

Identification of Parkinson's Disease Facial Symptoms By Utilizing Deep Learning Techniques

Ioannis Kyprakis

UFR Sciences et Techniques
Université de Bourgogne

A Thesis Submitted for the Degree of
MSc in Informatics (Health-AI)

· 2023-2024 ·

Abstract

Parkinson's Disease (PD) is a progressive neurological disorder characterized by motor and non-motor symptoms, including hypomimia, a significant reduction in facial expressiveness. Hypomimia, in conjunction with other symptoms, complicates the accurate assessment of depression in PD patients. This research presents a novel approach for estimating depression in PD patients by analyzing facial video recordings using deep learning techniques. The study leverages two datasets: a comprehensive video dataset comprising 1,875 facial recordings from 173 patients, and a detailed clinical dataset consisting of 140 patients and 18 features, including critical patient information such as UPDRS III scores and LEDD. Our results demonstrate that the 3D-CNN-LSTM model with attention outperforms other models, achieving an accuracy of 82% in binary classification and 71.17% in multiclass classification tasks. Furthermore, machine learning models applied to clinical data to reinforce the findings, highlighting key features such as UPDRS III scores and LEDD as significant predictors of depression. This research establishes a new benchmark for the automated detection of depression in PD patients, emphasizing the potential of integrating facial video analysis with clinical data for more accurate and early diagnosis.

Keywords: Parkinson's Disease; Depression Detection; Hypomimia; Deep Learning; Facial Video Analysis; Machine Learning; Neurological Disorders; XAI

Contents

Acknowledgments	vi
1 Introduction	vii
2 Background	xi
3 Material and methods	xiv
3.1 Data collection	xiv
3.1.1 Facial video recordings	xv
3.1.2 Speech recordings	xvii
3.1.3 Upper limb tremor measurements	xviii
3.1.4 Motion data & Plantar pressure data	xix
3.1.5 Questionnaire & Scale data	xix
3.2 Dataset	xx
3.2.1 Video Dataset	xx
3.2.2 Clinical Dataset	xxi
3.3 Proposed pipeline	xxiii
3.3.1 Preprocess	xxiii
3.3.2 Model architectures	xxiv
3.4 Experimental Setup and Implementation	xxx
4 Results	xxxii
4.1 Metrics	xxxiii
4.2 Results	xxxiv
4.2.1 Deep learning approach using facial video data	xxxiv

4.2.2	Machine learning approach	xxxvi
4.3	Comparison with state-of-the-art	xxxviii
5	Discussion and Conclusion	xxxix
5.1	Discussion	xxxix
5.2	Conclusion	xlii
A	The first appendix	xliv
	Bibliography	liv

List of Figures

3.1	Task open-close mouth.	xvi
3.2	Task smile-puckering.	xvi
3.3	Task maintaining a still posture.	xvi
3.4	Task moving tongue in and out.	xvii
3.5	Task moving tongue up and down.	xvii
3.6	Task moving tongue left and right.	xvii
3.7	3D CNN-LSTM with attention architecure.	xxvii
3.8	Vivit architecture [41].	xxix
3.9	Video Swim Transformer architecture [42].	xxx
4.1	Accuracy & Loss for training and validation fold	xxxvi
4.2	Feature importance for RandomForest and LightGBM based on the multiclass problem	xxxviii

List of Tables

3.1	Key technical features GoPro HERO9 Black.	xv
3.2	Class Distribution Binary	xxi
3.3	Class Distribution Multiclass	xxi
3.4	Clinical Features Used for Depression Estimation in PD	xxii
3.5	Clinical dataset's Class Distribution Binary	xxiii
3.6	Clinical dataset's Class Distribution Multiclass	xxiii
4.1	Final Results for binary classification	xxxiv
4.2	Final Results for multiclass classification	xxxiv
4.3	Best Hyperparameters for Binary Classification	xxxv
4.4	Best Hyperparameters for Multiclass Classification	xxxv
4.5	Binary results on Clinical Dataset	xxxvi
4.6	Multiclass results on Clinical Dataset	xxxvii
A.1	Experiments on Binary Classification 3D-CNN	xliv
A.2	Experiments on Multiclass Classification 3D-CNN	xliv
A.3	Experiments on Binary Classification 3D-CNN-LSTM with attention	xlv
A.4	Experiments on Multiclass Classification 3D-CNN-LSTM with attention	xlv
A.5	Experiments on Binary Classification ViVit	xlvi
A.6	Experiments on Multiclass Classification ViVit	xlvi
A.7	Experiments on Binary Classification Video Swin tiny	xlvi
A.8	Experiments on Multiclass Classification Video Swin tiny	xlvi
A.9	Experiments on Binary Classification Video Frame-wise	xlvi
A.10	Experiments on Multiclass Classification Frame-wise	xlvi
A.11	Binary results on Clinical Dataset with Grid Search tuner	xlvi

A.12 Multiclass results on Clinical Dataset with Grid Search tuner	xlvi
--	------

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Manolis Tsiknakis, for their invaluable guidance, support, and encouragement throughout the course of my research. Their insights and expertise were instrumental in shaping this work.

I am also deeply grateful to PhD-c Vasileios Skaramagkas, whose guidance and willingness to share their research greatly enriched my understanding and progress. Their mentorship has been a source of inspiration and motivation.

Additionally, I would like to thank the doctors, Zinovia Kefalopoulou & Kleio Spanaki, for their invaluable assistance and expertise. Their contributions provided a critical perspective and were essential in ensuring the clinical relevance and accuracy of this research.

I would also like to extend my sincere thanks to the Hellenic Mediterranean University and the Intelligent Systems & Computer Architecture Laboratory for providing access to the DGX A100 station. Their support and resources were crucial in the successful implementation of this research.

Lastly, I would like to thank my colleagues, friends, and family for their unwavering support and encouragement during this journey.

Chapter 1

Introduction

Parkinson Disease (PD) ranks as the second most prevalent neurological disorder following Alzheimer's, impacting 2-3% of individuals over the age of 65 worldwide [1]. The disease is characterized by neuronal degeneration in the substantia nigra, leading to dopamine deficiency in the striatum, and the presence of intracellular inclusions with α -synuclein aggregates [2]. A combination of environmental and lifestyle factors, along with sex, age, and genetic predispositions, play significant roles in the development of PD [3]. Additionally, several genetic mutations and polymorphisms have been linked to PD, including mutations in the SNCA, LRRK2, and PARK7 genes. Environmental factors, such as exposure to pesticides and heavy metals, have also been implicated in increasing the risk of developing PD [4].

Although PD currently has no cure, treatments focusing on dopamine replacement and surgical interventions like deep brain stimulation (DBS) significantly alleviate motor symptoms [5]. Mortality rates do not increase during the first decade after the disease's onset, but eventually, they double relative to the general population. Falls and aspiration pneumonia are the primary reasons for hospitalization and death [6].

PD presents with both motor and non-motor symptoms, significantly impacting the lives of those affected. Among the motor symptoms, bradykinesia is a hallmark feature, often accompanied by other primary motor manifestations such as rigidity and resting tremor, along with various supportive and exclusionary criteria [7]. As the disease progresses, additional motor symptoms emerge, including hypomimia—a reduction in facial expressions—along with micrographia (smaller handwriting), speech and voice impairments, and difficulties with swallowing [8]. These motor dysfunctions progressively interfere with daily activities, leading to a decline in patients' quality of life. As the severity of motor symptoms increases, gait abnormalities and postural instability further compromise patients' independence and

safety.

While motor symptoms are central to the clinical diagnosis of PD, most patients also experience non-motor symptoms likely linked to non-dopaminergic pathways. These include mood and cognitive changes such as depression, apathy, and reduced cognitive ability. Other common non-motor symptoms are fatigue, sleep disturbances, autonomic issues (like orthostatic hypotension, urogenital dysfunction, constipation, and excessive sweating), as well as sensory complaints, which are all part of the PD clinical spectrum [9]. These non-motor symptoms can appear early in the disease, sometimes preceding the primary motor symptoms by years or even decades [10]. As the disease progresses, non-motor symptoms become more prevalent and significantly impact quality of life [11], overall disability progression [12], and mortality.

Diagnosing PD primarily relies on clinical evaluation of the patient [13]. The Movement Disorder Society - Unified Parkinson's Disease Rating Scale (MDS-UPDRS) is currently considered the gold standard for assessing and monitoring PD [14], with the Hoehn and Yahr (H&Y) scale [15] also being widely used. Despite promising research avenues, no definitive biomarkers for PD have been identified, reinforcing the importance of clinical diagnosis. Early-stage detection is particularly challenging due to the diverse manifestation and progression of symptoms, resulting in complex clinical presentations of PD [16].

Deep Learning (DL), an advanced form of Machine Learning (ML), has further pushed the boundaries in various health fields, achieving state-of-the-art (sota) performance by automatically learning representations from raw, high-dimensional data through Deep Neural Networks (DNN) with minimal feature engineering [17,18]. DL strategies have become essential for handling complex data from sensor-enabled devices, driving significant progress in tracking, monitoring, and predicting PD progression [19–21]. Beyond sensor data, DL has been applied to biomarker datasets, such as dopamine transporter data and serum cytokines, to improve classification performance [22]. Moreover, AI has recently been used to identify PD based on nocturnal breathing patterns, accurately estimating disease severity and progression [23].

Current research into wearable sensory devices for detecting PD symptoms or characteristics has advanced significantly [24]. These sensors can be placed on different parts of the body and are associated with various modalities and their combinations. Proper sensor use and careful selection of modalities can enhance remote healthcare services and enable more personalized diagnoses [25]. Additionally, digital sensors allow for the autonomous and unobtrusive collection of real-world data. The real-time capabilities of wearable technology, combined with detailed insights into a patient's vital signs, allow physicians to detect disease early and provide accurate diagnoses, whether in a clinical setting or remotely [26].

Therefore, using sensors to gather biomarkers from various modalities is crucial for real-time monitoring and accurately reflecting everyday symptoms and their variations [27].

Hypomimia, characterized by reduced facial expression, is a significant motor symptom of PD that has profound implications on both social interactions and the emotional well-being of patients. This diminished ability to express emotions is closely linked to higher levels of depression in PD patients, as it can lead to social withdrawal and feelings of isolation [28]. The social impact of hypomimia extends beyond personal interactions, affecting how patients engage with their surroundings and contributing to their overall mental health.

In the context of technological advancements, AI and digital sensors offer promising potential in monitoring hypomimia in real-time. These tools can provide valuable insights into the progression of PD by detecting subtle changes in facial expressiveness [8]. Furthermore, the integration of AI and deep learning in computing interfaces can adapt these technologies to accommodate the needs of PD patients, ensuring that reduced facial expressiveness does not hinder their interaction with digital tools, particularly those that rely on facial recognition and expressions for user engagement and feedback.

Although the detection of depression in patients with PD is of tremendous importance [29–31], there is limited research focused on assessing or estimating depression levels in PD patients, particularly using facial videos. Estimating depression in PD patients is particularly challenging due to hypomimia. Hypomimia can obscure typical facial cues associated with emotional states, making it difficult for clinicians and technological systems to accurately assess depressive symptoms. This lack of expressiveness may lead to underdiagnosis or misinterpretation of the patient’s mental health status [32]. Moreover, traditional diagnostic methods, which rely heavily on facial and emotional expressiveness, may not be as effective for PD patients [33]. Thus, there is a critical need for developing alternative assessment tools and methods that account for the unique presentation of depression in individuals with PD.

Compounding this challenge is the lack of comprehensive datasets in PD patients. Existing datasets often do not account for the nuances of hypomimia or its impact on depression assessment, making it difficult to analyze subtle facial changes that could indicate depressive states. The limited availability of video data tailored to this intersection hampers the development of ML/DL models and other technological interventions that require robust datasets for training and validation. Without sufficient high-quality video datasets that capture the facial dynamics of PD patients, researchers face significant obstacles in improving diagnostic accuracy. Therefore, there is a pressing need for focused data collection efforts and research initiatives to build datasets that accurately reflect the experiences and symptoms of PD patients, ultimately aiding in the creation of more reliable diagnostic tools.

The primary objective of this research is to assess depression in PD patients using a novel and comprehensive multi-modal dataset. Despite the significant impact of depression on the quality of life in PD patients, the intricate connections between these conditions have not been thoroughly explored, particularly through the lens of facial expressiveness. Due to the inherent challenges posed by hypomimia—reduced facial expressiveness often seen in PD patients—there has been limited research interest in leveraging facial videos for depression assessment in this population. By utilizing this innovative dataset, we aim to fill a critical gap in current research, paving the way for new insights and advancements in understanding PD.

Furthermore, this research addresses a critical gap in the estimation of depression in PD, particularly concerning the role of hypomimia in depression assessment, an area that has received limited research focus. By leveraging AI and deep learning technologies, this study aims to provide novel insights and tools for a more accurate diagnosis and management of depression in PD patients. A key aspect of this work involves investigating the potential of diagnosing hypomimia through video analysis, which can be performed using handheld devices like smartphones and tablets. These devices offer the advantage of being widely accessible, enabling continuous and non-intrusive monitoring of facial expressions in real-world settings. This approach not only enhances the accuracy of hypomimia detection but also contributes to a more comprehensive understanding of its relationship with depression in PD. Ultimately, this research will significantly advance the field and improve the quality of life for individuals living with PD.

The rest of the thesis report is structured as follows: Chapter 2 provides a detailed review of the existing literature and state-of-the-art technologies in assessing depression of PD patients' research. Chapter 3 outlines the methodology for developing the multi-modal dataset and the proposed pipeline used for our data & models. Chapter 4 presents the metrics used, the results, and a comparison with state-of-the-art. Finally, Chapter 5 offers discussion-conclusions and recommendations for future research.

Chapter 2

Background

Depression is a critical and often overlooked symptom in PD, impacting a significant portion of patients. Research indicates that the prevalence of depression in PD patients ranges from 30% to 45% [34,35], with its presence being associated with a poorer quality of life, increased disability, and a faster progression of motor symptoms [30]. Despite the high prevalence and impact, depression in PD is under-diagnosed and under-treated [36,37], largely because it can manifest differently compared to the general population, often being mistaken for the effects of PD itself. This underscores the need for more nuanced and effective tools to detect and monitor depression in PD patients .

While there has been progress in developing speech-based models for detecting depression, there remains a significant gap in leveraging facial video analysis for this purpose. Research has shown that speech-based models can effectively identify depression by analyzing vocal biomarkers, focusing on aspects such as pitch, tone, and rhythm, which are often altered by depressive states. For example, Ozkanca et al. [38] demonstrated that vocal biomarkers could be effective for depression screening among Parkinson's Disease (PD) patients, achieving accuracies up to 77% by analyzing voice samples to predict depression states. Additionally, Soumaya et al. [39] proposed a model for detecting Parkinson's disease using a genetic algorithm (GA) combined with a Support Vector Machine (SVM) classifier. They utilized discrete wavelet transform (DWT) and Mel-frequency cepstral coefficients (MFCC) for feature extraction from speech signals, achieving a classification accuracy of 91.18%.

Similarly, machine learning techniques have been applied to enhance the detection of depression, as evidenced by [40], where a predictive model for depression in PD patients was developed using the support vector machine (SVM) algorithm. The study utilized data by analyzing the PD Epidemiology (PDE) and included 223 PD patients, incorporating various predictors such as sociodemographic factors,

health habits, motor symptoms, sleep behavior disorders, and neuropsychological indicators. The results indicated that the Gaussian kernel-based Nu-SVM model achieved the highest accuracy, with 95% overall accuracy. This model highlighted late motor complications, particularly levodopa-induced dyskinesia, as significant risk factors for depression in PD patients. Moreover in [41] the author explores the use of stacking ensemble methods to predict depression in PD patients. Using the PDE dataset consisting of 238 PD patients aged 65 and older, the research developed nine different machine learning models combined in various ways with logistic regression (LR). Among these, the random forest+LR model achieved the best predictive performance with a root mean square error (RMSE) of 0.16, an index of agreement (IA) of 0.73, and variance of errors (Ev) of 0.48. The study identified key predictors of depression in PD patients, including cognitive and motor assessments like K-MMSE, K-MoCA, and UPDRS scores. These findings highlight the potential of ensemble machine learning models in the early detection of depression in PD patients, which is crucial for timely intervention and care. However, the potential of facial expressions, which can also reveal crucial insights into an individual's emotional state, remains largely untapped, particularly in the context of PD. The application of video analysis to detect depression in PD patients is a relatively unexplored area, presenting an opportunity for new advancements that could complement or transcend existing speech-based methods.

Recent advancements in video analysis have significantly enhanced the detection of symptoms in PD. In [42] developed an automated video analysis method to detect Parkinson's Disease symptoms using machine learning classifiers like Support Vector Machines and Random Forests. The approach achieved an F1 score of 0.93, demonstrating the potential of using unstructured video data for PD symptom detection. Research has focused on using video recordings to analyze facial expressions, gait patterns, and other motor functions, enabling clinicians to assess the severity and progression of PD with greater accuracy. These studies often employ machine learning and computer vision techniques to extract meaningful features from video data, providing insights into the patients' motor capabilities and symptom fluctuations. For example, in [43] the researchers collected video data of smiles from PD patients and healthy controls, analyzing facial expression features such as amplitude and tremors using machine learning algorithms. The study found that these facial features could effectively distinguish PD patients from healthy individuals, achieving high precision and F1 values, particularly with the use of support vector machines (SVM) and long short-term memory (LSTM) models. Yin et al. [44] utilized a 3D Convolutional Neural Network (CNN) as a baseline for classifying PD severity and explored transfer learning from non-medical datasets due to the scarcity of clinical data. To address domain discrepancies, a Temporal Self-Attention (TSA) mechanism was introduced, focusing on subtle temporal cues like

tremor frequency. Seven tasks from the MDS-UPDRS part III were evaluated, with an emphasis on bradykinesia and postural tremors. Additionally, a multi-domain learning method was proposed for predicting patient-level PD severity. The TSA and task-assembling methods proved effective, achieving MCC scores of 0.55 for binary task-level classification and 0.39 for three-class patient-level classification on the PD video dataset.

While video analysis has been effectively used for estimating depression in the general population, there remains a significant gap in its application for PD patients. Several studies have demonstrated the potential of video-based systems to detect depression. For instance, in [45] the researchers proposed hierarchical framework in interpreting depression from the Self-Rating Depression Scale (SDS) questionnaire and its corresponding video recordings. The system achieved a high accuracy of 92.5% when combining both SDS scores and video data, significantly outperforming traditional methods that rely solely on questionnaire data or video analysis. The framework's use of a 3D CNN for local temporal feature extraction, combined with a redundancy-aware self-attention mechanism for global feature aggregation. In [46] the authors explore the use of facial action units (AUs) as biomarkers for diagnosing depression through the analysis of video data. The study found significant differences in the intensity of certain AUs associated with sadness and happiness between depressed and non-depressed individuals, suggesting these facial expressions could serve as objective indicators of depression. the ROCKET combined with Logistic Regression model achieved the highest accuracy of 70% in classifying depression based on facial expression data. Shangguan et al. [47] introduced a dual-stream multiple instance learning (MIL) approach that effectively captures both instance-level and bag-level information to classify depression from facial videos. Their model demonstrated a notable accuracy of 74.7%, highlighting the potential of weakly supervised learning frameworks in advancing automated depression detection.

However, despite these advancements, the unique challenges posed by PD—such as hypomimia, which reduces facial expressivity—have not been adequately addressed in depression assessment using video analysis. The current literature lacks studies that focus on adapting these methodologies for PD patients, who often exhibit overlapping motor and non-motor symptoms. This gap presents an opportunity to extend existing video analysis techniques to develop tailored tools for detecting depression in PD patients, which could enhance diagnostic accuracy and improve overall patient care.

Chapter 3

Material and methods

3.1 Data collection

The dataset was developed by the research team of BMI Lab, HMU (<https://bmi.hmu.gr/>), and clinicians from General University Hospital of Heraklion, and from General University Hospital of Patras. Data collection was conducted with PD patients from the two major university hospitals aforementioned. The dataset is the largest of its kind with respect to the number of modalities and participant number. It comprises a variety of modalities, providing a comprehensive resource for the study of PD. The included modalities are:

- Facial video recordings
- Speech/voice recordings
- Upper limb movement measurements
- Plantar pressure data
- Motion data
- Questionnaire & Scale data

The experimental procedure is structured into two cycles of exercises to evaluate the effects of medication on patients with Parkinson's Disease (PD).

In the first cycle, patients are assessed in the "OFF" state, which means they have not yet received their medication for the last 12 hours. This state allows for the evaluation of baseline motor and non-motor symptoms without the influence of pharmacological treatment. In the second cycle, patients are

evaluated in the "ON" state. This cycle begins one hour after each patient has received their prescribed medication, allowing sufficient time for the treatment to take effect. This state facilitates the assessment of symptom improvement and medication efficacy.

The structured approach of the two cycles allows for a comprehensive comparison of the patients' condition in both medicated and unmedicated states, providing valuable insights into the impact of treatment on the progression of PD symptoms.

3.1.1 Facial video recordings

The recording of movements and facial expressions is conducted using a camera mounted on a tripod and its position is adjusted to cover the whole of the of the patient's face. The selected camera is the GoPro Hero 9, which offers 5K video resolution and 20-megapixel still images, in addition to advanced features such as HyperSmooth 3.0 stabilization. Its robust construction and waterproof design make it suitable for capturing high-quality data in various research settings. The key technical feature of GoPro HERO9 can be see at Table 3.1.

Table 3.1: Key technical features GoPro HERO9 Black.

Image Sensor	CMOS	Resolution (MP)	23.6
Manufacturer	GoPro	Video Resolution	5K
Display	LCD touch screen	Shutter speed	1/2000 – 1/25
Water Deapth Protection	10m	Microphone	Yes
Speaker	Yes	WiFi	Yes
Batter	1720 mAh	Weight	158g
FPS	240	Power supply	Power over USB or via GPIO

For the video acquisition, patients were recorded at 4k with 30 frames per second in a hospital room with controlled lighting and a uniform background to ensure consistency and quality in the data collected. The recordings followed a series of tasks based on a protocol developed by the neurologists in collaboration with a speech therapist. These tasks are designed to elicit specific muscle movements similar to those involved in facial expressions, providing valuable data for assessing expressive difficulties such as hypomimia, which is a common feature of PD. The tasks recorded include:

- Opening and closing the mouth Figure 3.1
- Smiling and puckering the lips 3.2
- Maintaining a still posture Figure 3.3

- Moving the tongue in and out Figure 3.4
- Moving the tongue up and down Figure 3.5
- Moving the tongue left and right Figure 3.6

These tasks are crucial for analyzing the degree of facial expressiveness and muscular control in PD patients, contributing to a more comprehensive understanding of hypomimia.



Figure 3.1: Task open-close mouth.



Figure 3.2: Task smile-puckering.



Figure 3.3: Task maintaining a still posture.



Figure 3.4: Task moving tongue in and out.

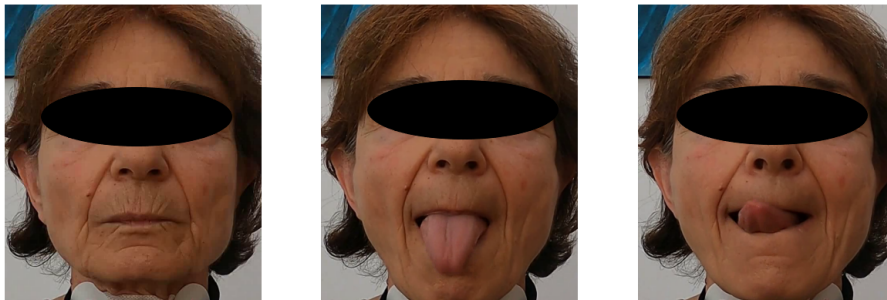


Figure 3.5: Task moving tongue up and down.

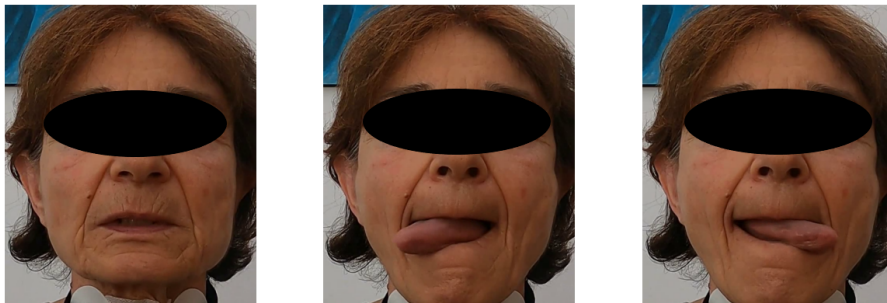


Figure 3.6: Task moving tongue left and right.

3.1.2 Speech recordings

The audio recordings were captured using the Blue Microphone Snowball, which connects to the computer via USB. The sound recording and analysis were conducted using the Praat software, version 6.0.36, configured to record mono sound at a sampling rate of 44,100 Hz. The Blue Microphone Snowball is designed with a dual-capacitor capsule, making it suitable for a variety of applications including voice recording, music recording, and podcasting.

This microphone offers multiple pickup patterns, including cardioid, omnidirectional, and cardioid with pad, providing flexibility for capturing audio in different environments. The package also includes an adjustable metal stand and a USB cable, ensuring ease of setup and use. The microphone is compatible

with both Windows and macOS operating systems, enhancing its versatility for different recording needs.

Speech recordings were conducted immediately following the video recordings, in the same room, which was free from external noise to ensure high-quality audio capture. The microphone was positioned on a table 20-30 cm away from the patient to optimize sound clarity. Similar to the video recordings, the speech tasks were structured according to a protocol developed by speech therapists, designed to assess various aspects of vocal function and expression. The tasks recorded include:

- Repeating a sentence.
- Two questions of the type "Good morning, how are you?" with the aim of provoking the colouring of the voice.
- Prolonged pronunciation of vowels a/e/o in 3 tones (low, medium, high) and in rising/falling pitch.
- Completion of Acoustic Voice Quality Index (AVQI). The AVQI uses some speech features (e.g., shimmer etc) to evaluate the voice quality.

These tasks are designed to provide comprehensive data on the patients' speech capabilities, allowing for the analysis of vocal quality and the identification of any speech-related impairments commonly associated with PD.

3.1.3 Upper limb tremor measurements

The Empatica EmbracePlus will be utilized to monitor, record, and collect data on the upper limb movements of patients with PD. The EmbracePlus integrates Empatica's precise digital biomarkers with advanced sensor technology. It features a comprehensive array of sensors, including an accelerometer, gyroscope, electrodermal activity (EDA) sensor, and skin temperature sensor. These sensors provide raw data in AVRO format, including high-frequency measurements of acceleration, angular velocity, skin temperature, electrodermal activity, blood volume pulse, systolic peaks, and wrist tracking.

In addition to raw sensor data, the EmbracePlus enables the extraction of digital biomarkers such as movement intensity, pulse rate, pulse rate variability, and respiratory rate. These biomarkers are derived from physiological parameters calculated by Empatica's sophisticated algorithms, offering valuable insights into the physical and physiological states of PD patients.

The EmbracePlus is positioned on the metacarpal region of the patient's "weak" hand and remains in place throughout the collection of motion, speech, and video data. This positioning ensures consistent

data acquisition across different modalities, allowing for a holistic analysis of the patients' motor functions and physiological responses during the study.

3.1.4 Motion data & Plantar pressure data

Participants are asked to perform a specific sequence of exercises in two cycles, spaced one hour apart. During these exercises, kinesiological data are collected from the monitoring devices. The exercises are categorized into five distinct types:

1. Simple Walking: Participants walk in a straight line to assess basic locomotion.
2. Timed Up and Go Test (TUG Test): This involves recording the time taken for participants to rise from a sitting position and begin walking, providing insights into transitional movements.
3. Static Analysis Exercise: This exercise focuses on analyzing the participants' ability to maintain a static posture.
4. Gait Freeze Control Exercise: This exercise assesses the participants' ability to manage and control episodes of gait freezing, a common symptom in PD.
5. Balance Analysis Exercise with Exogenous Activation: Participants undergo balance assessment with external stimuli to evaluate their response to balance challenges.

Prior to beginning the exercises, the researcher or physician integrates the insoles into the participants' shoes to facilitate data collection. If the participant agrees, they may be provided with shoes that have the insoles already integrated, ensuring comfort and convenience during the performance of the exercises.

These exercises were selected based on criteria of simplicity, safety, and the ability to provide comprehensive data, following an extensive review of relevant literature. This structured approach ensures that the collected data are robust and provide valuable insights into the motor functions and capabilities of participants with PD.

3.1.5 Questionnaire & Scale data

For this research we used a variety of questionnaires. Our main focus will be the Geriatric Depression Scale (GDS) [48]. The Geriatric Depression Scale (GDS) is a comprehensive diagnostic and therapeutic tool used in a multidimensional and multidisciplinary context to identify patients who are at risk of or currently experiencing depression. This scale is specifically designed for the elderly population, offering

a sensitive and effective means of detecting depression in this demographic. The GDS facilitates early intervention and tailored therapeutic approaches, thereby improving patient outcomes and enhancing quality of life. According to the original paper the scoring of the GDS is the following:

- 0 - 9 = Normal
- 10 - 19 = Mild Depression
- 20 - 30 Severe Depression

In addition to the Geriatric Depression Scale (GDS), several other questionnaires and scales were utilized to obtain comprehensive data on various aspects of the participants' health and well-being. These instruments include: MDS-UPDRS, MoCA Test, Non-Motor Symptoms Scale (NMSS), Non-Motor Scale Questionnaire (NMQ), SCOPA-AUT, State Trait Anxiety Inventory (STAI), REM Sleep Behaviour Disorder Screening Questionnaire (RBDSQ), Epworth Sleepiness Scale (ESS), PD ImpulsiveCompulsive Disorders Questionnaire (QUIP), Neuropsychiatric questionnaire (NPI-Q), The PD Questionnaire (PDQ-39), Schwab & England (S&E), Physical Assessment Scale for the Elderly (PASE), CRS, EAT-10.

During the data acquisition process, strict protocols were followed to ensure that all information collected from participants is treated with the highest level of confidentiality. The data acquisition involved collecting sensitive information, including personal and clinical data, which requires careful handling to protect participants' privacy. In compliance with the General Data Protection Regulation (GDPR) and applicable national legislation, all personal data were anonymized and securely stored. Access to this data was restricted to authorized personnel only, and measures were implemented to safeguard the information against unauthorized access or breaches. By adhering to these protocols, we ensured that the integrity and confidentiality of the data are maintained throughout the research, fostering trust and upholding ethical standards in the study.

3.2 Dataset

3.2.1 Video Dataset

Our primary focus is the detection of depression through the analysis of facial video data. To achieve this, we have curated the dataset consisting of videos collected from participants during the study. Following the application of rigorous filtering and exclusion criteria to ensure data quality and relevance, the final dataset comprises 173 patients and a total of 1,875 video recordings available. Video recordings consist

of RGB, captured at 30 frames per second (fps) with a resolution of 3840x2160 pixels videos. The videos varied in length depending on the individual sessions and specific tasks performed by each patient. The depression distribution in our dataset is:

- No depression 56
- Mild Depression 92
- Severe Depression 25

The initial dataset is structured to include patient labels indicating the presence or absence of depression, categorizing individuals into two classes: “Depression” and “No Depression.” This classification is based on the generalization of the GDS measurements, where patients with mild and severe depression scores are labeled as “Depression,” while those with normal scores are labeled as “No Depression.” The class distribution as a Depression/No Depression problem can be seen at Table 3.2.

Table 3.2: Class Distribution Binary

Class	Number of Videos
No Depression	612
Depression	1263

To further investigate depression assessment in patients with PD, we employed a three-class evaluation using the GDS measurement. This classification differentiates patients into three categories: “Normal”, “Mild Depression”, and “Severe Depression,” providing a more nuanced understanding of the depression spectrum within the PD population. This three-class distribution allows for a more detailed analysis of depression severity and its impact on patient outcomes. The class distribution as No Depression/Mild Depression/Severe Depression problem can be seen at Table 3.3.

Table 3.3: Class Distribution Multiclass

Class	Number of Videos
No depression	612
Mild Depression	994
Severe Depression	269

3.2.2 Clinical Dataset

This dataset consists of clinical features provided by the doctors (Table 3.4), We will analyze them separately to provide a detailed, multi-dimensional understanding of the patients’ conditions. This

approach will enable us to identify specific patterns or correlations that may not be immediately apparent, ultimately leading to more informed clinical decisions.

Table 3.4: Clinical Features Used for Depression Estimation in PD

Feature	Description
Med_state	Refers to the medication state of the patient, typically indicating whether the patient is in an "ON" state (when the medication is active and symptoms are better controlled) or "OFF" state (when the medication has worn off, and symptoms are more pronounced).
Med state subjective	Refers to the patient's subjective assessment of their medication state, reflecting how they feel or perceive their symptom control.
H&Y	A commonly used clinical rating scale that measures the severity of PD. It stages the disease progression from 1 (mild symptoms) to 5 (severe disability).
Hypomimia	Hypomimia refers to reduced facial expressiveness, a common symptom in PD.
Hypomimia_class	A categorical classification of hypomimia severity based on the score in the previous column.
UPDRS_III	This part of the UPDRS specifically measures motor function and is widely used in clinical studies and treatment evaluations to assess the severity of motor symptoms in PD.
UDPRS_III_dif	Represents the difference in the UPDRS Part III score, before and after treatment or between "ON" and "OFF" medication states.
UDPRS_III_dif_%	The percentage change in the UPDRS Part III score, indicating the relative improvement or worsening of motor symptoms.
UPDRS_III_dif_class	A categorical classification based on the difference in UPDRS Part III scores, classifying the degree of improvement or worsening.
Overall_fluct	Refers to overall symptom fluctuations, indicating the variability in symptoms experienced by the patient.
Dyskinesia	Involuntary movements that can occur as a side effect of long-term levodopa use in PD treatment.
Fluctuations	Refers to fluctuations in the patient's symptoms, particularly motor symptoms.
LEDD	A measure used to standardize the dose of different PD medications to their equivalent dose in levodopa.
Drug_naive	Indicates whether the patient has not yet been treated with Parkinson's medications (i.e., "drug-naive").
LD	Levodopa, the primary medication used to treat PD by replenishing dopamine.
DA	Refers to a class of drugs (Dopamine Agonists) used to treat PD by mimicking dopamine in the brain.
Adv_Ther	Refers to more advanced therapeutic approaches, which could include options like deep brain stimulation (DBS) or infusion therapies.
Type_Adv_Ther	Specifies the type of advanced therapy used if the patient is undergoing such treatment.

After applying the exclusion criteria for missing values, the final dataset consists of 140 patients and 16 features. Features Adv_Ther and Type_Adv_Ther had to be skipped due to many missing values. The class distributions are shown in Table 3.5 for binary and in Table 3.6 for multiclass problem.

Table 3.5: Clinical dataset’s Class Distribution Binary

Class	Patients
No Depression	50
Depression	90

Table 3.6: Clinical dataset’s Class Distribution Multiclass

Class	Patients
No depression	50
Mild Depression	80
Severe Depression	10

3.3 Proposed pipeline

3.3.1 Preprocess

Initially, MTCNN [49] was utilized for face detection and alignment within each video frame. MTCNN is a robust model that not only detects faces but also trims the image accordingly to ensure that the detected face is centered and properly scaled. This process involves adjusting the orientation and position of the face to achieve a standard pose, which is critical for maintaining consistency across all input data. This uniformity in face alignment allows the model to focus on learning relevant features without being affected by variations in orientation, scale, or position.

Given the unique challenges associated with patient data, particularly involving individuals with PD, some cases required manual intervention for face alignment. These cases often included pronounced postural deviations or significant movements characteristic of PD, which the automated system occasionally failed to align accurately. Manual correction ensured that even in these challenging cases, the faces were correctly aligned to maintain the integrity of the input data.

In addition to face alignment, further preprocessing steps were applied to enhance the quality of the video data. These steps included normalizing pixel values to a range of [0,1], which is essential for ensuring consistent input features and aiding model convergence during training. Additionally, histogram equalization was performed to improve the contrast of the video frames, thereby enhancing the ability

of the convolutional neural networks to extract meaningful features from the data.

Furthermore, during the preprocessing phase, all videos were uniformly divided into 10-second segments, with each segment consisting of 30 frames per second, resulting in a total of 300 frames per clip. This standardization provided a uniform temporal structure across all video samples, allowing the model to examine facial expressions and movements within a fixed time frame. By maintaining a consistent frame count and clip duration, the preprocessing process ensured that the input data had uniform temporal features, which facilitated more accurate feature extraction and analysis. This approach was essential in eliminating biases that could arise from varying video lengths, thereby enabling the model to concentrate on the key features relevant to assessing PD patients.

3.3.2 Model architectures

Frame-Wise classification using Google-Net and 1d transformer

Inspired by [50], we developed a model tailored for frame-wise classification of video data, designed to efficiently capture both spatial and temporal features within a unified architecture. The model is structured in two key components, each responsible for extracting different types of information crucial for accurate video analysis.

The first component employs GoogleNet (InceptionV3) [51] to extract spatial features from each frame of the video. GoogleNet, is a deep convolutional neural network renowned for its capability to capture intricate visual patterns, processes each frame individually. This ensures that the model captures the detailed spatial structures and nuances within each frame, which are essential for understanding the content of the video at a granular level. By leveraging a pre-trained version of GoogleNet, the model benefits from robust feature extraction capabilities that have been fine-tuned on large-scale image datasets, ensuring high-quality spatial feature representation.

The second component of the model is a 1D transformer, which is responsible for capturing the temporal dynamics between the frames. While GoogleNet excels at understanding the content of individual frames, the 1D transformer is designed to model the sequential relationships between these frames, effectively learning how the visual information evolves over time. The transformer achieves this by applying attention mechanisms, allowing the model to focus on the most relevant frames and their temporal connections. This dual approach, combining the spatial prowess of GoogleNet with the temporal modeling capabilities of the transformer, results in a highly effective video classification model. The integration of these two components allows the model to not only understand what is happening in each frame but

also how these events unfold across the video, leading to more accurate and context-aware predictions.

3D-CNN

Three-dimensional Convolutional Neural Networks (3D CNNs) [52] are an advanced extension of the traditional 2D CNNs designed to process data with an additional temporal dimension, making them particularly effective for video analysis tasks such as action recognition. Unlike 2D CNNs that perform convolutions over the spatial dimensions (height and width), 3D CNNs extend these operations into the temporal dimension, capturing motion information by convolving over sequences of frames. This enables the network to learn spatial and temporal features simultaneously, which is critical for understanding dynamic activities within video sequences.

A typical 3D CNN architecture consists of multiple layers of 3D convolutions, each followed by non-linear activation functions, pooling layers for dimensionality reduction, and fully connected layers for final classification. For example, a 3D CNN model might include a series of convolutional and pooling layers that progressively extract higher-level spatiotemporal features from raw video data.

Advantages:

- **Comprehensive Feature Extraction:** The ability to capture temporal dependencies in addition to spatial features allows 3D CNNs to effectively model motion patterns and temporal dynamics, offering a significant advantage in video-related applications
- **Rich Representation:** 3D CNNs generate a rich feature representation by combining multiple types of information from adjacent frames, which enhances their performance in tasks such as human action recognition

Disadvantages:

- **Increased Computational Complexity:** The extension to three dimensions significantly increases the computational requirements, demanding more processing power and memory, which can lead to longer training and inference times.
- **Large Dataset Requirements:** Effective training of 3D CNNs often requires large, labeled video datasets, which can be challenging to acquire, especially in specialized domains

Despite these challenges, 3D CNNs have demonstrated superior performance in real-world applications, outperforming traditional methods in action recognition tasks by effectively leveraging spatiotemporal features.

Based on the above we developed a 3D Convolutional Neural Network (3D CNN) for video classification, designed to capture spatiotemporal features essential for dynamic content analysis. The model consists of three convolutional blocks, each featuring a 3D convolutional layer with increasing filter sizes (64, 128, and 256), followed by max pooling, batch normalization, and dropout layers. A dense layer with 512 units aggregates the features, and the output layer uses either a sigmoid or softmax activation depending on the classification task. Regularization techniques, including dropout, batch and layer normalization, are employed throughout the model to enhance feature extraction, to ensure robust learning and also prevent overfitting. This architecture balances complexity and performance, providing a strong baseline for video classification tasks.

3D CNN-LSTM with attention layer

In recent years, 2D CNNs combined with Long Short-Term Memory (LSTM) layers have been widely adopted for various video analysis applications due to their ability to capture spatial and temporal patterns effectively [53–55]. However, as the complexity of video data increases, researchers have explored more sophisticated architectures such as 3D CNNs combined with LSTMs [56]. This combination leverages the strengths of both models: 3D CNNs are adept at capturing spatiotemporal features across video frames, while LSTMs excel at modeling temporal dependencies and sequences. The integration of these architectures has proven particularly beneficial in challenging tasks, such as action recognition and video classification, where understanding both spatial features and temporal dynamics is critical. Furthermore, the addition of attention mechanisms in these models enhances their capability to focus on important aspects of the data, leading to improved performance in complex scenarios [57]. As research in this area progresses, the development of such hybrid models continues to contribute to advancements in video analysis, offering promising results across a range of applications. Based on [56] we utilized a 3D CNN-LSTM architecture with an attention layer Figure 3.7. This sophisticated approach is particularly potent because it captures both the depth of visual features and the sequence of motions effectively. The inclusion of an attention layer further refines the model’s focus, allowing it to prioritize more relevant information within the video sequence. This can lead to better learning efficiency and improved accuracy in tasks that require nuanced understanding of complex behaviors or events over time, such as action recognition in videos, patient monitoring in healthcare settings, or any scenario where contextual continuity is crucial. The attention mechanism enhances model interpretability by highlighting which frames most influenced its predictions, offering insights into the decision-making process of the network.

Our model is architected with multiple 3D convolutional layers that sequentially escalate in com-

plexity, starting with 32 filters and doubling in deeper layers up to 256 to capture intricate spatial and temporal dynamics. Each convolutional stage except the last one comprises of a $3 \times 3 \times 3$ kernel followed by max pooling with a $2 \times 2 \times 2$ window to reduce spatial dimensions while retaining essential features. The last layer is composed by 2x256 filters, a $2 \times 2 \times 2$ kernel followed by a max pooling layer with a $2 \times 2 \times 2$ window. Batch normalization follows each convolutional and pooling step to ensure efficient and stable training dynamics by normalizing layer inputs. To further refine feature representation, the flattened convolutional outputs are processed through fully connected layers with 1024 and 512 nodes, each incorporated with a dropout rate of 20% to mitigate overfitting. Temporal dependencies are captured via LSTM layers with 32 units, facilitating the understanding of temporal progression within video data. The inclusion of an attention layer post-LSTM allows the model to focus selectively on the most salient sequences, thereby optimizing the learning of relevant temporal patterns crucial for accurate classification. The inclusion of an additional LSTM layer after the attention mechanism provides a critical function in the network's architecture. This LSTM layer, with 16 units, processes the refined, attended sequences. The attention layer selectively emphasizes key temporal features, focusing the model on the most informative parts of the data. The subsequent LSTM layer is then able to capture these enhanced features, integrating them over time to construct a more nuanced understanding of the sequence's dynamics. This setup is particularly effective in scenarios like video data analysis, where temporal coherence and the identification of important moments can drastically improve the performance of tasks such as action recognition or event detection in videos. The layered approach ensures that both spatial and temporal features are not only captured but also prioritized effectively, leading to more precise and contextually aware outputs.

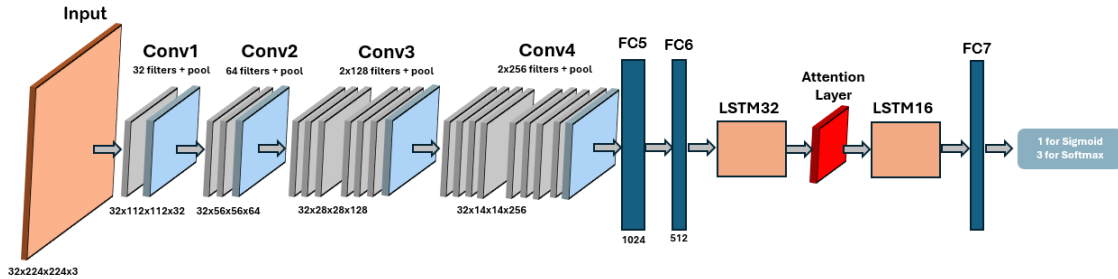


Figure 3.7: 3D CNN-LSTM with attention architecture.

Video Vision Transformer

The Video Vision Transformer (ViViT) model [58], leverages a purely transformer-based architecture tailored for video classification tasks. This model primarily processes spatio-temporal tokens extracted from input videos, which are then processed by transformer layers to capture both spatial and temporal dynamics effectively Figure 3.8. To manage the extensive sequence of tokens typical in video data, the ViViT employs various efficient strategies to factorize the processing across spatial and temporal dimensions, enhancing both computational efficiency and scalability.

The core functionality of the ViViT lies in its use of self-attention mechanisms, which unlike traditional convolutional models, allow the system to handle long-range interactions directly without the limitations of a confined receptive field. This ability to attend to all parts of the input sequence universally allows ViViT to outperform existing deep 3D convolutional models on several benchmark datasets.

Moreover, the architecture’s flexibility is augmented by effective regularization strategies during training, which enable it to perform well even on smaller datasets by leveraging pretrained image models. This method of training compensates for the typical requirement of transformers for large amounts of data, adapting the model to practical applications where extensive video data may not be available. Additionally, ViViT’s performance on a range of video classification benchmarks sets new standards, particularly when the model is pretrained on large-scale image datasets, showcasing its robustness and adaptability to various video understanding tasks.

The hyperparameters of our ViViT model were optimized using Optuna. The best configuration identified through this tuning process is as follows:

- **Image Patch Size:** 8
- **Frame Patch Size:** 4
- **Embedding Dimension:** 128
- **Spatial Transformer Depth:** 4
- **Temporal Transformer Depth:** 4
- **Number of Attention Heads:** 2
- **MLP Dimension:** 2048

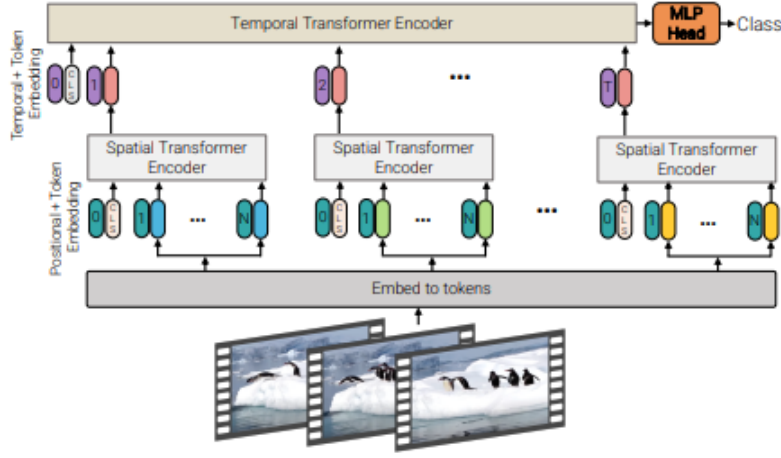


Figure 3.8: ViViT architecture [41].

Video Swin Transformer

The Video Swin Transformer [59] integrates a novel architectural approach to video recognition, leveraging the inherent spatiotemporal locality of videos by adapting Swin Transformer [60] principles to video tasks. This model capitalizes on the hierarchical and shifted window self-attention mechanisms, originally designed for images, and extends them to the video domain to process sequences effectively. By employing non-overlapping 3D windows that shift across layers, the Video Swin Transformer achieves significant reductions in computational requirements compared to global attention models. This method not only ensures a more efficient processing of video data but also maintains high accuracy by preserving important spatiotemporal relationships within the video content. This approach is further refined with a strong initialization from pre-trained image models, which accelerates the training process and enhances model performance on challenging video recognition benchmarks like Kinetics-400 and Kinetics-600, making it highly suitable for dynamic scene understanding.

For our implementation, and in light of the limited amount of data available, we employed the tiny version of the Video Swin Transformer using the PyTorch framework. Adhering to the principles of transfer learning, we utilized a pretrained version of the model on Kinetics-600 and subsequently fine-tuned it with our dataset.

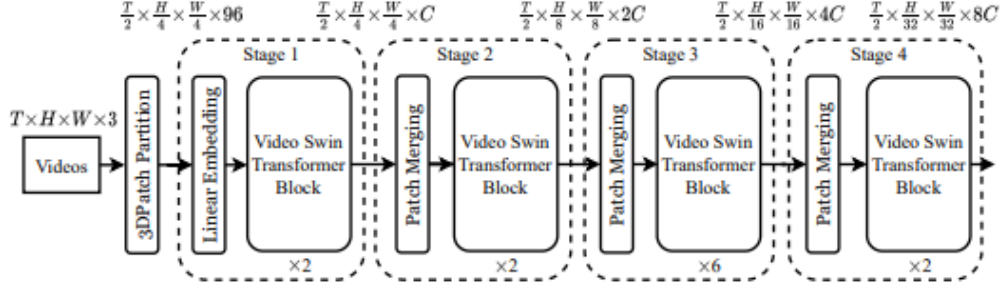


Figure 3.9: Video Swin Transformer architecture [42].

ML on Clinical data

For this implementation, we utilized explainable machine learning algorithms to ensure that the insights generated from the data are transparent and actionable for medical professionals. By focusing on models that provide clear decision paths, such as LightGBM, RandomForest, and XGBoost, we were able to analyze the clinical features with a high degree of interpretability. Additionally, we implemented an ensemble voting model that combines the strengths of these three algorithms (RandomForest, LightGBM, and XGBoost), allowing for a more robust and comprehensive analysis of the data. This ensemble approach enhances the reliability of predictions by aggregating the outputs of multiple models, thereby reducing the risk of overfitting and improving generalizability to unseen data. All models underwent hyperparameter tuning using the Optuna framework [61]. This process involved optimizing the parameters for each model to achieve the best possible performance, ensuring that the ensemble method benefits from models that are finely tuned to the dataset. The use of Optuna allowed for an efficient exploration of the hyperparameter space, leading to robust and well-calibrated models. The use of these models facilitates a collaborative environment, enabling clinicians to engage with the findings and validate the machine learning-driven insights in a clinical context.

3.4 Experimental Setup and Implementation

Various python versions compatible with the libraries used to build each model were utilized throughout the implementation and experimentation process, ensuring optimal performance and compatibility. The system -provided by Hellenic Mediterranean University- used for experimentation equipped with an AMD EPYC 7742 64-Core Processor, four NVIDIA A100 Tensor Core GPUs with 320 GB total GPU memory, and 1 TB of DDR4 ECC RAM. Both TensorFlow-GPU with the Keras-GPU frontend and PyTorch were

employed for developing, training, and evaluating the deep learning models, allowing for a comprehensive comparison of frameworks.

During the experiments, we tested three different loss functions: binary cross-entropy (BCE) for the binary classification part of the problem, sparse categorical cross-entropy (SCE), and focal loss for the multiclass classification tasks. For optimizing the models, we experimented with Adam, AdamW, and SGD optimizers, adjusting parameters to find the most effective combination.

Batch sizes varied across experiments, ranging from 4 to 32, to determine the impact on model performance. Additionally, we applied down-sampling techniques on video frames to reduce data volume while retaining critical information. We also, experimented with different video frame lengths, processing the first 30, 60, 90, up to 300 frames, corresponding to the full length of 10-second videos.

Chapter 4

Results

To assess the performance of our models, we employed the Leave-One-Group-Out (LOGO) validation method to evaluate the models' performance. The LOGO validation method provides a rigorous evaluation framework that reflects the model's ability to generalize to new, unseen patients. Given the use of 173 patients and leaving out groups of 15 patients for testing in each iteration, this approach ensures that the results are indicative of the model's performance in real-world clinical settings, where it must accurately diagnose new patients who were not included in the training process. Additionally, based on validation loss we utilized an early stopping callback function with a patience level set to 5, which prevented overfitting and improved all models' generalization capability. We also implemented a learning rate schedule using ReduceLROnPlateau with a monitoring metric of validation loss, reducing the learning rate by a factor of 0.5 after 3 epochs of no improvement, further optimizing all models performance.

Advantages:

- This approach ensures that the model is tested on completely unseen patients, providing a robust assessment of its generalization capabilities.
- By avoiding any overlap of patients between the training and test sets, we mitigate the risk of the model overfitting to specific individuals, leading to more reliable and realistic performance metrics.

4.1 Metrics

In our study, several metrics were employed to evaluate the performance of the models in detecting depression from facial video data. These metrics provide a comprehensive assessment of the models' accuracy, precision, recall, F1 score, and overall effectiveness. Below is a detailed description of each metric used:

Accuracy

Definition: Accuracy is the ratio of correctly predicted instances to the total number of instances. It measures the overall effectiveness of the model in correctly identifying both positive and negative instances.

Formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

Precision

Definition: Precision, also known as Positive Predictive Value, is the ratio of correctly predicted positive observations to the total predicted positives. It indicates how many of the instances predicted as positive are actually positive.

Formula:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.2)$$

Recall (Sensitivity or True Positive Rate)

Definition: Recall is the ratio of correctly predicted positive observations to all observations in the actual positive class. It measures the model's ability to correctly identify all relevant instances.

Formula:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.3)$$

F1 Score

Definition: The F1 Score is the harmonic mean of precision and recall, providing a single metric that balances the two. It is particularly useful when the class distribution is imbalanced, as it combines the benefits of both precision and recall.

Formula:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.4)$$

4.2 Results

In this section, we present the results of our comprehensive analysis, focusing on the performance of deep learning models trained on video data and machine learning models trained on clinical data. The outcomes provide insights into the efficacy of each approach in capturing and classifying relevant features associated with PD.

4.2.1 Deep learning approach using facial video data

The 3D-CNN model with LSTM and an attention layer emerged as the most effective approach in this study. Given the limited amount of available data, the ViVit and 3D-Swin tiny models were unable to fully demonstrate their potential capabilities. In contrast, the 3D-CNN model performed robustly, delivering strong results, while the frame-wise classification approach also yielded satisfactory performance. The final results are presented in Table 4.1 for binary classification and in Table 4.2 for multiclass classification.

Table 4.1: Final Results for binary classification

Model	Accuracy	Precision	Recall	F1-Score
3D-CNN	80%	80.02%	80.05%	80.058%
3D-CNN-LSTM with attention	82%	81.5%	83%	82%
ViVit	70.3%	69.1%	70.1%	69.6%
3D Swin tiny	76.5%	76.3%	77.5%	74.77%
Frame-wise	75.4%	75.9%	77.66%	75.28%

Table 4.2: Final Results for multiclass classification

Model	Accuracy	Precision	Recall	F1-Score
3D-CNN	68.9%	65.9%	67.6%	65.5%
3D-CNN-LSTM with attention	71.17%	66.79%	69.93%	66.48%
ViVit	65.2%	66.0%	64.0%	65%
3D Swin tiny	69.45%	66.1%	64.2%	65.1%
Frame-wise	66.4%	66.9%	67.66%	67.27%

More experiments and results are presented at Appendix A.

The optimal combination of hyperparameters for each model is presented in Table 4.3 for the binary task and Table 4.4 for the multiclass task.

Table 4.3: Best Hyperparameters for Binary Classification

Model	Epochs	Loss	Optimizer	lr	Batch Size	Input Frames	Dropout	Resolution
3D-CNN	50	Focal	Adam	1e-3	8	First 50	15%	224x224
3D-CNN-LSTM with attention	50	BCE	Adam	1e-3	8	First 50	15%	224x224
ViVit	50	BCE	Adam	1e-4	4	First 60	15%	224x224
3D Swin tiny	50	BCE	Adam	1e-4	8	2 frames per 30	20%	224x224
Frame-wise	50	BCE	Adam	1e-5	8	300	15%	220x220

Table 4.4: Best Hyperparameters for Multiclass Classification

Model	Epochs	Loss	Optimizer	lr	Batch Size	Input Frames	Dropout	Resolution
3D-CNN	50	SCE	Adam	1e-3	8	First 50	15%	224x224
3D-CNN-LSTM with attention	50	SCE	Adam	1e-3	8	First 32	15%	224x224
ViVit	50	SCE	Adam	1e-4	4	First 60	15%	224x224
3D Swin tiny	50	SCE	Adam	1e-4	8	2 frames per 30	20%	224x224
Frame-wise	50	SCE	Adam	1e-5	8	300	15%	220x220

In figure 4.1 we can see the best model’s training and validation curves given a LOGO fold. The training accuracy shows a consistent upward trend, indicating effective learning, while the training loss steadily decreases, reflecting good error minimization and convergence over time. This combination suggests the model is progressively improving its performance on the training data. The validation curves exhibit noticeable variability, with the validation accuracy generally increasing, indicating better generalization, but fluctuating significantly across epochs. Similarly, the validation loss shows large fluctuations, although it trends downward overall. This instability could suggest that the model’s performance on unseen data is inconsistent, possibly due to the inherent variability in the validation data. Additionally, the internal adjustments that are made by the camera to enhance video quality may be inadvertently contributing to these fluctuations by introducing subtle variations that the model struggles to generalize across as described in [62].

Another reason this instability might occur is due to the reduced and subtle facial movements characteristic of hypomimia itself. Since hypomimia results in less pronounced facial expressions, the model may find it difficult to detect and interpret the subtle cues necessary for accurate classification. This can lead to inconsistencies when the model is exposed to different levels of facial expressiveness in the validation data, causing it to perform well on some instances while struggling with others. This variability in expression poses a significant challenge for models relying on facial features, as the lack of distinct, expressive signals can make it harder for the model to generalize its learning to new, unseen data. Therefore, addressing the specific challenges posed by hypomimia is crucial for improving the robustness and reliability of the model in real-world applications.

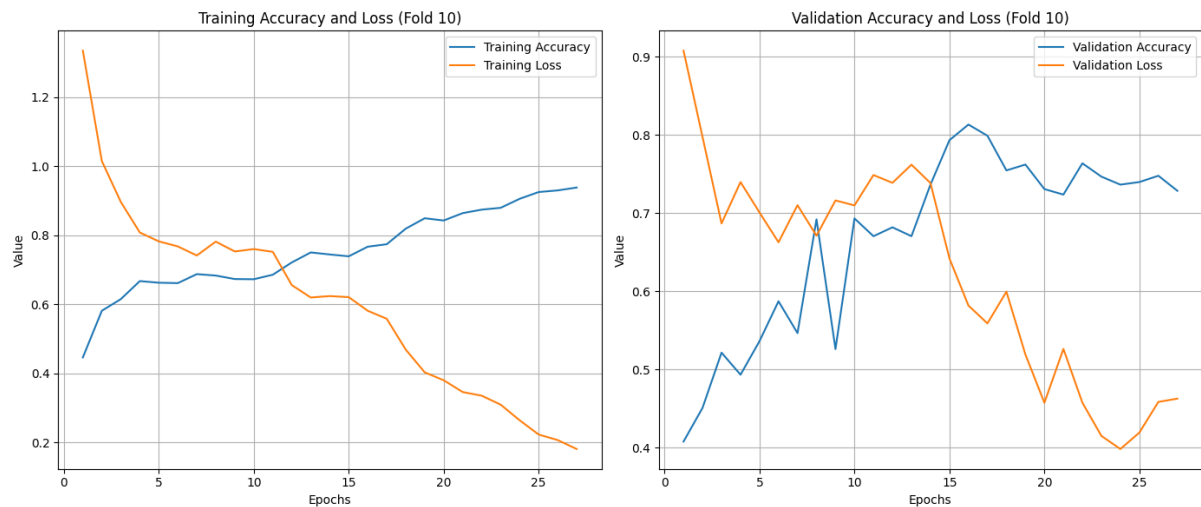


Figure 4.1: Accuracy & Loss for training and validation fold

4.2.2 Machine learning approach

Once again to assess the performance of our models, we employed the LOGO validation method to evaluate the models' performance. Our machine learning models demonstrated strong performance in both binary and multiclass depression estimation tasks, leveraging structured Clinical data to accurately differentiate between varying levels of depressive symptoms in PD patients. The models were able to capture key features associated with depression, resulting in high accuracy across different classification scenarios.

For binary classification, the models successfully distinguished between depressed and non-depressed patients, while in the multiclass setting, they effectively categorized patients into more granular levels of depression severity. The tables below summarize the accuracy statistics for both the binary and multiclass tasks, highlighting the robustness of our machine learning approach in this domain. The results of our models can be seen on Tables 4.5, 4.6.

Table 4.5: Binary results on Clinical Dataset

Model	Accuracy	Recall	Precision	F1-Score
Random Forest	85%	85%	86%	85.49%
LightGBM	88%	90%	88%	88.98%
XGBoost	87%	87%	89%	87.98%
Ensemble	88%	90%	88%	88.98%

Table 4.6: Multiclass results on Clinical Dataset

Model	Accuracy	Recall	Precision	F1-Score
Random Forest	79%	79%	83%	80.95%
LightGBM	86%	88%	87%	87.49%
XGBoost	85%	87%	85%	85.98%
Enseble	86%	88%	87%	87.49%

The ensemble model outperforms the other individual models, as demonstrated in both Binary and Multiclass Clasification. This improvement is expected because the ensemble method effectively combines the strengths of all the individual models at the decision level. By aggregating the predictions from Random Forest, LightGBM, and XGBoost, the ensemble model mitigates the weaknesses of each model while amplifying their strengths, leading to more robust and accurate predictions across the classification tasks.

In addition to evaluating the performance of our machine learning models, we employed Explainable AI (XAI) techniques to gain deeper insights into the decision-making processes of these models. By analyzing the key features (Figure 4.2) that contributed most significantly to the models' predictions, we were able to extract valuable information about the underlying factors associated with depression in PD patients. This interpretability is crucial for ensuring that the models' outputs are not only accurate but also understandable and actionable for clinicians. The insights derived from XAI provide a transparent view into how the models weigh different Clinical data inputs.

Notably, features such as UPDRS III scores, LEDD, and H&Y scale emerged as the most significant predictors across the models. The high importance of these features aligns with clinical understanding, where the severity and progression of PD are closely linked to the presence of depressive symptoms. This finding reinforces the validity of the models' predictions and highlights the potential of these features as critical indicators for clinical assessments.

The feature importance rankings, show a consistent pattern across both RandomForest and LightGBM models. However, some differences were observed in the weighting of specific features. For instance, while RandomForest placed greater emphasis on motor symptoms such as hypomimia and dyskinesia, LightGBM gave slightly more weight to medication-related features. These variations suggest that different models may capture distinct aspects of the clinical data, further justifying the use of an ensemble approach to leverage the strengths of each individual model.

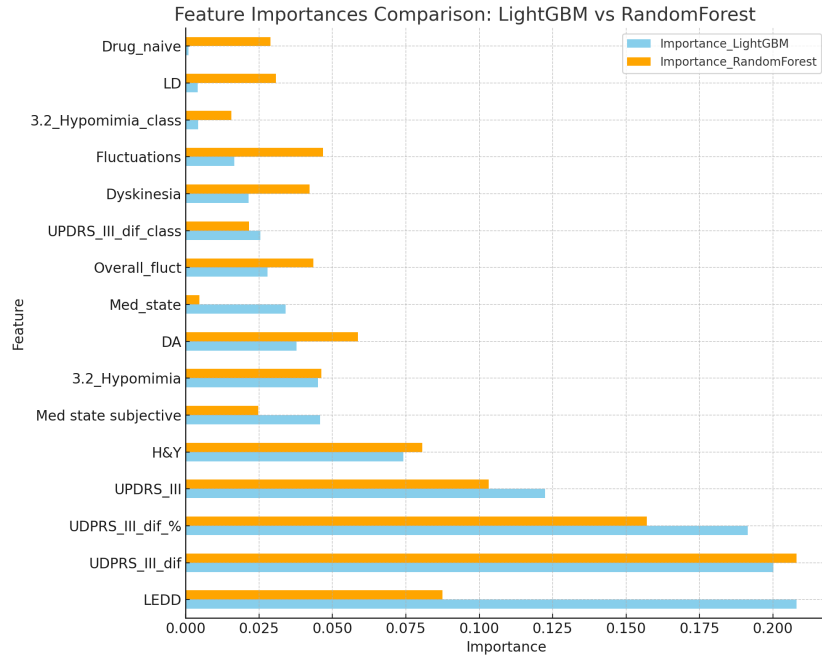


Figure 4.2: Feature importance for RandomForest and LightGBM based on the multiclass problem

These findings have significant clinical implications. By understanding which features are most influential in predicting depression, clinicians can better monitor these factors in PD patients, potentially leading to earlier identification and more targeted interventions for depression. The use of XAI techniques not only enhances the transparency of the models but also ensures that their application in clinical settings is both effective and trustworthy.

4.3 Comparison with state-of-the-art

Unfortunately, no prior work has been conducted on the estimation of depression in PD patients using similar methodologies, making it impossible to present a direct comparison with the state-of-the-art. Nevertheless, we are fortunate enough to establish a current state-of-the-art benchmark with our study.

Chapter 5

Discussion and Conclusion

5.1 Discussion

The results of our study open up a new avenue for estimating depression in PD patients using facial videos, which is an area that has not been explored before. Our best model showed high accuracy, precision, recall, and F1-score, indicating that it is possible to reliably detect depression in PD patients through facial analysis.

In addition to facial video analysis, we also conducted an in-depth analysis of clinical data using XAI (Explainable AI) machine learning models. By employing these models, we were able to identify and understand the key clinical features that contribute to the prediction of depression in PD patients. The integration of these clinical features, such as UPDRS scores, LEDD, and other PD-related metrics, provided further insights and validation for our video-based approach, highlighting the interplay between the clinical severity of PD and depressive symptoms.

Based on the feature importance analysis, we can make several key assumptions regarding the prediction of depression in patients:

- **Severity of PD Correlates with Depression:**
 - **UPDRS III Metrics and H&Y Scale:** The prominence of features related to the UPDRS III and the H&Y scale suggests that the severity and progression of PD are closely associated with the presence of depression. It can be assumed that as the clinical severity of Parkinson's increases, so does the likelihood or severity of depressive symptoms.

- **Motor and Non-Motor Symptoms Influence Depression:**

- **Motor Fluctuations, Hypomimia, Dyskinesia:** Features related to motor symptoms, such as fluctuations, hypomimia (reduced facial expressiveness), and dyskinesia (involuntary movements), are important predictors. This indicates that both motor impairments and their variability may contribute to depressive states, possibly due to the increased burden and reduced quality of life they impose on patients.
- **Subjective Medication State:** The importance of subjective assessments of medication state suggests that how patients perceive their treatment and its effects can influence their mental health. Negative perceptions or dissatisfaction with medication efficacy may be linked to higher depression rates.

- **Medication Dosage and Management Are Linked to Depression:**

- **LEDD:** The significant role of LEDD implies that the dosage and management of dopaminergic medications are related to depressive symptoms. Higher or imbalanced dosages might contribute to mood alterations, or conversely, depression might affect medication adherence and effectiveness.

- **Subjective Well-Being and Perceptions Are Critical:**

- **Med State Subjective:** Features capturing patients' subjective experiences and perceptions of their medication state highlight the importance of psychological and subjective well-being in predicting depression. This suggests that patients' self-reported experiences and mental states are valuable indicators of their depressive status.

- **Interactions Between Clinical Features Affect Depression:**

- The effectiveness of ensemble models, which combine multiple algorithms, indicates that interactions between various clinical features play a significant role in predicting depression. It can be assumed that depression in patients is influenced by a complex interplay of multiple factors rather than isolated features.

While the XAI-driven insights provide valuable understanding, it is important to acknowledge the limitations of this approach. For instance, the reliance on the specific features available in the dataset

may limit the generalizability of the findings to broader populations. Future work could explore the inclusion of additional data modalities, such as genetic markers or neuroimaging, to further enhance the predictive power and interpretability of the models.

One of the significant challenges we faced was dealing with hypomimia, a common symptom in PD that reduces facial expressiveness. This symptom can obscure subtle facial cues related to emotional states, making it more difficult for automated systems to accurately assess depression. Despite this challenge, our model was able to capture relevant features that contributed to the assessment of depression.

Another challenge was the frequent and involuntary movements caused by Parkinson’s motor symptoms. These movements often disrupted the stability of the facial region in video frames, making it harder for our models to focus on the subtle expressions related to depression. Overcoming these obstacles required careful preprocessing and robust model design to ensure that the system could still perform effectively under these challenging conditions.

Distinguishing between mild and severe depression also proved difficult. The subtle differences in facial expressions and behaviors between these two states often led to overlapping features, complicating the classification process. This difficulty highlights the need for further refinement of our model and potentially integrating additional modalities or data sources to improve the sensitivity and specificity of depression detection across different severity levels. Additionally, creating videos specifically designed for depression estimation, with targeted prompts and scenarios that elicit relevant emotional responses, could further enhance the accuracy and reliability of our model’s predictions. Such tailored video content could help capture more distinct and informative facial cues, ultimately leading to better differentiation between varying levels of depression severity in PD patients.

While our study has successfully demonstrated the potential of using facial video analysis to estimate depression in PD patients, there are several areas for future work that could further enhance the accuracy and robustness of our models.

One promising direction is to integrate additional modalities beyond facial video data. For example, incorporating speech analysis, body movement data, or even physiological signals such as heart rate variability could provide a more comprehensive view of the patient’s emotional and physical state. This multi-modal approach could help capture aspects of depression that are not as easily seen in facial expressions alone, leading to more accurate assessments.

Further pre-processing of the videos before introducing them into the models could enhance the quality of input data. Techniques such as noise reduction, and stabilization could mitigate potential artifacts and variability, thus improving the models’ performance.

Another area for future exploration involves using the facial landmarks we have already extracted using MTCNN. These precise landmarks could be used for a finer analysis of facial movements and expressions. By tracking subtle changes in facial muscle activity associated with emotional states, these landmarks could further enhance our model's ability to detect depression.

Furthermore, optical flow algorithms could be used to analyze the motion patterns in video frames, capturing how facial expressions change over time. Understanding how these movements evolve could help distinguish between different levels of depression severity, especially in cases where traditional static analysis might struggle. Expanding the scope of predictions to include other mental health conditions such as stress, anxiety, and emotional dysregulation could also provide a more holistic assessment of the patient's psychological well-being.

Incorporating these advanced techniques, additional data sources, and expanding predictions could significantly improve the sensitivity and specificity of mental health detection models. Moreover, exploring the use of foundation models trained on diverse datasets with specific transfer learning to adapt them for PD-related applications could further enhance the generalizability and effectiveness of these tools. This comprehensive approach would not only refine the current methods but also solidify the viability of non-invasive, video-based tools for mental health assessment in PD patients, leading to better outcomes in clinical applications.

Overall, our work shows the potential of using facial video analysis as a non-invasive and accessible tool for assessing depression in PD patients, offering a promising direction for future research and clinical applications. Moreover, the analysis of clinical data showed that key clinical features, such as UPDRS scores and LEDD, play a significant role in predicting depression in PD patients. This reinforces the importance of considering both facial and clinical data in future studies to develop more comprehensive and accurate models for mental health assessment. Together, these findings highlight the value of integrating multiple approaches to enhance the understanding and management of depression in PD patients.

5.2 Conclusion

This research has made significant strides in addressing the challenging task of estimating depression in PD patients through the use of facial video analysis and comprehensive clinical data analysis. Leveraging a novel dataset and advanced deep learning models, we have demonstrated that it is possible to reliably detect depression in this specific population, despite the complexities introduced by symptoms such as hypomimia.

Our approach utilized a combination of state-of-the-art deep learning techniques, including the tiny version of the Video Swin Transformer, 3D-CNN models, and a 3D-CNN-LSTM with an attention layer, ViVit, and an implementation of Google-Net with a 1d-transformer. Among these, the 3D-CNN-LSTM model with an attention mechanism emerged as the most effective, achieving high accuracy, precision, recall, and F1 scores. This indicates the model's robust capability to handle the intricacies of facial expressions in PD patients, even with limited data.

In addition to the video-based analysis, we conducted a detailed evaluation of clinical data using machine learning models and XAI techniques. This dual approach allowed us to identify key clinical features, such as UPDRS III scores, LEDD, and the H&Y scale, that significantly contribute to depression in PD patients. The integration of clinical data provided a more holistic understanding of the factors influencing depression, enhancing the interpretability and reliability of our predictions.

A key achievement of this work is the establishment of a new state-of-the-art (SOTA) in the field of depression estimation in PD patients. Prior to this study, there was a notable lack of research focused on this specific intersection of PD and depression, particularly using facial video data combined with clinical data analysis. By developing and fine-tuning models that are tailored to this unique application, we have laid the groundwork for future research and clinical applications, effectively creating a new area of study within the domain of PD.

In summary, this study not only advances the technological capabilities in assessing depression in PD patients but also highlights the importance of integrating diverse data sources, such as facial video and clinical metrics, to provide a more comprehensive and reliable tool. This non-invasive approach offers significant potential to improve the quality of life for individuals affected by PD and opens up new avenues for research in this critical area of healthcare.

Appendix A

The first appendix

Supplementary Metrics for 3D-CNN

Table A.1: Experiments on Binary Classification 3D-CNN

Epochs	Loss	Opt	lr	Batch Size	Frames	Dropout	Res	Acc
50	BCE	Adam	1e-3	4	First 120	15%	128x128	71%
50	BCE	Adam	1e-4	8	2 frames per 30	15%	128x128	69%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	224x224	76%
50	BCE	AdamW	1e-3	8	First 50	15%	224x224	75%
50	BCE	SDG	1e-4	16	First 30	15%	224x224	74%
50	Focal	Adam	1e-3	8	1 frame per 30	15%	224x224	77%

Table A.2: Experiments on Multiclass Classification 3D-CNN

Epochs	Loss	Opt	lr	Batch Size	Frames	Dropout	Res	Acc
50	SCE	Adam	1e-3	4	First 120	15%	128x128	60.5%
50	SCE	Adam	1e-4	8	2 frames per 30	15%	128x128	56%
50	Kullback Leibler	Adam	1e-4	4	2 frames per 30	20%	224x224	63.6%
50		AdamW	1e-3	8	First 50	15%	224x224	62.22%
50	SCE	SDG	1e-4	16	First 30	15%	224x224	61.11%
50	SCE	Adam	1e-3	8	1 frame per 30	15%	224x224	65.57%

Supplementary Metrics for 3D-CNN-LSTM with attention

Table A.3: Experiments on Binary Classification 3D-CNN-LSTM with attention

Epochs	Loss	Opt	lr	Batch Size	Frames	Dropout	Res	Acc
50	BCE	Adam	1e-3	4	First 120	15%	128x128	75.59%
50	BCE	Adam	1e-4	8	2 frames per 30	15%	128x128	71.16%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	224x224	76%
50	BCE	AdamW	1e-3	16	First 30	15%	224x224	76.2%
50	BCE	SDG	1e-4	8	First 50	15%	224x224	75.9%
50	Focal	Adam	1e-3	8	1 frame per 30	15%	224x224	78.85%

Table A.4: Experiments on Multiclass Classification 3D-CNN-LSTM with attention

Epochs	Loss	Opt	lr	Batch Size	Frames	Dropout	Res	Acc
50	SCE	Adam	1e-3	4	First 120	15%	128x128	62.6%
50	SCE	Adam	1e-4	8	2 frames per 30	15%	128x128	60.1%
50	Kullback Leibler	Adam	1e-4	4	2 frames per 30	20%	224x224	65.9%
50	SCE	AdamW	1e-3	8	First 50	15%	224x224	65.22%
50	SCE	SDG	1e-4	16	First 30	15%	224x224	64.11%
50	SCE	Adam	1e-3	8	1 frame per 30	15%	224x224	68.57%

Supplementary Metrics for ViVit

Table A.5: Experiments on Binary Classification ViVit

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	BCE	Adam	1e-3	4	First 120	15%	73%
50	BCE	Adam	1e-4	8	2 frames per 30	15%	71.12%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	72%
50	BCE	AdamW	1e-3	16	First 30	15%	70%
50	BCE	SDG	1e-4	8	First 50	15%	70.66%
50	Focal	Adam	1e-3	8	1 frame per 30	15%	69.77%

Table A.6: Experiments on Multiclass Classification ViVit

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	SCE	Adam	1e-3	4	First 120	15%	65.62%
50	SCE	Adam	1e-4	8	2 frames per 30	15%	63%
50	Kullback Leibler	Adam	1e-4	4	2 frames per 30	20%	63.3%
50	SCE	AdamW	1e-3	16	First 30	15%	62%
50	SCE	SDG	1e-4	8	First 50	15%	62.39%
50	SCE	Adam	1e-3	8	1 frame per 30	15%	61%

Supplementary Metrics for Video Swin

Table A.7: Experiments on Binary Classification Video Swin tiny

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	BCE	Adam	1e-3	4	First 120	15%	71%
50	BCE	Adam	1e-4	8	2 frames per 30	15%	73.1%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	74.42%
50	BCE	AdamW	1e-3	16	First 30	15%	73.91%
50	BCE	SDG	1e-4	8	First 50	15%	75.55%
50	Focal	Adam	1e-3	8	1 frame per 30	15%	74.44%

Table A.8: Experiments on Multiclass Classification Video Swin tiny

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	SCE	Adam	1e-3	4	First 120	15%	66.9%
50	SCE	Adam	1e-4	8	2 frames per 30	15%	65.2%
50	Kullback Leibler	Adam	1e-4	4	2 frames per 30	20%	66.9%
50	SCE	AdamW	1e-3	16	First 30	15%	66%
50	SCE	SDG	1e-4	8	First 50	15%	66.85%
50	SCE	Adam	1e-3	8	1 frame per 30	15%	67.72%

Supplementary Metrics for Frame-Wise

Table A.9: Experiments on Binary Classification Video Frame-wise

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	BCE	Adam	1e-3	4	First 150	15%	71%
50	BCE	Adam	1e-4	8	5 frames per 30	15%	73.1%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	74.42%
50	BCE	AdamW	1e-3	16	Last 150	15%	73.91%
50	BCE	SDG	1e-4	8	10 frames per 30	15%	75.55%
50	Focal	Adam	1e-3	8	300	15%	74.44%

Table A.10: Experiments on Multiclass Classification Frame-wise

Epochs	Loss	Opt	lr	Batch	Frames	Dropout	Acc
50	BCE	Adam	1e-3	4	First 150	15%	71%
50	BCE	Adam	1e-4	8	5 frames per 30	15%	73.1%
50	Focal	Adam	1e-4	4	2 frames per 30	20%	74.42%
50	BCE	AdamW	1e-3	16	Last 150	15%	73.91%
50	BCE	SDG	1e-4	8	10 frames per 30	15%	75.55%
50	Focal	Adam	1e-3	8	300	15%	74.44%

Supplementary Metrics for ML on Clinical Data

Table A.11: Binary results on Clinical Dataset with Grid Search tuner

Model	Accuracy	Recall	Precision	F1-Score
Random Forest	85%	85%	86%	85.49%
LightGBM	84%	83%	85%	83.98%
XGBoost	81%	79%	78%	78.49%
Ensemble	86%	86%	85%	85.49%

Table A.12: Multiclass results on Clinical Dataset with Grid Search tuner

Model	Accuracy	Recall	Precision	F1-Score
Random Forest	79%	79%	83%	80.95%
LightGBM	80%	80%	82%	80.98%
XGBoost	77%	77%	80%	78.47%
Ensemble	80%	80%	83%	81.47%

Bibliography

- [1] L. Marsili, G. Rizzo, and C. Colosimo. Diagnostic criteria for parkinson's disease: From james parkinson to the concept of prodromal disease. *Frontiers in Neurology*, 9:156, 2018.
- [2] L. Stefanis. α -synuclein in parkinson's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(2):a009399, 2012.
- [3] E. M. Kline, M. C. Houser, M. K. Herrick, P. Seibler, C. Klein, A. West, and M. G. Tansey. Genetic and environmental factors in parkinson's disease converge on immune function and inflammation. *Movement Disorders: Official Journal of the Movement Disorder Society*, 36(1):25–36, 2021.
- [4] M. Cruts K. Nuytemans, J. Theuns and C. Van Broeckhoven. Genetic etiology of parkinson disease associated with mutations in the snca, park2, pink1, park7, and lrrk2 genes: A mutation update. *Human Mutation*, 31(7):763–780, 2010.
- [5] T. Subramanian D. Nemade and V. Shivkumar. An update on medical and surgical treatments of parkinson's disease. *Aging and Disease*, 12(4):1021–1035, 2021.
- [6] K. Han D. W. Ryu and A. H. Cho. Mortality and causes of death in patients with parkinson's disease: A nationwide population-based cohort study. *Frontiers in Neurology*, 14:1236296, 2023.
- [7] J. Jankovic. Parkinson's disease: Clinical features and diagnosis. *Journal of Neurology, Neurosurgery, and Psychiatry*, 79(4):368–376, 2008.
- [8] M. Alborghetti M. Simonelli F. D'Audino C. Onelli E. Pegolo E. Bianchini, D. Rinaldi and F. E. Pontieri. The story behind the mask: A narrative review on hypomimia in parkinson's disease. *Brain Sciences*, 14(1):109, 2024.
- [9] H. A. González-Usigli, G. G. Ortiz, C. Charles-Niño, M. A. Mireles-Ramírez, F. P. Pacheco-Moisés, B. M. G. Torres-Mendoza, J. J. Hernández-Cruz, D. L. D. C. Delgado-Lara, and L. J. Ramírez-

- Jirano. Neurocognitive psychiatric and neuropsychological alterations in parkinson's disease: A basic and clinical approach. *Brain Sciences*, 13(3):508, 2023.
- [10] G. Li Z. Chen and J. Liu. Autonomic dysfunction in parkinson's disease: Implications for pathophysiology, diagnosis, and treatment. *Neurobiology of Disease*, 134:104700, 2020.
- [11] A. Albishi A. Al-Onazi S. Aseeri F. Alotaibi Y. Almazroua N. Al-Khammash, N. Al-Jabri and M. Albloushi. Quality of life in patients with parkinson's disease: A cross-sectional study. *Cureus*, 15(1):e33989, 2023.
- [12] G. M. Earhart et al. L. E. Dibble, J. T. Cavanaugh. Charting the progression of disability in parkinson disease: study protocol for a prospective longitudinal cohort study. *BMC Neurology*, 10:110, 2010.
- [13] T. Rajabally P. Vinayakray-Jani J. Gandhi, A. Gadekar and D. Ambawade. Detection of parkinson's disease via a multi-modal approach. In *Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, July 2021.
- [14] S. R. Shaftman G. T. Stebbins S. Fahn P. Martinez-Martin W. Poewe C. Sampaio M. B. Stern R. Dodel B. Dubois R. Holloway J. Jankovic J. Kulisevsky A. E. Lang A. Lees S. Leurgans P. A. LeWitt D. Nyenhuis C. W. Olanow et al. C. G. Goetz, B. C. Tilley. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results. *Movement Disorders: Official Journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.
- [15] M. M. Hoehn and M. D. Yahr. Parkinsonism. *Neurology*, 17(5):427–427, 1967.
- [16] C. Tremblay R. B. Postuma A. Dagher R. D. Markello, G. Shafiei and B. Misic. Multimodal phenotypic axes of parkinson's disease. *npj Parkinson's Disease*, 7(1):6, 2021.
- [17] A. Pepe K. L. Pomykala F. Jonske M. Kurz J. Li J. Egger, C. Gsaxner and J. Kleesiek. Medical deep learning-a systematic meta-review. *Computer Methods and Programs in Biomedicine*, 221:106874, 2022.
- [18] A. J. Humaidi A. Al-Dujaili Y. Duan O. Al-Shamma J. Santamaría M. A. Fadhel M. Al-Amidie L. Alzubaidi, J. Zhang and L. Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021.

-
- [19] L. Klingelhoefer S. Bostantjopoulou K. R. Chaudhuri K. Kyritsis S. Hadjidimitriou V. Charisis L. J. Hadjileontiadis A. Papadopoulos, D. Iakovakis and A. Delopoulos. Unobtrusive detection of parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific Reports*, 10(1):21370, 2020.
- [20] F. Amato I. Rechichi C. Ramos-Romero A. Cárdenas L. Gascó L. Sigcha, L. Borzi and G. Olmo. Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review. *Expert Systems with Applications*, 229:120541, 2023.
- [21] C. Spanaki E. Michou G. Karamanis-Z. Kefalopoulou V. Skaramagkas, I. Boura and M. Tsiknakis. Detecting minor symptoms of parkinson's disease in the wild using bi-lstm with attention mechanism. *Sensors*, 23:7850, 2023.
- [22] L. Chen J.-Y. Lu-J.-J. Ge F.-T. Liu J.-T. Yu W. Lin C.-T. Zuo Y. Wu, J.-H. Jiang and J. Wang. Use of radiomic features and support vector machine to distinguish parkinson's disease cases from normal controls. *Annals of Translational Medicine*, 7(23), 2019.
- [23] G. Zhang H. Wang-Y. C. Chen-Y. Liu C. G. Tarolli D. Crepeau J. Bukartyk M. R. Junna A. Vidnovic T. D. Ellis M. C. Lipford R. Dorsey Y. Yang, Y. Yuan and D. Katabi. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature Medicine*, 28(10):2207–2215, 2022.
- [24] X. Li Y. Fan-W. Zeng Y. Tan K. Ren W. Chen R. Lu, Y. Xu and X. Cao. Evaluation of wearable sensor devices in parkinson's disease: A review of current status and future prospects. *Parkinson's Disease*, page 4693019, 2020.
- [25] J. Howard, V. Murashov, E. Cauda, and J. Snawder. Advanced sensor technologies and the future of work. *American Journal of Industrial Medicine*, 65(1):3–11, 2022.
- [26] E. Du-M. A. Leavitt S. M. A. Iqbal, I. Mahgoub and W. Asghar. Advances in healthcare wearable devices. *npj Flexible Electronics*, 5(1):9, 2021.
- [27] X. Ye R. Liu and T. Cui. Recent progress of biomarker detection sensors. *Research (Washington, D.C.)*, page 7949037, 2020.
- [28] M. Saint-Hilaire-C. A. Thomas S. D. Gunnery, B. Habermann and L. Tickle-Degnen. The relationship between the experience of hypomimia and social wellbeing in people with parkinson's disease and their care partners. *Journal of Parkinson's Disease*, 6(3):625–630, 2016.

-
- [29] Parkinson's Foundation. Depression. <https://www.parkinson.org/understanding-parkinsons/non-movement-symptoms/depression>. Accessed: 2024-08-01.
- [30] Y. Elserogy-A. F. Zaki E. M. Khedr, A. A. Abdelrahman and A. Gamea. Depression and anxiety among patients with parkinson's disease: frequency, risk factors, and impact on quality of life. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, 56(1):116, 2020.
- [31] M. Kanimozhi N. Kumar S. Samanta, M. Bisht and S. S. Handu. Association of depression with disease duration, quality of life and adherence in parkinson's disease: A cross sectional study. *Journal of Family Medicine and Primary Care*, 12(7):1406–1411, 2023.
- [32] H. D. Weiss and G. M. Pontone. "pseudo-syndromes" associated with parkinson disease, dementia, apathy, anxiety, and depression. *Neurology: Clinical Practice*, 9(4):354–359, 2019.
- [33] S. Neargarder U. S. Clark and A. Cronin-Golomb. Specific impairments in the recognition of emotional facial expressions in parkinson's disease. *Neuropsychologia*, 46(9):2300–2309, 2008.
- [34] J. L. Cummings. Depression and parkinson's disease: A review. *The American Journal of Psychiatry*, 149(4):443–454, 1992.
- [35] L. Li, Z. Wang, Z. You, and J. Huang. Prevalence and influencing factors of depression in patients with parkinson's disease. *Alpha Psychiatry*, 24(6):234–238, 2023.
- [36] Z. Goodarzi and Z. Ismail. A practical approach to detection and treatment of depression in parkinson disease and dementia. *Neurology Clinical Practice*, 7(2):128–140, 2017.
- [37] L. M. Shulman, R. L. Taback, A. A. Rabinstein, and W. J. Weiner. Non-recognition of depression and other non-motor symptoms in parkinson's disease. *Parkinsonism & Related Disorders*, 8(3):193–197, 2002.
- [38] Y. Ozkanca, M. G. Öztürk, M. N. Ekmekci, D. C. Atkins, C. Demiroglu, and R. H. Ghomi. Depression screening from voice samples of patients affected by parkinson's disease. *Digital Biomarkers*, 3(2):72–82, 2019.
- [39] Z. Soumaya, B. Drissi Taoufiq, B. Nsiri, Y. Korkmaz, and A. Ammoumou. The detection of parkinson disease using the genetic algorithm and svm classifier. *Applied Acoustics*, 171:107528, 2021.
- [40] H. Byeon. Development of a depression in parkinson's disease prediction model using machine learning. *World Journal of Psychiatry*, 10(10):234–244, 2020.

-
- [41] H. Byeon. Development of a stacking-based ensemble machine learning for detection of depression in parkinson's disease: Preliminary research. *Biology and Life Sciences Forum*, 9(1):5, 2021.
- [42] J. Mifsud, K. Embry, R. Macaluso, L. Lonini, R. Cotton, T. Simuni, and A. Jayaraman. Detecting the symptoms of parkinson's disease with non-standard video. *Journal of NeuroEngineering and Rehabilitation*, 21, 2024.
- [43] B. Jin, Y. Qu, L. Zhang, and Z. Gao. Diagnosing parkinson disease through facial expression recognition: Video analysis. *Journal of Medical Internet Research*, 22(7):e18697, 2020.
- [44] Z. Yin, V. J. Geraedts, Z. Wang, M. F. Contarino, H. Dibeklioglu, and J. van Gemert. Assessment of parkinson's disease severity from videos using deep architectures. *IEEE Journal of Biomedical and Health Informatics*, 26(3):1164–1176, 2022.
- [45] W. Xie et al. Interpreting depression from question-wise long-term video recording of sds evaluation. *IEEE Journal of Biomedical and Health Informatics*, 26(2):865–875, 2022.
- [46] A. Parikh, M. Sadeghi, and B. Eskofier. Exploring facial biomarkers for depression through temporal analysis of action units. *arXiv preprint arXiv:2407.13753*, 2024.
- [47] Z. Shangguan, Z. Liu, G. Li, Q. Chen, Z. Ding, and B. Hu. Dual-stream multiple instance learning for depression detection with facial expression videos. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:554–563, 2023.
- [48] T. L. Rose O. Lum V. Huang M. Adey J. A. Yesavage, T. L. Brink and V. O. Leirer. Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research*, 17(1):37–49, 1982.
- [49] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23, 2016.
- [50] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, Los Alamitos, CA, USA, June 2015. IEEE Computer Society.
- [51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, Boston, MA, USA, 2015. IEEE.

-
- [52] M. Yang S. Ji, W. Xu and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.
- [53] A. Yenter and A. Verma. Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis. In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pages 540–546, 2017.
- [54] M. Habaebi M. Halbouni M. Kartiwi A. Ha, T. Gunawan and R. Ahmad. Cnn-lstm: Hybrid deep neural network for network intrusion detection system. *IEEE Access*, PP:1–1, 2022.
- [55] N. T. Nguyen N. T. Nguyen, D. Q. Tran and H. Q. Nguyen. A cnn-lstm architecture for detection of intracranial hemorrhage on ct scans. *ArXiv*, abs/2005.10992, 2020.
- [56] J. You and J. Korhonen. Deep neural networks for no-reference video quality assessment. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2349–2353, 2019.
- [57] S. Ahmed M. Z. Uddin S. A. Alyami S. Al-Ashhab H. F. Akhdar A. Khan A. Azad M. A. Khatun, M. A. Yousuf and M. A. Moni. Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:2700316, 2022.
- [58] G. Heigold C. Sun M. Lucic A. Arnab, M. Dehghani and C. Schmid. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021.
- [59] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2022.
- [60] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.
- [61] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

-
- [62] S. Paul, K. Rao, G. Coviello, M. Sankaradas, O. Po, Y. C. Hu, and S. Chakradhar. Why is the video analytics accuracy fluctuating, and what can we do about it? In *Computer Vision – ECCV 2022 Workshops*, pages 430–448, Cham, 2023. Springer Nature Switzerland.