorders that cause tremors made them incomparable to current clinical diagnosis procedures. Instead, the strengths of our models

	Predicted Negative	Predicted Positive
Actual Negative	10	5
Actual Positive	1	14

Table 1: Confusion matrix for CNN-based model. Visual comparison of the predicted vs actual diagnoses for the images in the 30 image testing set used to evaluate the CNN-based model.

https://doi.org/10.59720/24-048

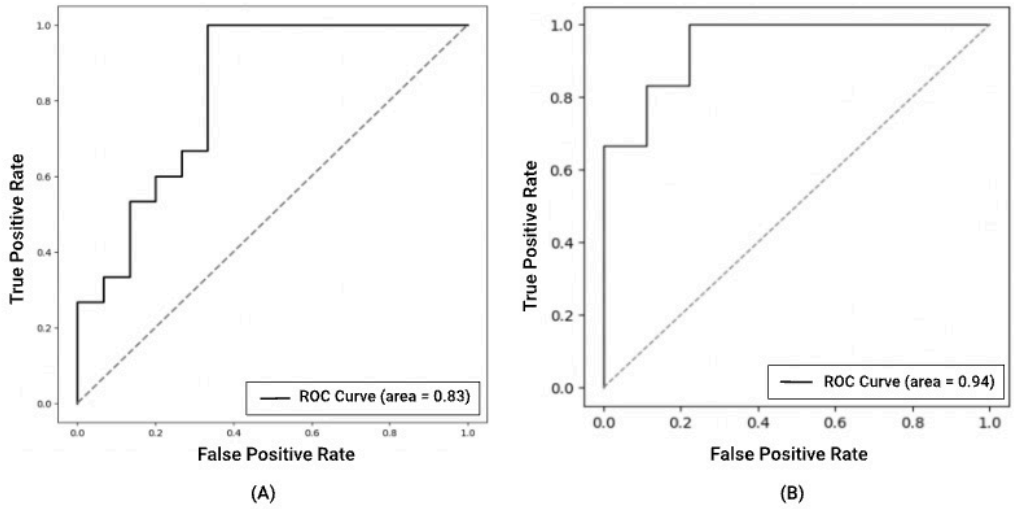

Figure 1: Receiver operating characteristic (ROC) curve for CNN and feature extraction-based model. (A) ROC curve for CNN-based model. (B) ROC curve for feature extraction-based model. ROC curve plots the true positive and false positive rates at different classification thresholds. The true positive rate is the rate at which the model correctly diagnoses a PD patient with PD. The false positive rate is the rate at which the model incorrectly diagnoses a non-PD patient with PD. The definite integral from 0 to 1 shows the area under the curve (AUC) for (A) the CNN-based model (AUC = 0.83) and (B) the feature extraction-based model (AUC=0.94). The dashed line shows the AUC of a model with no ability to differentiate between the two classes (AUC = 0.5).

lie in their ability to provide a baseline diagnosis of whether the patient has a neurocognitive disorder, many of which deteriorate a patient's ability to perform tasks that require fine motor skills. For instance, returning to the example used above, our models can be used to identify that the patient with CBS has a neurocognitive disorder. However, they are not able to pinpoint which specific neurocognitive disorder the patient has. Further examples of neurocognitive disorders that may cause tremors in the hand are multiple sclerosis (MS) and essential tremor (ET) (19).

To improve the models' ability to distinguish between patients with and without a neurocognitive disorder, a more common approach would be training the models with drawings made by individuals with a variety of neurocognitive disorders. This change, however, would likely have a minimal effect on the models' ability to fulfill this goal because the PD, MS, ET, CBS, etc. drawings would look very similar because of their similar symptoms; thus, using drawings from only one disorder (as done in this paper) would not be substantially detrimental to model generalization (18).

The precision, recall, f1-score, AUC, and accuracy characterize both models' ability to act as a baseline diagnosis tool for neurocognitive disorders. The model was marginally stronger at identifying the PD patients (recall) at the expense of the accuracy of the positive diagnoses (precision). The larger emphasis on identifying all PD patients at the small cost of the accuracy of the positive diagnoses is optimal in this context because a false negative implies an undetected disorder, which in its progressive nature, can cause PD symptoms to worsen (20). The f1-score of 0.86, however, factors in both metrics and indicates that the model performed strongly in both categories (15). The metrics, as a whole, illustrate the model's ability to act as a baseline tool for the general diagnosis of neurocognitive disorders but its inability to directly diagnose PD because of the limitations discussed above.

Both models proposed in this paper, particularly the SVM model, potentially provide an accurate diagnosis of a neurocognitive disorder (e.g., PD, MS, ET, CBS) for further medical examination. The stronger performance of the SVM-based model in comparison to its CNN counterpart exemplifies the strengths of using feature extraction over a strictly computer vision-based approach to identify the more unsteady drawings expected from a PD patient.

Our feature extraction-based model, rather than replace the current clinical diagnosis process, can complement it to enhance the efficiency and accessibility of the process. A patient, for example, can use our model to gain an accurate understanding of whether they have a neurocognitive disorder. They would simply need to draw a spiral, take a picture of the spiral, and our model will return its diagnosis with an accuracy of 86%. Further medical attention can be sought to identify which specific neurocognitive disorder the patient has. Future research should focus on creating an

	Predicted Negative	Predicted Positive
Actual Negative	8	1
Actual Positive	2	10

Table 2: Confusion matrix for HOG feature extraction-based model. Visual comparison of the predicted vs actual diagnoses in the 21 image testing set used to evaluate the feature extraction-based model.

https://doi.org/10.59720/24-048

algorithm that can differentiate between tremors caused by different neurocognitive disorders. This would allow for the direct, automated diagnosis of PD.

MATERIALS AND METHODS

In the development and application of both models, the programming language Python was used along with a variety of its libraries: NumPy, Matplotlib, scikit-learn, and scikit-image. The development of the CNN-based model utilized the deep learning framework PyTorch and its corresponding package torchvision. The dataset was obtained from Kaggle.com (11). Five metrics were used to evaluate both models: accuracy, precision, recall, f1-score, and AUC. Recall was calculated by dividing the number of true positives by the number of true positives and false negatives (14). Precision was calculated by dividing the number of true positives by the number of true positives and false positives (14). F1-score, which factors in both precision and recall, was calculated using the following equation.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

CNN-Based Model

Prior to building the model itself, the data was transformed using data augmentation (**Figure 2**). Data augmentation plays a significant role in the improvement of model generalization and the prevention of overfitting by artificially increasing data variability (21). Data variability is increased by taking existing data, in this case images, and altering them (21). Transforms

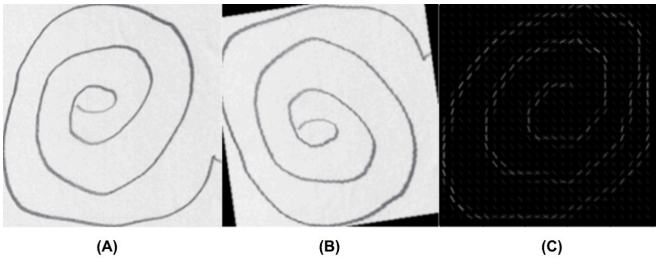

Figure 2: Visual representation of original and augmented spirals drawn by an individual with PD. (A) Original image. (B) Image post data augmentation (10° rotation, vertically flipped) for CNN model, using library torchvision's transforms module. (C) Image post feature extraction for SVM model, using library scikit-image's "hog" function.

were used to rotate the images by 10° in either direction, horizontally flip 30% of the images, vertically flip another 30% of the images, and convert the images into tensors, which is a numerical form of the PIL Image. Transforms improve model generalization by training the model on images that haven't been perfectly captured, allowing it to perform stronger on real-world data that may not be "perfect" either (21).

The model architecture was then built, utilizing one convolutional and one classifier block. The convolutional block consisted of three sets of 2D convolutional (conv2D), Rectified Linear (ReLU), and MaxPool2D layers. The 2D convolutional layer iterated over 2D sections of the image data, performing element-wise multiplication, and thus transforming the 2D

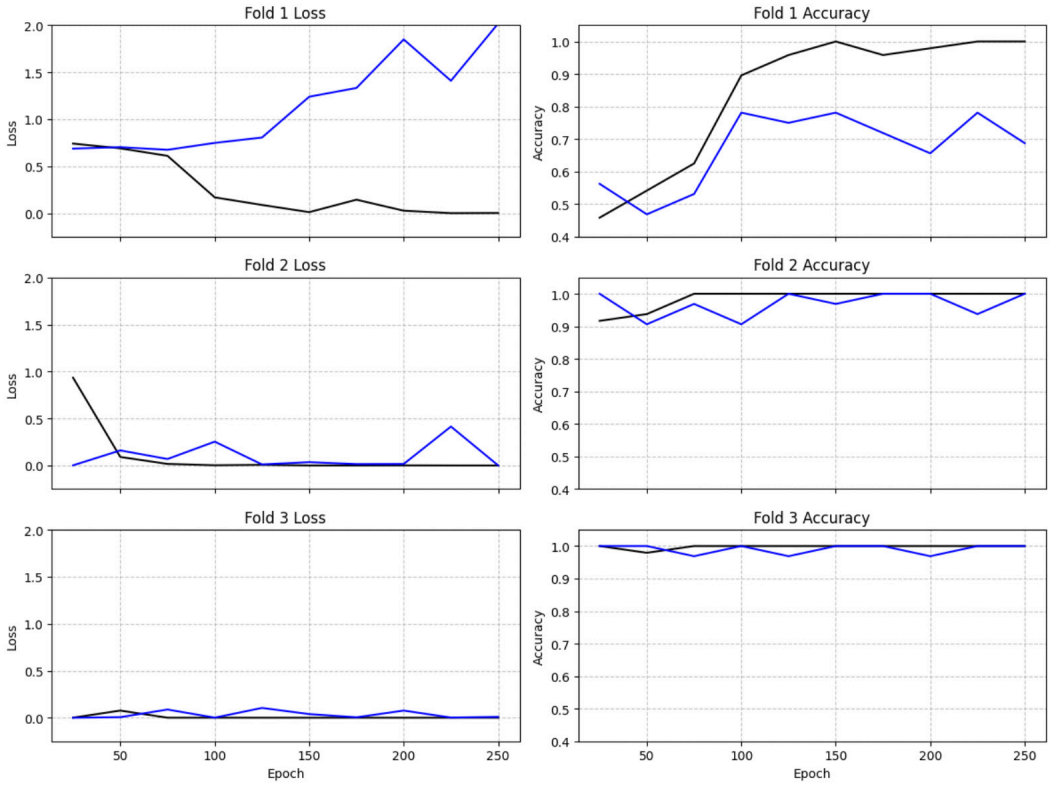

Figure 3: Graphical representation of k-fold cross-validation accuracy and loss curves. Graphs show loss (left) and accuracy (right) curves for k-fold cross-validation (k=3). Black lines represent train loss and accuracy curves; blue lines represent validation loss and accuracy curves. There was a general increase in accuracy and decrease in loss across all 3 folds. Graphs were created using Python library Matplotlib.

https://doi.org/10.59720/24-048

matrix while maintaining its size (22). The ReLU activation layer iterated over each data point, converting it to 0 if it was negative (23). The ReLU activation layer introduced the model to non-linearity, allowing the model to get accustomed to the naturally non-linear spiral drawings it was tested on (23). The max pooling (MaxPool2D) layer reduced the number of image data points by selecting the maximum value in every 2 by 2 subset of data (24).

The classifier block flattened the data into a single 1D array and then used 3 sets of alternating linear and ReLU layers. Each linear layer along with its corresponding ReLU activation function reduced the number of features until only 2 remained, representing the probability that the image is drawn by a PD or non-PD patient. The greater of the two is the returned diagnosis.

The model was trained using k-fold cross-validation. Three folds were used, each iterating over the training and validation set 250 times (250 epochs). For each fold, the training set was randomly split into a smaller training and validation set. The model was trained on the smaller training set and progress was tracked using the validation set. Utilizing k-fold cross-validation prevented overfitting and provided a more robust assessment of how the model would perform on unseen data (25). The model was trained using the Adam optimizer (learning rate of 0.001) and the multiclass cross entropy loss function, which are both commonly used for medical image classification (26). The accuracies on the validation set after each fold were 0.66, 1.0, and 1.0, with a validation loss reaching as low as 0.003 (**Figure 3**). The testing set was then passed into the model and the final accuracy, precision, recall, f1-score, and AUC were recorded.

Feature Extraction-Based SVM Model

To build the feature extraction-based SVM model, the images were taken and passed through scikit-image's HOG function. HOG is a feature extraction method that calculates a gradient for every pixel and uses the magnitude and orientation of that gradient to compute histograms from which features are extracted (27). HOG feature extraction works well in the analysis of spiral drawings because of its ability to simplify an image, extracting only its most important features where changes in gradient are largest (27). This simplification process allows it to extract the unsteady parts of the drawings, making differentiation between the two classes far easier. 11 orientations, 10 pixels per cell, and 3 cells per block (arguments for scikit-image's HOG function) were found to work well with the 256 by 256 pixel images used to test both models and were used to extract 52,371 numerical features from each image. A visual representation of the resultant of this simplification process can be seen in **Figure 2**. 80% of the feature-extracted image data was used to create a training set, and the other 20% was used for a testing set. The training set was then passed through a support vector machine (SVM), which found a hyperplane that best separated the data from the two classes (28). An SVM was used because of its effectiveness when working with a training set that has more features (52,371) than data points (102 images) (28). The testing set was then passed into the trained SVM, and the final accuracy, precision, recall, and AUC were recorded.

Received: January 25, 2024
Accepted: June 05, 2024
Published: August 09, 2024

REFERENCES

1. "Parkinson's Disease: Challenges, Progress, and Promise." *National Institute of Neurological Disorders and Stroke.* www.ninds.nih.gov/current-research/focus-disorders/parkinsons-disease-research/parkinsons-disease-challenges-progress-and-promise. Accessed 23 Dec. 2023.
2. "Parkinson's Disease." *American Association of Neurological Surgeons*. www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Parkinsons-Disease. Accessed 17 Dec. 2023.
3. "How Parkinson's Disease Is Diagnosed." *Johns Hopkins Medicine*. www.hopkinsmedicine.org/health/treatment-tests-and-therapies/how-parkinson-disease-is-diagnosed. Accessed 16 Jan. 2024.
4. "Parkinson's Disease." *Mayo Clinic*. www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062. Accessed 23 Dec. 2023.
5. "Quarter of Parkinson's sufferers were wrongly diagnosed, says charity." *The Guardian*. www.theguardian.com/society/2019/dec/30/quarter-of-parkinsons-sufferers-were-wrongly-diagnosed-says-charity. Accessed 26 Dec. 2023.
6. "Parkinson's Disease." *Mayo Clinic*. www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055. Accessed 23. Dec. 2023.
7. "Tremor." *Parkinson's Foundation*. www.parkinson.org/understanding-parkinsons/movement-symptoms/tremor. Accessed 3 Jan. 2024.
8. "What is machine learning?" *IBM*. www.ibm.com/topics/machine-learning. Accessed 18 Jan. 2024.
9. "What is supervised learning?" *IBM*. www.ibm.com/topics/supervised-learning. Accessed 18 Jan. 2024.
10. Davenport, Thomas and Ravi Kalakota. "The potential for artificial intelligence in healthcare." *Future Healthcare Journal*, vol. 6, no. 2, 13 June 2019, pp. 94-8, https://doi.org/10.7861/futurehosp.6-2-94.
11. "Parkinson's Drawings." *Kaggle*. www.kaggle.com/datasets/kmader/parkinsons-drawings. Accessed 20 Nov. 2023.
12. Zham, Poonam, et al. "Distinguishing Different Stages of Parkinson's Disease Using Composite Index of Speed and Pen-Pressure of Sketching a Spiral." *Frontiers in Neurology*, vol. 8, 06 Sep. 2017, https://doi.org/10.3389/fneur.2017.00435.
13. "Accuracy, precision, and recall in multi-class classification." *EvidentlyAI*. www.evidentlyai.com/classification-metrics/multi-class-metrics. Accessed 3 Jan. 2024.
14. "Classification: Precision and Recall." *Google*. developers.google.com/machine-learning/crash-course/classification/precision-and-recall. Accessed 24 Dec. 2023.
15. "What is the F1-score?" *Educative*. www.educative.io/answers/what-is-the-f1-score. Accessed 3 Jan. 2024.
16. "Classification: ROC Curve and AUC." *Google*. developers.google.com/machine-learning/crash-course/classification/roc-and-auc. Accessed 24 Dec. 2023.

https://doi.org/10.59720/24-048

17. "Classification: Check Your Understanding (ROC and AUC)." *Google*. developers.google.com/machine-learning/crash-course/classification/check-your-understanding-roc-and-auc. Accessed 3 Jan. 2024.
18. "Conditions that Mimic Parkinson's." *Parkinson's Foundation*. www.parkinson.org/understanding-parkinsons/getting-diagnosed/conditions-that-mimic-parkinsons. Accessed 26 Dec. 2023.
19. "Tremor." *Beaumont*. www.beaumont.org/conditions/tremor. Accessed January 3, 2024.
20. "Progression of Parkinson's." *Parkinson Canada*. www.parkinson.ca/about-parkinsons/progression-of-parkinsons/. Accessed July 13, 2024.
21. "What are Data augmentation techniques: [2024 update]." *ubiAI*. ubiai.tools/what-are-the-advantages-anddisadvantages-of-data-augmentation-2023-update/. Accessed 4 Jan. 2024.
22. "Convolutional Layer." *Databricks*. www.databricks.com/glossary/convolutional-layer. Accessed 28 Dec. 2023.
23. "An Introduction to the ReLU Activation Function." *BuiltIn.* builtin.com/machine-learning/relu-activation-function. Accessed 31 Dec. 2023.
24. "Introduction To Pooling Layers In CNN." *Towards AI*. towardsai.net/p/l/introduction-to-pooling-layers-in-cnn. Accessed 28 Dec. 2023.
25. "K-fold Cross-validation." *Shiksha Online*. www.shiksha.com/online-courses/articles/k-fold-cross-validation/. Accessed 4 Jan. 2024.
26. Rajaraman, Sivaramakrishnan, et al. "Novel loss functions for ensemble-based medical image classification." *Public Library of Science One,* Dec. 2021, https://doi.org/10.1371/journal.pone.0261307.
27. "What is Histogram of Oriented Gradients (HOG)?" *Educative*. www.educative.io/answers/what-is-histogram-of-oriented-gradients-hog. Accessed 31 Dec. 2023.
28. "SVM Machine Learning Tutorial – What is the Support Vector Machine Algorithm, Explained with Code Examples." *FreeCodeCamp*. www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/. Accessed 28 Dec. 2023.