Multiple-Fine-Tuned Convolutional Neural Networks for Parkinson's Disease Diagnosis From Offline Handwriting

Matej Gazda, Máté Hireš, and Peter Drotár, *Member, IEEE*

***Abstract*—Existing decision support system frameworks for diagnosing Parkinson's disease (PD) through handwriting, speech, or gait characteristics share very similar pipelines. Although in some cases, patient data can be captured by commercially available devices, specialized devices or even custom-made prototypes are often required for such tasks. Captured data are used for extracting features that are carefully designed on the basis of domain and problem knowledge. These features are then fed to classifiers that provide a final decision. In this article, we present an approach in which end-to-end processing by a convolutional neural network (CNN) is utilized to diagnose PD from handwriting images, without the use of additional signals. This eliminates any need for specialized devices or feature engineering. To improve the performance of the proposed pretrained CNN, we propose the idea of multiple fine tuning to bridge the gap between semantically different source and target datasets and facilitate more efficient transfer learning. The proposed architecture, which is based on multiple fine tuning and an ensemble of multiple-fine-tuned CNNs, achieves 94.7% accuracy in the classification of PD from offline handwriting.**

***Index Terms*—Convolutional neural network (CNN), decision support system, fine tuning, handwriting, Parkinson's disease (PD), transfer learning (TL).**

I. Introduction

PARKINSON'S disease (PD) is the second-most common chronic progressive neurodegenerative disorder, and is characterized by motor symptoms and various psychiatric manifestations. PD is caused by a loss of neurons that produce dopamine, which functions as a neurotransmitter. Dopamine contributes to smooth, coordinated movement, and the functioning of muscles. Dopamine deficiency creates biochemical imbalances that result in the poor balance and poor motor coordination that characterize PD. Typical motor symptoms include resting tremors, rigidity, postural instability, and bradykinesia, i.e., slowness of the spontaneous movement [1]. These symptoms, together with psychiatric symptoms, negatively impact patients lives in numerous aspects and increase the risk of falls and further health complications.

Unfortunately, there is no definitive test for the diagnosis of PD; it must be diagnosed based on clinical criteria. The diagnosis accuracy reported in the literature ranges from 75% to 92%, depending on the expertise of the examiner [2], [3]. In addition to the relatively low diagnosis accuracy, the cost of diagnosis is quite high because patients are usually examined by specialists, such as neurologists or geriatricians. The high cost and low accuracy of diagnosis calls for an objective diagnostic approach that can support the current clinical evaluation.

Computational approaches to support PD diagnosis and rehabilitation are very active areas of research. Several modalities can be used to support PD diagnosis and evaluation. Freezing of the gait (FoG) is one of the most typical symptoms and has been closely associated with PD since its early description by James Parkinson. Most diagnostic system approaches rely on different types of sensors and machine learning methods to detect FoG [4]–[6]. However, it is quite difficult for clinicians to review and analyze sensor data; therefore, the video of patient movements is often recorded along with sensor data. Recent advances in video processing have facilitated the extension of automatic FoG detection methods using video data [7], [8].

Most patients suffering from PD experience voice impairments, such as reduced loudness, breathiness, hoarseness, imprecise articulation, and vocal tremor. As such, vocal measurements can also be an effective diagnostic and monitoring tool. Because speech processing is a popular and mature domain, there are numerous papers that examine different aspects of speech in PD [9]–[12]. The reported classification accuracy is as high as 90%; however, the resulting accuracy depends on the dataset under investigation [10]. Less prevalent approaches supporting the evaluation and treatment of PD utilize surface electromyography [13] and touch-screen devices [14], [15]. Finally, an approach based on handwriting is receiving significant attention because it does not require expensive equipment, tedious preprocessing, or annotation, and can be administered within a short period of time [10], [16], [17].

Handwriting analysis for the detection of pathological handwriting is based on rich knowledge and experience that

Manuscript received April 9, 2020; revised July 27, 2020, September 18, 2020, and October 30, 2020; accepted December 29, 2020. Date of publication January 18, 2021; date of current version December 17, 2021. This work was supported in part by the Slovak Research and Development Agency under Contract APVV-16-0211; in part by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic; and in part by the Slovak Academy of Sciences under Contract VEGA 1/0327/20. This article was recommended by Associate Editor X. Wang. *(Corresponding author: Peter Drotár.)*

The authors are with the Intelligent Information Systems Laboratory, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, 04201 Kosice, Slovakia (e-mail: peter.drotar@tuke.sk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2020.3048892.

Digital Object Identifier 10.1109/TSMC.2020.3048892

emerged from several decades of research on handwriting recognition. Recognizable scripts started with Arabic numerals and expanded to the Latin alphabets [18]. Later advances also brought the recognition of Japanese Katakana and Kanji characters, Chinese characters, and Indian and Arabic scripts [19]. Currently, the main challenges of handwriting recognition are represented by scripts (such as Gurmukhi) that are not so explored as Latin or oriental scripts [20], [21]. The most recent studies report over 90% accuracy for Indic and Chinese text recognition [22]–[24] with multilanguage systems being intensively developed [25]. There are two approaches to recognition: 1) online and 2) offline handwriting. For the online approach, the 2-D coordinates of successive points of writing as a function of time are available. The newest devices can also provide additional information, such as pressure, tilt, etc. On the other hand, the offline handwriting is available only as a completed image. Whether it is online or offline handwriting, the most recent approaches mostly employ the hidden Markov models [26], [27], and recurrent neural networks [28]–[30] that currently achieve state-of-the-art performance. In contrast to handwriting recognition, where a sufficient amount of data is available to train the neural networks, data in the biomedical domain are often limited to a few hundred samples or even less.

Utilizing neural networks for small sample domains became feasible with recent methodological breakthroughs and outreach from the transfer learning (TL) domain to the deep learning domain. Because datasets in small sample domains do not contain sufficient amounts of data, they are not suitable for training neural networks from scratch. Here, TL facilitates the transfer of knowledge gained in one domain to improve the prediction accuracy in another domain. Several methods may be considered during TL depending on the sizes of, and semantical similarities between, the source and target datasets. The general approach is to train a source network and then copy its lower layers to the lower layers of the target network. There is no deterministic method of selecting the layers that should be copied, apart from running extensive experiments with different numbers of layers transferred. The transferred layers can be fine tuned using backpropagation errors from the target task or they can be frozen (and therefore not affected by the target task). Although deep TL has shown significant potential and was successfully utilized with several small sample domains, such as [31] and [32], it is still considered to be a black-box model. There is no clear definition specifying which features should be transferred, or explaining the relationship between the source and target datasets; it is only known that the source and target datasets should be semantically similar. Unfortunately, there is a wide range of domains that are characterized by very specific image semantics, with significant differences between the source domain (mostly large datasets of general images) and a particular target domain.

This was the motivation for proposing novel multiple-fine-tuning approach using a mediator dataset, where a network is first trained on a source dataset and then further fine tuned using a carefully selected mediator dataset, and finally fine tuned using target task data. We show that this approach positively impacts the prediction accuracy of convolutional neural networks (CNNs), and that a multiple-fine-tuned CNN can outperform traditional fine-tuned CNNs in terms of pathological handwriting classification. Although this new approach is specific to the task of offline handwriting evaluation, its potential usability is much wider. In general, multiple fine tuning can be used in every domain in which there exists a dataset that can be, from a semantical similarity point of view, placed between the source and target datasets. For example, consider the task of detecting an object, such as a power line insulator, from aerial images [33]. Using multiple fine tuning, a network pretrained on some large dataset, such as ImageNet [34] or MIT-Places [35], can be utilized and further fine tuned by a dataset of aerial images, which would act as the mediator dataset. Subsequently, the network can be utilized with the target dataset for power-line insulator defects. Such an approach can also be used for the detection of pathological speech and gait analysis for disease classification.

To further boost performance, we introduce a computationally efficient ensemble of multiple-fine-tuned CNNs. The diversity is achieved through multiple fine tuning using different mediator datasets. Through this new approach, the multiple CNNs do not need to be trained from scratch on different data; we can use CNNs having the same architecture and trained on the same source data.

Finally, from the application-domain point of view, we propose a system in which the processing pipeline is replaced by a CNN architecture to the greatest extent possible. This approach removes any need for often tedious feature engineering. Moreover, because the input of the CNN is the image itself, there is no need for a graphical tablet or any advanced capturing device. This facilitates a very effective implementation into decision support systems or telemonitoring applications. By employing the proposed system, it is possible to capture and evaluate a handwriting/drawing sample using only a smartphone.

The remainder of this article is organized as follows. In the following section, we provide the literature survey on the handwriting analysis to support PD diagnosis. Section III describes all datasets used in this study. Then, in Section IV, we introduce the approach based on multiple fine tuning of the CNN, and derive the transferability used for capturing the benefit of multiple fine tuning. Finally, we present the results of experiments on two Parkinsonian handwriting datasets and discuss the results.

II. Handwriting Analysis to Support Parkinson's Disease Diagnosis

From a data processing point of view, we can differentiate three main strategies in computational handwriting analysis. The first strategy is based on the feature extraction from recorded signals. Typically, several types of signals are recorded depending on the device used for acquisition. This can be either a specialized pen [36] or, more frequently, a graphical digitizing tablet [37], [38]. The tablets can acquire x and y coordinates, timestamps, pressure, altitude, and azimuth. This allows the computation of basic kinematic and spatiotemporal handwriting parameters, such as velocity,

acceleration, jerk, stroke height, stroke width, and movement time [38]–[42]. To increase the performance of machine learning models, a variety of new features have been proposed; these include fractional derivatives [43], extended velocity features utilizing log-normal models, discrete transformations [16], [44], and various nonlinear features [17], [45]. A prediction accuracy of as high as 98% can be achieved on the PD handwriting dataset (PaHaw) [43] using this approach. The drawback of this approach is the need for feature engineering. Designing new features is time consuming and difficult, because it requires domain knowledge as well as extensive knowledge of signal processing and data mining.

Another approach involves using a neural network to process raw time-series data or time-series data transformed into images. With the advent of deep neural networks, feature extraction from raw data is no longer necessary because neural networks can discover patterns in raw data. The considerable success of CNNs in image recognition [46], skin cancer classification [47], and other domains indicates their suitability for processing handwriting data. Pereira *et al.* [36] proposed using a CNN for PD diagnosis from online handwriting. In addition to a microphone recording, signals, such as finger grip, axial pressure, and tilt and acceleration in the *x*/*y*-directions were captured by a special pen and recorded as a time series, and then transformed into images. A similar approach was explored by Vasquez-Correa *et al.* [10]. Here, the authors focused on the onset and offset aspects of handwriting, and achieved 67% accuracy in PD classification-based solely on online handwriting. However, this approach still requires special equipment. Either a pen or tablet capable of acquiring several different modalities of handwriting must be used to obtain accurate results.

The ultimate goal of handwriting analysis for diagnosis of PD is identifying deteriorated handwriting using imagery alone. This is the most recent approach to handwriting analysis. In this case, the only information available is visual information acquired by a camera, smartphone, or similar commercially available device. Because CNNs are currently the best option for many computer vision problems, their utilization in this task is a natural extension of CNN applications. There have been various attempts at using visual information alone to diagnose PD on the basis of offline handwriting. Naseer *et al.* [48] took advantage of a pretrained AlexNet CNN, claiming an accuracy of as high as 98%. However, it should be noted that this reported accuracy is overly optimistic, because the data samples were first augmented and then divided into training and validation sets. As a result, samples from the same subject were present in both the training and validation sets, leading to biased results. Additionally, some of the reported results were obtained through prediction on the training dataset, and as such cannot be considered to be generalizable. Similarly, as Naseer *et al.* [48] and Moetesum *et al.* [49] used a pretrained AlexNet as a feature extractor; however, in addition to raw images, they also extracted features from median filter residuals of the raw images, and edge information from images was used for classification. A support vector machine-based classifier trained on these data achieved an 83% prediction accuracy on the PaHaw dataset.

III. Data

To evaluate the proposed approach, we used two existing datasets: 1) the Parkinsonian handwriting dataset (PaHaw) [17] and 2) the NewHandPD dataset [36]. These datasets, which are the most frequently used in the literature on this topic, provide a balanced ratio between PD patients and healthy controls (HCs) and contain a sufficient number of data samples. The PaHaw dataset contains eight completed handwriting tasks from each subject: 1) Archimedean spiral; 2) letter *l*; 3) syllable *le*; 4) Czech words *les*; 5) *lektorka*; 6) *porovnat*; 7) *nepopadnout*; and 8) a sentence *Tramvaj dnes uz nepojede*. We do not use the sentence writing task because it is different from single word writing. Seventy-five subjects (37 PD, 38 HC) completed the handwriting tasks [with the exception of spiral drawing, which was successfully completed by only 69 subjects (33 PD, 36 HC)]. The NewHandPD dataset contains four repetitions of Archimedean spiral drawings and meander drawings from 66 subjects (31 PD, 35 HC).

The samples in the PaHaw dataset were acquired using a digitizing tablet whereas the NewHandPD samples were acquired using a specialized pen. Moreover, in the acquisitions for PaHaw, subjects drew spirals in bounded rectangles without any other constraints; whereas, in the acquisition of spiral and meander drawings in NewHandPD, subjects followed tracings preprinted on paper.

Both datasets provide the *x* and *y* coordinates of drawings as well as numerous other modalities. However, because our aim is to focus purely on images without using any additional signals, we only utilize output images. As such, our approach can be considered as a basic framework for decision support systems-based solely on image processing.

In addition to two Parkinsonian handwriting datasets, we employ the popular MNIST [50] and UJIpenchars2 [51] datasets. The MNIST dataset contains 70 000 handwritten numbers provided by The National Institute of Standards and Technology. The UJIpenchars2 is compiled from handwritten characters from 60 writers, totaling more than 11 000 samples.

A. Data Preprocessing and Augmentation

A relatively low number of samples in NewHandPD and PaHaw datasets can cause model overfitting. To improve performance, we artificially extended the dataset by using label-preserving transformations. The original image and newly created images are highly interdependent, and they are part of the training subset. Augmented images were shifted to a random side up to 5 pixels, rotated randomly up to 60°, or zoomed up to 20%. The tasks composed from writing words l, le, les, lektorka, porovnat, and nepopadnout were rotated only up to 20°, since this should be sufficient for human handwriting.

The NewHandPD and PaHaw datasets were resized to 224×224 pixels by nearest-neighbor interpolation. In addition, for each simulation, we selected the optimal image transformation

method from among the median filtering, Gaussian blurring, and erosion.

The MNIST dataset [50] consists of raster images of size 28×28 that were resized to 224×224 using bilinear interpolation. Because the resizing procedure causes distortion of objects, we used Zhang's skeletonizing algorithm [52] to skeletonize the image and remove the noise caused by resizing.

IV. Proposed Approach

The proposed approach is built on CNNs, with weights trained using the newly proposed multiple-fine-tuning approach; these are employed to build an ensemble of CNNs.

A. Convolutional Neural Networks

A CNN is a type of neural network suitable for tasks with input data in the form of multiple arrays, such as time series and 2-D images. CNNs consist of one input layer, one output layer, and multiple hidden layers between them. Although hidden layers are mostly convolutional layers, in some popular CNNs, such as VGG16 [53] and AlexNet [46], a few fully connected layers are added immediately before the output layer. A convolutional layer applies a convolutional operation to produce a set of linear activations that is passed into nonlinear activation functions. The output is then further modified by pooling functions, such as max pooling and min pooling, to reduce the number of parameters and the computational complexity of the network [54].

1) CNN Learning Mechanism: The neural network optimization is based on minimization of the objective function $J(\boldsymbol{\theta})$, which is perceived as an average loss over all training samples. It can be defined as

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)\sim\hat{p}_{\text{data}}} L(f(\mathbf{x};\boldsymbol{\theta}), y) \tag{1}$$

where $\hat{p}_{\text{data}}$ is an empirical distribution, L is a per-example loss function, $f(\cdot)$ is a function that generates the prediction when the input is $\mathbf{x}$, and y is the actual label [54].

The objective function $J(\boldsymbol{\theta})$ is minimized by updating the parameters $\boldsymbol{\theta}$ in the opposite direction of its gradient. There are three popular alternatives for optimizing an objective function: 1) batch gradient descent; 2) stochastic gradient descent; and 3) minibatch gradient descent [55]. The choice of a particular approach is determined by the amount of training data. Batch gradient descent processes the full training dataset simultaneously in a large batch. Stochastic gradient descent processes only a single example at a time. Most deep learning algorithms utilize minibatch gradient descent, which processes only a small portion of the dataset at once. The parameter update in minibatch gradient descent can be computed as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \frac{1}{B} \sum_{i=Bt+1}^{B(t+1)} \frac{\partial L(f(x_i;\boldsymbol{\theta}), y_i)}{\partial \boldsymbol{\theta}} \tag{2}$$

where L is a loss function of one example from the batch sampled at time t, B is size of the batch, and η is a hyperparameter known as the learning rate. Note that if B is equal to the number of samples in the training dataset, it indicates batch gradient descent, and if $B = 1$, it indicates stochastic gradient descent. The size of the batch B has a significant impact on the training process and was thoroughly discussed in [56].

Various gradient descent optimization algorithms exist, and from the model perspective, they are often treated as hyperparameters. We used the Adadelta [57] optimization algorithm, which is an extension of Adagrad [58]. Adadelta has a number of useful features, such as robustness to architectural choice, insensitivity to hyperparameters, and no manual learning rate setting.

B. Multiple-Fine-Tuned CNN

Deep CNNs provide unprecedented performance in numerous computer vision tasks. However, to achieve state-of-the-art performance, they need a substantial amount of data for training. Unfortunately, these are rarely available in some domains, including bioinformatics and biomedicine. Until recently, the utilization of CNNs for smaller datasets typically provided suboptimal results. This has changed with the advent of deep TL. When employing TL, the training data and test data are not required to be identically distributed. The network is first trained on a source dataset containing a sufficiently large amount of data and afterward, the network processes a target dataset containing a small sample of data. Using this approach, the weights of the network do not need to be learned from scratch on the target task but are only fine tuned to learn the specifics of the target task. The choice of the source task can significantly impact network performance during processing of the target dataset. The source and target tasks should appear semantically relevant for TL to work efficiently [59].

Here, we propose the concept of multiple-fine-tuned CNNs to bridge the gap between source and target tasks. In our approach, the network is first trained on a massive source dataset. In the next step, to bring the trained network closer to the target task, all layers of the network are fine tuned on the mediator dataset. The mediator dataset should have a sufficient number of samples and should be semantically closer to the target dataset than to the source dataset. Finally, because the target dataset contains a small sample of the data, some layers of the network are frozen and the remaining layers are fine tuned using the target dataset. This approach allows the network to learn, through the mediator dataset, features that are not as general as features learned from the source dataset, but more specific to the target task. The multiple-fine-tuning procedure utilized in this article is illustrated in Fig. 1.

1) Transferability: To analyze the potential of the multiple-fine-tuning approach, we utilize the concept of transferability [60]. Transferability captures the usefulness of the source CNN with regard to the target CNN. Because transferability, as proposed in [60], cannot be directly used for target tasks built using small sample datasets, we derive a modified version of this measure that can accommodate multiple fine tuning. The original version of transferability assumes that the network can be trained on the target dataset. This is not feasible in many cases, because recent CNNs are usually several layers deep and there is insufficient data for training such a deep network using target data.

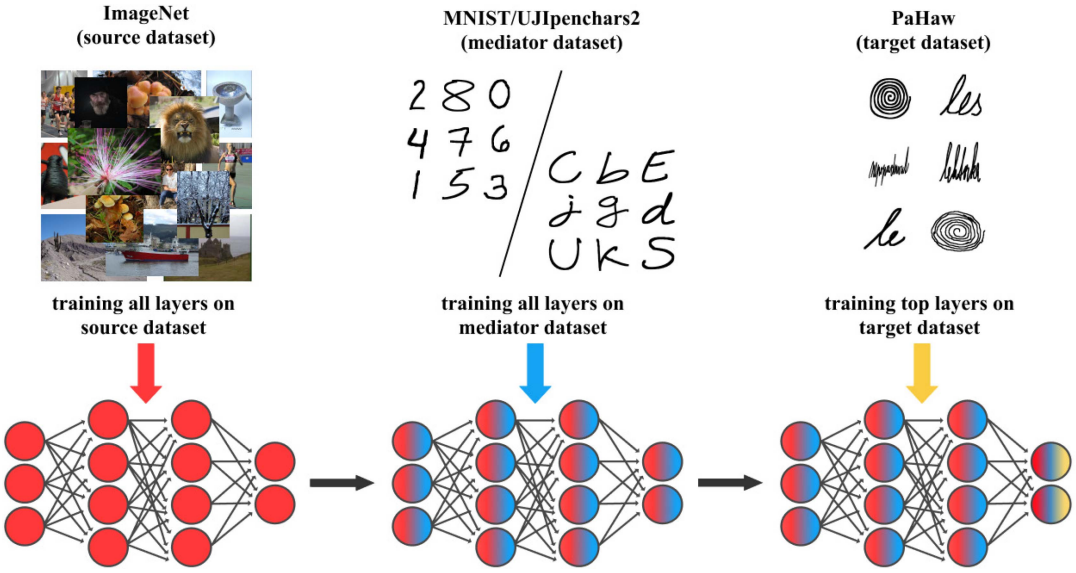

Fig. 1. Network is first trained on big Imagenet dataset. Then, all layers of the network are fine-tuned by mediator dataset. Finally, only top layers of the CNN are fine-tuned on target Parkinsonian dataset.

Let us denote the target dataset as $\mathcal{D}^T = \{\mathcal{X}^T, \mathcal{Y}^T\}$, where $\mathcal{X}^T$ are data samples and $\mathcal{Y}^T$ denotes the corresponding labels. The dataset $\mathcal{D}^T$ can be split into a training dataset $\mathcal{D}_a^T = \{\mathcal{X}_a^T, \mathcal{Y}_a^T\}$ and a testing dataset $\mathcal{D}_e^T = \{\mathcal{X}_e^T, \mathcal{Y}_e^T\}$ such as $\mathcal{D}^T = \mathcal{D}_a^T \cup \mathcal{D}_e^T$. Then, uncertainty in predicting $\mathcal{Y}_e^T$ can be expressed by entropy $H(\mathcal{Y}_e^T)$. Now, assume a CNN CNN_I that is pretrained on some large dataset, in this case, the ImageNet dataset. When this network is further fine tuned with D_a^T and used for prediction, additional information $N_{I,T}(\mathcal{X}_e^T)$ can be obtained with regard to prediction $\mathcal{Y}_e^T$. This is expressed as conditioning that reduces entropy such that $H(\mathcal{Y}_e^T) \geq H(\mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T))$.

Now, let us define the source network $\text{CNN}_{I,S}$. In general, $\text{CNN}_{I,S}$ covers two cases. In the first case, the source network is pretrained only on a large source dataset, i.e., $\text{CNN}_{I,S} = \text{CNN}_I$. As a second case, we consider the multiple-fine-tuned network, i.e., the network is first trained on a large source dataset and then all layers are further fine-tuned with another dataset—the mediator dataset. In this study, we utilize two different mediator datasets: 1) UJIpenchars2 and 2) MNIST. Because training on another dataset represents another source of additional information $N_{I,S}$, entropy is reduced by conditioning and

$$\begin{aligned} H(\mathcal{Y}_e^T) &\geq H(\mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T)) \\ &\geq H(\mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T), N_{I,S}(\mathcal{X}_e^T)). \end{aligned} \tag{3}$$

Reduction in entropy $H(\mathcal{Y}_e^T) - H(\mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T), N_{I,S}(\mathcal{X}_e^T))$ is known as information gain IG, which can be rewritten in the form of mutual information $I()$ as

$$\begin{aligned} IG = H(\mathcal{Y}_e^T) - \big[& H(\mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T)) - \\ & - I(N_{I,S}(\mathcal{X}_e^T); \mathcal{Y}_e^T | N_{I,T}(\mathcal{X}_e^T)) \big]. \end{aligned} \tag{4}$$

Equation (4) can be rewritten using the definition and properties of mutual information as

$$IG = I(\mathcal{Y}_e^T; N_{I,T}(\mathcal{X}_e^T)) + I(\mathcal{Y}_e^T; N_{I,S}(\mathcal{X}_e^T) | N_{I,T}(\mathcal{X}_e^T)). \tag{5}$$

Equation (5) consists of two parts. The first term $I(\mathcal{Y}_e^T; N_{I,T}(\mathcal{X}_e^T))$ captures mutual information considering two variables $\mathcal{Y}_e^T$ and $N_{I,T}(\mathcal{X}_e^T)$. Here, $\mathcal{Y}_e^T$ denotes labels from testing dataset $\mathcal{D}_e^T$ and $N_{I,T}(\mathcal{X}_e^T)$ is additional information that can be extracted from testing data $\mathcal{X}_e^T$ by applying feature representations learned by fine-tuned network $CNN_{I,T}$. The second part $I(\mathcal{Y}_e^T; N_{I,S}(\mathcal{X}_e^T) | N_{I,T}(\mathcal{X}_e^T))$ is conditional mutual information of random variables $\mathcal{Y}_e^T$ and $N_{I,S}(\mathcal{X}_e^T)$ given $N_{I,T}(\mathcal{X}_e^T)$, defined in general as $I(X; Y|Z) = H(X|Z) - H(X|Y, Z)$ [61]. It represents the transferability. Transferability captures the gain realized by fine tuning through the mediator dataset. For the special case in which $\text{CNN}_{I,S} = \text{CNN}_I$, transferability is equal to 0 because there is no additional source of information.

We hypothesize that multiple fine tuning provides a benefit for the target task, in comparison with the fine tuning in which only the target task is used. Transferability values for multiple fine tuning using UJIpenchars2 and MNIST mediator datasets are provided in Table I. ImageNet is used as a source dataset, and the two Parkinsonian handwriting datasets PaHaw and NewHandPD are used as the target datasets. The transferability values for both mediator datasets are nonzero for all evaluated tasks obtained from both target datasets. This indicates that multiple fine tuning provides additional information that is beneficial for the target task, thus providing some initial confidence in this approach. Another observation is that there are different transferability values for different mediator datasets. This can be used as a guideline for selecting a suitable mediator dataset. A higher transferability value indicates a greater benefit from the mediator dataset to the target task.

2) Explaining CNN Decisions: To further illustrate the decision process of CNNs that are multiple fine tuned using different datasets, we employ deep Taylor decomposition that exploits the structure of the network by backpropagating the explanations from the output layer to the input layer [62].

TABLE I
TRANSFERABILITY VALUES FOR TWO DIFFERENT MEDIATOR DATASETS. IMAGENET IS USED AS SOURCE DATASET

Handwriting task	VGG Mediator dataset UJIpenchars2	VGG Mediator dataset MNIST	SqueezeNet Mediator dataset UJIpenchars2	SqueezeNet Mediator dataset MNIST
Spiral (NewHandPD)	0.12	0.09	0.24	0.16
Meander (NewHandPD)	0.03	0.04	0.12	0.17
Spiral	0.11	0.19	0.20	0.36
l	0.08	0.10	0.14	0.07
le	0.15	0.15	0.18	0.15
les	0.06	0.05	0.14	0.13
lektorka	0.13	0.14	0.15	0.9
porovnat	0.09	0.18	0.11	0.9
nepopadnout	0.17	0.18	0.105	0.13

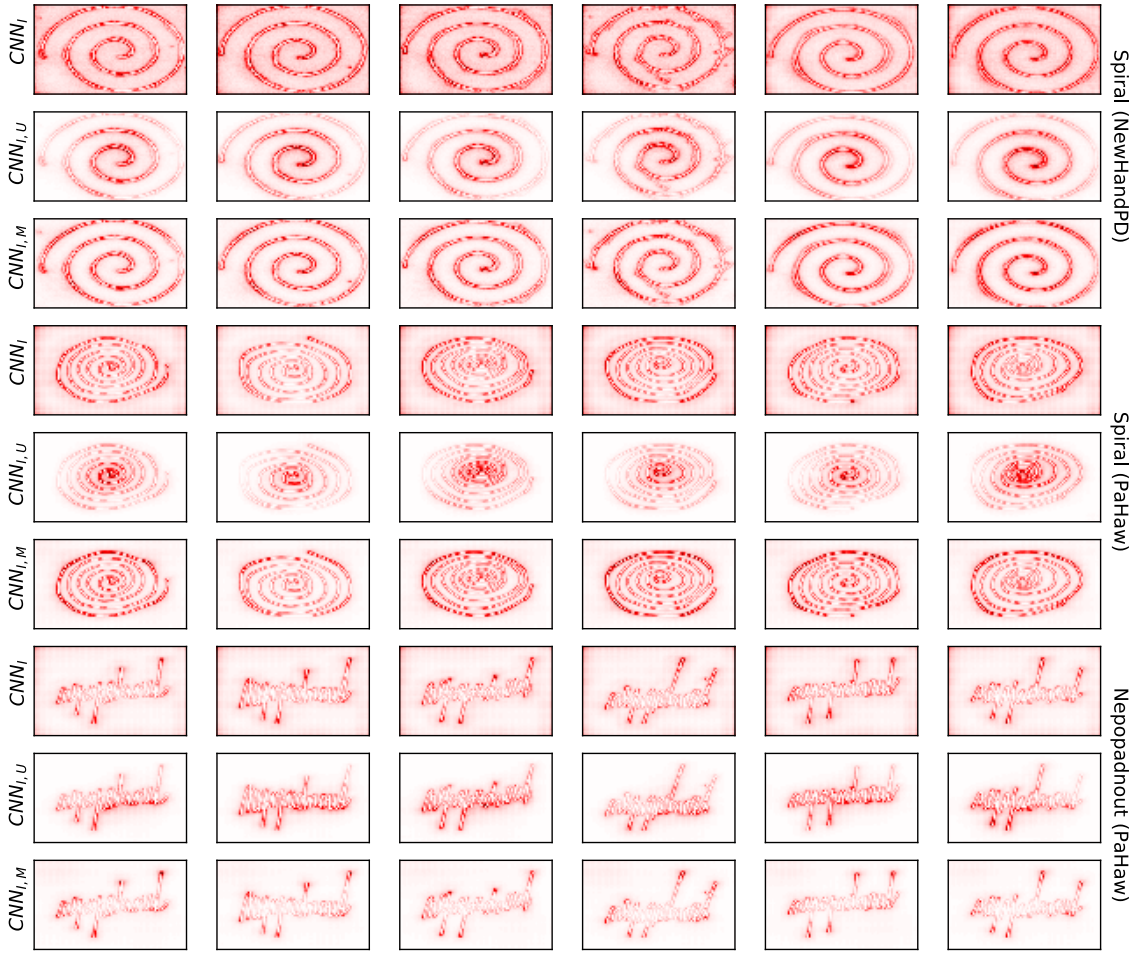

Fig. 2. Heatmaps obtained by deep Taylor decomposition applied to CNNs for six handwriting samples from three different tasks.

Our aim is to identify the differences between the decision processes of differently multiple-fine-tuned CNNs. We consider three networks: CNN_I trained on the source ImageNet dataset, $\text{CNN}_{I,M}$ similarly trained on the source ImageNet dataset, but multiple-fine-tuned using MNIST as a mediator dataset, and finally $\text{CNN}_{I,U}$, which was trained on ImageNet as a source dataset and further fine tuned on UJIpenchars2. In the last step, all three networks were fine tuned on target Parkinsonian datasets (PaHaw or NewHandPD). Fig. 2 shows an analysis of six examples from three different tasks: the spiral from the NewHandPD dataset and the spiral and writing of *nepopadnout* from the PaHaw dataset. The samples were processed by different networks and are depicted in different rows of the figure.

There are several notable differences between the heatmaps obtained from the different networks. The examples obtained from CNN_I (1st, 4th, and 7th rows), which was not multiple-fine-tuned, show a significant amount of incorrect relevance in the background. There are also differences between the results from the two multiple-fine-tuned networks. As the figure shows, $\text{CNN}_{I,M}$ (2nd, 5th, and 8th rows) puts more relevance on the middle of the handwritten symbol while $\text{CNN}_{I,U}$

(3rd, 6th, and 9th row) enhances the relevance of the pixels on the outer edges. Additionally, samples yielded by $\text{CNN}_{I,U}$ still show some small relevance in the background pixels, although it is significantly lower than that yielded by CNN_I.

C. Ensemble of Multiple-Fine-Tuned CNNs

Using an ensemble of classifiers is an approach that combines the decisions of multiple base classifiers to obtain better accuracy. It is crucial to obtain diversity between base classifiers to receive meaningful results. The most straightforward means of achieving diversity is to use different classifiers. In an ensemble of multiple-fine-tuned CNNs, the same CNN architecture is used. Diversity is achieved using various mediator datasets. Fine tuning using mediator datasets changes the weights significantly, thus influencing the feature extraction process and the structure of the CNN.

The idea of an optimal number of base classifiers was introduced in [63], which also coined the term "law of diminishing returns in ensemble construction." If the number of base classifiers is too high, performance loss may occur in the aggregation rule; conversely, a number that is too low prevents the construction of an efficient ensemble. The proposed ensemble is constructed from three CNNs, which appears to be a reasonable tradeoff between computational power and performance gain. The first CNN, CNN_I, is trained on the ImageNet dataset and later fine tuned using the Parkinsonian datasets. The other two CNNs, $\text{CNN}_{I,U}$ and $\text{CNN}_{I,M}$, are multiple-fine-tuned neural networks that are fully trained on the ImageNet dataset, then fully fine tuned using the UJIpenchars2 dataset or MNIST dataset and later partially fine tuned using the Parkinsonian datasets.

We use the majority aggregation rule to obtain predictions from the ensemble of CNNs. Here, $\hat{y}_i \in \{0, 1\}$ is the binary label predicted by the ith classifier. Then, $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{n+1})$, where n is the number of mediator datasets used, is a binary vector of predictions generated by all CNNs. The final prediction by the ensemble is $\hat{y}_e = \text{mode}(\hat{\mathbf{y}})$. Ties are resolved by random choice.

TABLE II
VGG ARCHITECTURE USED IN THIS STUDY

Layer	Size	Kernel Size	Stride
Input layer	$224 \times 224 \times 3$	-	-
Convolutional	$224 \times 224 \times 64$	3×3	1
Convolutional	$224 \times 224 \times 64$	3×3	1
Max Pooling	$224 \times 224 \times 64$	2×2	2
Convolutional	$112 \times 112 \times 128$	3×3	1
Convolutional	$112 \times 112 \times 128$	3×3	1
Max Pooling	$112 \times 112 \times 128$	2×2	2
Convolutional	$56 \times 56 \times 128$	3×3	1
Convolutional	$56 \times 56 \times 256$	3×3	1
Convolutional	$56 \times 56 \times 256$	3×3	1
Max Pooling	$56 \times 56 \times 256$	2×2	2
Convolutional	$28 \times 28 \times 256$	3×3	1
Convolutional	$28 \times 28 \times 512$	3×3	1
Convolutional	$28 \times 28 \times 512$	3×3	1
Max Pooling	$28 \times 28 \times 512$	2×2	2
Convolutional	$14 \times 14 \times 512$	3×3	1
Convolutional	$14 \times 14 \times 512$	3×3	1
Convolutional	$14 \times 14 \times 512$	3×3	1
Max Pooling	$14 \times 14 \times 512$	2×2	2
Global Pooling			
Dense Layers	$\langle 16 - 1024 \rangle \times 1$		
Output layer	1×1		

TABLE III
SQUEEZENET ARCHITECTURE USED IN THIS STUDY

Layer	Size	Kernel Size / Squeeze (for fire module)	Stride/ Expand (for fire module)
Input layer	$224 \times 224 \times 3$	-	-
Convolutional	$111 \times 111 \times 96$	7×7	2
Max Pooling	$55 \times 55 \times 96$	3×3	2
Fire	$55 \times 55 \times 128$	16	64
Fire	$55 \times 55 \times 128$	16	64
Fire	$55 \times 55 \times 256$	32	128
Max Pooling	$27 \times 27 \times 256$	3×3	2
Fire	$27 \times 27 \times 256$	32	128
Fire	$27 \times 27 \times 384$	48	192
Fire	$27 \times 27 \times 384$	48	192
Fire	$27 \times 27 \times 512$	64	256
Max Pooling	$13 \times 12 \times 512$	3×3	2
Fire	$13 \times 13 \times 512$	64	256
Convolutional	$13 \times 13 \times 1000$	1×1	1
Average Pooling	$1 \times 1 \times 1000$	13×13	1
Output layer	1×1		

V. NUMERICAL EXPERIMENTS

To evaluate the proposed approach, we used the PaHaw and NewHandPD Parkinsonian handwriting datasets. We employed two CNN architectures: 1) VGG and 2) SqueezeNet [64], both pretrained on the ImageNet dataset. VGG is a well-known CNN, proved to be efficient for many computer vision tasks. SqueezeNet is a smaller compact neural network that is suitable for mobile phones or similar hardware-constrained devices. For both networks, we used the MNIST and UJIpenchars2 datasets as mediator datasets. Additionally, we trained the VGG and SqueezeNet network on MNIST from scratch and multiple fine tuned this network using the UJIpenchars2 dataset. This allows for the evaluation of multiple fine tuning on two different source datasets.

The architectures of the VGG network and SqueezeNet network are outlined in Tables II and III, respectively. For VGG, when the network was trained on MNIST and UJIpenchars2, the global pooling layer was replaced by the flatten layer. The number of neurons on the output layer corresponds to the number of classes. For predictions on the target task (i.e., binary classification), we employ a sigmoid activation function. For nonbinary classification, the softmax activation function was used instead. Similarly, for the SqueezeNet, we adapted the last convolutional layer and pooling layer to reflect number of classes in mediator datasets and target datasets. As the loss function, we selected a cross-entropy function that frequently leads to faster training and better generalization. Because the target datasets were small, we employed early stopping to avoid overfitting. The complete set of hyperparameters used at different stages of learning is provided in Table IV. We mostly opted to use default learning parameters. For the SqueezeNet architecture we selected smaller learning rate that provided better results during exploratory experiments. Since the SGD optimizer yielded satisfactory results we used it for larger datasets. However, for small-scale datasets, PaHaw and NewHandPD,

TABLE IV
PARAMETERS USED FOR CNNS TRAINING AT DIFFERENT STAGES. EARLY STOPING PATIENCE WAS SET TO 60 FOR ALL EXPERIMENTS

Network	⟶	source dataset	⟶	mediator dataset	⟶	target dataset
VGG	pretrained	imagenet	SGD, lr=0.01, 300 epochs	MNIST UJIpenchars2	Adadelta, 300 epochs	PaHaw NewHandPD
	SGD, lr=0.01, 300 epochs	MNIST	SGD, lr=0.01, 300 epochs	UJIpenchars2		PaHaw NewHandPD
SqueezeNet	pretrained	imagenet	SGD, lr=0.001, 600 epochs	MNIST UJIpenchars2	Adadelta, 300 epochs	PaHaw NewHandPD
	SGD, lr=0.001, 600 epochs	MNIST	SGD, lr=0.001, 600 epochs	UJIpenchars2		PaHaw NewHandPD

TABLE V
PREDICTION ACCURACY AND CONFUSION MATRIX (CM) OF DIFFERENT VGG NETWORKS ON ALL EVALUATED HANDWRITING TASKS FROM NEWHANDPD AND PAHAW DATASETS. THE ELEMENTS OF CM ARE GIVEN IN FOLLOWING ORDER: TP/TN/FP/FN

	Metric	CNN_I	$CNN_{I,U}$	$CNN_{I,M}$	E-CNN	CNN_M	$CNN_{M,U}$
Spiral (NewHandPD)	Acc.	88.9 ± 5.9	92.0 ± 4.0	89.6 ± 8.0	$\mathbf{92.7 \pm 5.8}$	81.3 ± 8.4	82.5 ± 8.1
	CM	112/123/17/12	117/126/14/7	116/121/19/8	**117/128/12/7**	108/107/33/16	108/110/30/16
Meander (NewHandPD)	Acc.	88.9 ± 10.2	92.3 ± 6.5	92.7 ± 7.2	$\mathbf{94.7 \pm 7}$	89.0 ± 8.5	89.0 ± 7.5
	CM	118/117/23/6	114/130/10/10	122/123/17/2	**123/127/13/1**	114/121/19/10	110/125/15/14
Spiral	Acc.	78.5 ± 11.6	81.6 ± 8.6	83.0 ± 8.6	$\mathbf{85.8 \pm 7}$	79.5 ± 6.1	85.4 ± 4.7
	CM	24/30/6/9	26/30/6/7	23/34/2/10	**26/33/3/7**	23/32/4/10	27/32/4/6
l	Acc.	64.5 ± 6.3	67.6 ± 6.4	66.8 ± 6.2	$\mathbf{68 \pm 4}$	65.0 ± 3.2	65.0 ± 4.6
	CM	126/115/77/56	110/143/49/72	82/168/24/100	**122/132/60/60**	98/145/47/84	84/159/33/98
le	Acc.	73.5 ± 7.9	71.3 ± 9.5	68.0 ± 8.3	$\mathbf{74.7 \pm 6.9}$	65.1 ± 7.1	64.9 ± 3.4
	CM	116/161/29/69	128/141/49/57	124/132/58/61	**116/164/26/69**	115/130/60/70	97/147/43/88
les	Acc.	70.8 ± 4.0	69.9 ± 6.4	70.8 ± 4.1	$\mathbf{72.7 \pm 4.7}$	68.5 ± 7.1	68.6 ± 2.3
	CM	131/134/56/53	136/126/64/48	134/131/59/50	**140/132/58/44**	134/123/67/50	136/121/69/48
lektorka	Acc.	72.7 ± 6.2	74.7 ± 4.2	73.4 ± 7.8	$\mathbf{76.1 \pm 2.8}$	68.4 ± 6.2	72.2 ± 8.3
	CM	56/53/23/18	62/50/26/12	52/58/18/22	**58/56/20/16**	54/49/27/20	59/49/27/15
porovnat	Acc.	68.1 ± 9.1	67.8 ± 10.0	68.7 ± 10.6	$\mathbf{76 \pm 6}$	64.7 ± 6.2	68.5 ± 7.1
	CM	56/46/30/18	37/65/11/37	49/54/22/25	**57/57/19/17**	55/42/34/19	52/51/25/22
nepopadnout	Acc.	75.8 ± 4.2	78.4 ± 6.0	$\mathbf{78.5 \pm 3.7}$	$\mathbf{78.5 \pm 9.4}$	72.1 ± 4.1	77.3 ± 5.3
	CM	31/26/12/6	29/30/8/8	**29/30/8/8**	**29/30/8/8**	23/31/7/14	24/34/4/13

we selected Adadelta to avoid the need to tune the optimizer hyper-parameters on small data.

All layers of the network were fine tuned by mediator datasets; however, when training on the target task, we froze all layers up to the classification layers and trained only the top layers. For all experiments, we utilized the Keras library with TensorFlow library as a backend.

To validate the model, we employed stratified fivefold cross-validation while ensuring that handwriting samples from one subject were used only in the training dataset or testing dataset, and not in both. In each fold, 80% of data samples are used to train the model and 20% of data samples are used for model validation. For tasks containing multiple repetitions of the curve/syllable/word, each repetition is treated as single sample. The results presented in Tables V and VI constitute the average prediction accuracy over five folds. To provide more detailed view on the results, we also show the number of true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs). The tables depict the performance of six networks. Two networks were fine tuned by conventional methods using PaHaw or NewHandPD; one was pretrained on ImageNet (CNN_I), and one was pretrained on MNIST (CNN_M). Two multiple-fine-tuned networks, $CNN_{I,U}$ and $CNN_{I,M}$, were pretrained on ImageNet and then multiple-fine-tuned using MNIST or UJIpenchars2 as a mediator dataset and PaHaw or HandPD as a target dataset. Additionally, multiple-fine-tuned network $CNN_{M,U}$ was pretrained on MNIST and multiple-fine-tuned using the UJIpenchars2 and Parkinsonian datasets. Finally, we employed an ensemble of multiple-fine-tuned CNNs constructed from CNN_I, $CNN_{I,U}$, and $CNN_{I,M}$, denoted as *E*-CNN.

VI. RESULTS AND DISCUSSION

Our first observation is that the prediction accuracy achieved with the NewHandPD dataset is noticeably higher than that achieved with PaHaw. This is true for both (VGG and SqueezeNet) network architectures. The differences between these two datasets may stem from the different acquisition procedures, or simply from the different assemblage of participants. Disregarding the target dataset, we can see that the highest prediction accuracy is achieved by the proposed ensemble of multiple-fine-tuned networks. This confirms our assumption that the utilization of the mediator dataset provided additional information that created diversity in the ensemble. The advantage of multiple fine tuning is also clearly visible in the results achieved by single multiple-fine-tuned networks. In most cases, the multiple-fine-tuned networks $CNN_{I,U}$, $CNN_{I,M}$, and $CNN_{M,U}$ outperform the conventionally fine-tuned networks. This confirms our initial hypothesis, drawn in Section IV-B1, that the utilization of a mediator dataset is beneficial for predictions on the target task.

The greatest benefit of multiple fine tuning is exhibited by the SqueezeNet pretrained on the ImageNet dataset for

TABLE VI
PREDICTION ACCURACY AND CM OF DIFFERENT SQUEEZENET NETWORKS ON ALL EVALUATED HANDWRITING TASKS FROM NEWHANDPD AND PAHAW DATASETS. THE ELEMENTS OF CM ARE GIVEN IN FOLLOWING ORDER: TP/TN/FP/FN

	Metric	CNN_I	$CNN_{I,U}$	$CNN_{I,M}$	*E-CNN*	CNN_M	$CNN_{M,U}$
Spiral (NewHandPD)	Acc.	85.1 ± 4.5	88.2 ± 4.2	86.5 ± 8.9	$\mathbf{89.2 \pm 6}$	82.0 ± 5.7	87.7 ± 6.3
	CM	118/107/33/6	112/121/19/12	120/109/31/4	**119/117/23/5**	98/119/21/26	119/113/27/5
Meander (NewHandPD)	Acc.	84.4 ± 7.6	91.6 ± 9.4	$\mathbf{92.0 \pm 7.8}$	91.6 ± 10.3	87.0 ± 4.1	88.1 ± 10.0
	CM	116/107/33/8	119/123/17/5	**111/132/8/13**	119/123/17/5	108/122/18/16	118/115/25/6
Spiral	Acc.	76.9 ± 10.4	79.8 ± 6.9	77.1 ± 7.9	$\mathbf{85.7 \pm 4}$	73.5 ± 14.0	75.3 ± 5.9
	CM	21/32/4/12	25/30/6/8	20/33/3/13	**26/33/3/7**	22/29/7/11	24/28/8/9
l	Acc.	63.9 ± 5.8	64.6 ± 4.2	64.3 ± 3.4	$\mathbf{66.3 \pm 5.3}$	64.7 ± 5.9	65.1 ± 6.0
	CM	97/142/50/85	102/139/53/80	86/154/38/96	**91/156/36/91**	92/150/42/90	102/142/50/80
le	Acc.	66.9 ± 4.0	67.2 ± 6.0	66.9 ± 5.7	$\mathbf{71.4 \pm 4}$	66.6 ± 4.8	66.7 ± 3.7
	CM	122/129/61/63	115/137/53/70	96/156/34/89	**129/139/51/56**	125/125/65/60	107/143/47/78
les	Acc.	66.8 ± 4.2	67.9 ± 7.4	67.0 ± 5.9	$\mathbf{68.74 \pm 6}$	67.0 ± 5.3	67.5 ± 6.8
	CM	112/138/52/72	101/154/36/83	109/142/48/75	**111/147/43/73**	139/112/78/45	137/116/74/47
lektorka	Acc.	65.2 ± 8.9	70.6 ± 3.8	63.3 ± 7.4	$\mathbf{71.9 \pm 9.6}$	69.3 ± 2.8	73.4 ± 5.5
	CM	37/61/15/37	44/62/14/30	65/30/46/9	**54/54/22/20**	51/53/23/23	46/64/12/28
porovnat	Acc.	64.6 ± 8.8	65.8 ± 5.6	66.9 ± 6.8	$\mathbf{69.4 \pm 8}$	64.2 ± 6.2	66.6 ± 4.0
	CM	57/40/36/17	50/49/27/24	52/48/28/22	**51/53/23/23**	54/42/34/20	42/58/18/32
nepopadnout	Acc.	70.4 ± 11.1	71.8 ± 8.5	75.0 ± 7.6	$\mathbf{76.1 \pm 4.9}$	71.6 ± 6.6	72.0 ± 5.0
	CM	28/25/13/9	28/26/12/9	32/24/14/5	**32/25/13/5**	24/30/8/13	25/29/9/12

meander drawing tasks from the NewHandPD dataset; here, multiple fine tuning resulted in a performance increase from 84.4% to 92%. The highest increase in accuracy after multiple fine tuning for the VGG network is visible for spiral drawing tasks from the PaHaw dataset. This confirms that the multiple fine tuning has benefits for both investigated network architectures and both examined datasets.

The accuracy of the CNNs based on the VGG architecture is in majority cases higher than that of the SqueezeNet-based CNNs. This is more-less expected since VGG is more complex architecture that scored higher on benchmark Imagenet dataset that is used to get initial weight of both network architectures in our experiments. Another observation is that for most of the handwriting tasks VGG-based CNNs, pretrained on the ImageNet dataset, achieved higher score than the CNNs trained from scratch on MNIST dataset. It is interesting to note that even though the MNIST dataset is semantically more similar to the target task it almost always yielded poorer results than the CNN pretrained on ImageNet. This shows that similarity between target and source task in TL is not the only determining factor and other aspects have to be taken into the account. However, the results obtained for SqueezeNet architecture are noticeably different. Here, the scores of CNNs pretrained on ImageNet and those trained from scratch on MNIST are more equiponderant. This is probably due to the compact size of the SqeueezeNet. The VGG architecture needs large amount of data to train so the huge size of the ImageNet dataset is clear advantage. Whereas in case of SqueezeNet the smaller MNIST dataset is sufficient for network to learn the weights and get the performance comparable to performance of VGG-based CNNs.

For both Parkinsonian datasets and both CNN architectures, optimal performance was obtained by *E*-CNN, which yielded prediction accuracies of 94.7% for the meander task in NewHandPD and 85.8% for the spiral drawing task in PaHaw. Similarly as in the case of individual CNNs, also the ensembles performed better when the VGG architecture was utilized.

The confusion matrices provide information about number of correctly classified PD subjects (TP) and the number of correctly classified HC (TN). The models perform well on both of these classes. Moreover, there is very low FN rate, for NewHandPD datasets, indicating, that the CNN ensembles are able to identify patients suffering from disease that is crucial for each medical diagnostic system.

Comparing our results with other state-of-the-art methods is tricky since the majority of the previous papers deal with online handwriting. Online handwriting provides not only image data but also additional modalities that can be used for classification so the direct comparison would be unfair. The PD diagnosis from offline handwriting was analyzed only recently by Moetesum *et al.* [49]. Authors achieved accuracy of 83% on combination of all task from PaHaw dataset. The accuracy for single tasks is ranging from 51% on task *porovnat* to 76% on spiral drawing task. The accuracies of multiple-fine-tuned networks proposed in this article are significantly higher showing clear benefit of this approach. Not mentioning the ensemble *E*-CNN that further improved the prediction results. Another recent paper describing application of CNN to imagery data was published by Naseer *et al.* [48]. Even though they report as high as 98.28% accuracy on spiral task from the PaHaw dataset, it should be noted that the reported results are over-optimistic. The authors used the augmented versions of the same images for training and also for testing, so these results are not generalizable. Using the same data for training and testing the classifier results in high positive bias as can be seen in this case. As a such, approach proposed in our paper provides the highest achieved accuracy of PD diagnosis from offline handwriting on two popular datasets: 1) Pahaw and 2) NewHandPD.

A detailed analysis of CNN decisions revealed that for the particular patient, the decision for spiral drawing often differs from prediction for other tasks (e.g., letter and word writing). Prediction on spiral task yielded significantly higher accuracies indicating that the spiral task is more suitable for offline handwriting evaluation than the handwritten letters/words. We have investigated errors in the prediction on spiral drawing task to get a better insight on handwriting evaluation. Fig. 3 shows some examples, where multiple CNNs failed to provide

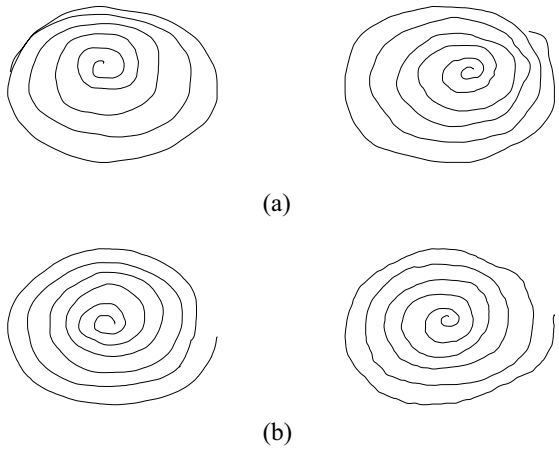

Fig. 3. Errors in PD/HC classification. (a) HC samples predicted as PD. (b) PD samples predicted as HC.

correct classification. In the case of HC that were classified as PD [Fig. 3(a)] is the reason probably the deformation of the Archimedean spiral. Both samples presented in Fig. 3(a) exhibit some deterioration. In the first case, multiple lines are blending together, and in the right-hand side figure, there is visible deformation on the right-hand side of the spiral. This may have influenced the CNN decision since we expect that the healthy subject would draw the spiral without some defects. Two false-negative examples in Fig. 3(b) illustrate different situation. Even though both these samples were acquired from PD patients, their shape is regular. The left-hand side sample does not show any visible deterioration of drawing due to the PD. PD is a complex disease that manifests itself in multiple different ways, and the deterioration of the handwriting does not need to be necessary present for every patient. The patients like this would be hardly detected by an approach based on offline handwriting. The possible option in this situation is to use some of the online handwriting features or to utilize multimodal input to a decisions support system. On the second FP sample, there is small visible deterioration due to the tremor. We hypothesize that in this case, convolutional filters in network lowered the resolution, so the network was not able to capture this deterioration and evaluate it as PD. The accuracy can be increased by considering multiple repetitions of the task or including some other modality, such as speech or online handwriting.

It was found that transferability, which was introduced in the previous section, could indicate the benefits of multiple-fine-tuning. The transferability is used as the measure to evaluate benefits of mediator datasets. We compared MNIST and UJIpenchars2, which were used for the fine-tuning of the CNN_I network. In six out of nine cases, the higher transferability value for a particular dataset yielded higher prediction performance after multiple fine tuning. However, there are some datasets, such as *le* task, where transferability was nonzero but there was no improvement in the prediction performance. We hypothesize that this may be a result of not finding the optimal weights of the network or maybe further hyperparameter tuning is necessary to find network providing better result. Unfortunately, the CNN related behaviors are often challenging to explain so it is hard to draw definitive conclusion. Transferability does not account for semantical similarity of the data, neither consider the size of the data nor the network itself, so there is a space for improvement. The topic of the suitable dataset selection for TL is still open and intensively studied. So, further developments in this area will help to identify the suitable mediator datasets for the multiple-fine-tuning.

We utilized two CNN architectures that proved notable improvements of the multiple-fine-tuning approach. However, we can expect that some of the new CNN architectures, such as Efficientnet [65], can boost the performance even further.

VII. Conclusion

In this article, we presented an efficient approach that uses static images of handwriting samples to detect handwriting deterioration due to PD. The highest prediction accuracy was 94.7% on a single drawing task. To the best of our knowledge, this is the highest accuracy achieved with the NewHandPD dataset. The prediction accuracy achieved with the PaHaw dataset was a few percentage points lower than the highest published accuracy; however, the previous approaches considered a wide range of different signal modalities, such as pressure and kinematics of pen movement. Because the target task for the CNN (i.e., offline handwriting evaluation) was semantically very different from the source task, we used an approach for fine tuning of the CNN. The multiple-fine-tuning approach proved to be beneficial and improved the prediction performance of the CNNs. Moreover, the multiple-fine-tuning approach can be used to create diversity and build an ensemble of CNNs that further boosts prediction performance. There are still some open questions that should be answered by future research. Although there is clear intuition behind the selection of the mediator dataset, there should be an approach to formalize its selection. The transferability metric used in this study can be a basis for further exploration. Additionally, although multiple fine tuning was used for the specific task of detecting pathological offline handwriting, its application domain can be significantly wider. In future research, we plan to further investigate the mechanism behind multiple fine tuning and compare it with other similar deep TL methods and different target tasks. Moreover, as future work, the larger datasets and also different network architectures need to be investigated to validate the proposed approach.

Acknowledgment

The authors would like to thank Dr. Ján Buša for valuable comments.

References

[1] A. Samii, J. G. Nutt, and B. R. Ransom, "Parkinson's disease," *Lancet*, vol. 363, no. 9423, pp. 1783–1793, 2004.

[2] A. Schrag, Y. Ben-Shlomo, and N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?" *J. Neurol. Neurosurg. Psychiat.*, vol. 73, no. 5, pp. 529–534, 2002.

[3] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *J. Neurol. Neurosurg. Psychiat.*, vol. 79, no. 4, pp. 368–376, 2008.

[4] V. Mikos *et al.*, "A wearable, patient-adaptive freezing of gait detection system for biofeedback cueing in Parkinson's disease," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 3, pp. 503–515, Jun. 2019.

[5] T. D. Pham and H. Yan, "Tensor decomposition of gait dynamics in Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 8, pp. 1820–1827, Aug. 2018.

[6] N. Naghavi and E. Wade, "Prediction of freezing of gait in Parkinson's disease using statistical inference and lower–limb acceleration data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 947–955, May 2019.

[7] E. Stack *et al.*, "Identifying balance impairments in people with Parkinson's disease using video and wearable sensors," *Gait Posture*, vol. 62, pp. 321–326, May 2018.

[8] K. Hu *et al.*, "Vision-based freezing of gait detection with anatomic directed graph representation," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 4, pp. 1215–1225, Apr. 2020.

[9] P. Harar *et al.*, "Towards robust voice pathology detection," *Neural Comput. Appl.*, vol. 32, pp. 15747–15757, Oct. 2020.

[10] J. C. Vasquez-Correa, T. Arias-Vergara, J. Orozco-Arroyave, B. M. Eskofier, J. Klucken, and E. Noth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.

[11] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 1, pp. 181–190, Jan. 2014.

[12] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.

[13] Z. Qin, Z. Jiang, J. Chen, C. Hu, and Y. Ma, "sEMG-based tremor severity evaluation for Parkinson's disease using a light-weight CNN," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 637–641, Apr. 2019.

[14] T. Arroyo-Gallego *et al.*, "Detection of motor impairment in Parkinson's disease via mobile touchscreen typing," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 1994–2002, Sep. 2017.

[15] M. Memedi *et al.*, "Validity and responsiveness of at-home touch screen assessments in advanced Parkinson's disease," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 6, pp. 1829–1834, Nov. 2015.

[16] C. D. Stefano, F. Fontanella, D. Impedovo, G. Pirlo, and A. S. di Freca, "Handwriting analysis to support neurodegenerative diseases diagnosis: A review," *Pattern Recognit. Lett.*, vol. 121, pp. 37–45, Apr. 2019.

[17] P. Drotar, J. Mekyska, I. Rektorova, L. Masarova, Z. Smekal, and M. Faundez-Zanuy, "Decision support framework for parkinsons disease based on novel handwriting markers," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 3, pp. 508–516, May 2015.

[18] C. Y. Suen, M. Berthod, and S. Mori, "Automatic recognition of handprinted characters—The state of the art," *Proc. IEEE*, vol. 68, no. 4, pp. 469–487, 1980.

[19] H. Fujisawa, "Forty years of research in character and document recognition—An industrial perspective," *Pattern Recognit.*, vol. 41, no. 8, pp. 2435–2446, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320308000964

[20] S. Singh, V. K. Chauhan, and E. H. B. Smith, "A self controlled rdp approach for feature extraction in online handwriting recognition using deep learning," *Appl. Intell.*, vol. 50, no. 7, pp. 2093–2104, 2020. [Online]. Available: https://doi.org/10.1007/s10489-020-01632-4

[21] S. Singh and A. Sharma, "Online handwritten Gurmukhi words recognition: An inclusive study," *ACM Trans. Asian Low Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–55, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3282441

[22] P. P. Roy, A. K. Bhunia, A. Das, P. Dey, and U. Pal, "HMM-based indic handwritten word recognition using zone segmentation," *Pattern Recognit.*, vol. 60, pp. 1057–1075, Dec. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316300450

[23] Z. Xie, Z. Sun, L. Jin, H. Ni, and T. Lyons, "Learning spatial-semantic context with fully convolutional recurrent network for online handwritten chinese text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1903–1917, Aug. 2018.

[24] S. Sen, S. Chowdhury, M. Mitra, F. Schwenker, R. Sarkar, and K. Roy, "A novel segmentation technique for online handwritten bangla words," *Pattern Recognit. Lett.*, vol. 139, pp. 26–33, Nov. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865518300436

[25] D. Keysers, T. Deselaers, H. A. Rowley, L. Wang, and V. Carbune, "Multi-language online handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1180–1194, Jun. 2017.

[26] E. F. B. Tasdemir and B. Yanikoglu, "A comparative study of delayed stroke handling approaches in online handwriting," *Int. J. Doc. Anal. Recognit. (IJDAR)*, vol. 22, no. 1, pp. 15–28, 2019. [Online]. Available: https://doi.org/10.1007/s10032-018-0313-2

[27] S. Mandal, S. R. M. Prasanna, and S. Sundaram, "An improved discriminative region selection methodology for online handwriting recognition," *Int. J. Doc. Anal. Recognit.*, vol. 22, no. 1, pp. 1–14, 2019. [Online]. Available: https://doi.org/10.1007/s10032-018-0314-1

[28] Y. Chherawala, P. P. Roy, and M. Cheriet, "Feature set evaluation for offline handwriting recognition systems: Application to the recurrent neural network model," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 2825–2836, Dec. 2016.

[29] X. Wu, Q. Chen, J. You, and Y. Xiao, "Unconstrained offline handwritten word recognition by position embedding integrated resnets model," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 597–601, Apr. 2019.

[30] X. Zhang, F. Yin, Y. Zhang, C. Liu, and Y. Bengio, "Drawing and recognizing chinese characters with recurrent neural network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 849–862, Aug. 2018.

[31] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1419–1434, Jul. 2019.

[32] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 136–144, Jan. 2019.

[33] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 4, pp. 1486–1498, Apr. 2020.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[35] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[36] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, "Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics," in *Proc. 29th SIBGRAPI Conf. Graph. Patterns Images (SIBGRAPI)*, 2016, pp. 340–346.

[37] P. Drotar, J. Mekyska, Z. Smekal, I. Rektorova, L. Masarova, and M. Faundez-Zanuy, "Contribution of different handwriting modalities to differential diagnosis of Parkinson's disease," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, May 2015, pp. 344–348.

[38] P. Zham, S. P. Arjunan, S. Raghav, and D. K. Kumar, "Efficacy of guided spiral drawing in the classification of Parkinson's disease," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1648–1652, Sep. 2018.

[39] E. J. Smits *et al.*, "Standardized handwriting to assess bradykinesia, micrographia and tremor in parkinson's disease," *PLoS ONE*, vol. 9, no. 5, 2014, Art. no. e97614.

[40] J. Danna *et al.*, "Digitalized spiral drawing in Parkinson's disease: A tool for evaluating beyond the written trace," *Human Movement Sci.*, vol. 65, pp. 80–88, Jun. 2019.

[41] P. Drotar, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Analysis of in-air movement in handwriting: A novel marker for parkinson's disease," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 405–411, 2014.

[42] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomed. Signal Process. Control*, vol. 31, pp. 174–180, Jan. 2017.

[43] J. Mucha *et al.*, "Identification and monitoring of Parkinson's disease dysgraphia based on fractional-order derivatives of online handwriting," *Appl. Sci.*, vol. 8, no. 12, p. 2566, 2018.

[44] D. Impedovo, "Velocity-based signal features for the assessment of parkinsonian handwriting," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 632–636, Apr. 2019.

[45] C. Rios-Urrego, J. Vásquez-Correa, J. Vargas-Bonilla, E. Nöth, F. Lopera, and J. Orozco-Arroyave, "Analysis and evaluation of handwriting in patients with Parkinson's disease using kinematic, geometrical, and non-linear features," *Comput. Methods Programs Biomed.*, vol. 173, pp. 43–52, May 2019.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[48] A. Naseer, M. Rani, S. Naz, M. I. Razzak, M. Imran, and G. Xu, "Refining Parkinson's neurological disorder identification through deep transfer learning," *Neural Comput. Appl.*, vol. 32, no. 3, pp. 839–854, Feb. 2020.
[49] M. Moetesum, I. Siddiqi, N. Vincent, and F. Cloppet, "Assessing visual attributes of handwriting for prediction of neurological disorders—A case study on Parkinson's disease," *Pattern Recognit. Lett.*, vol. 121, pp. 19–27, Apr. 2019.
[50] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[51] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
[52] T. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Commun. ACM*, vol. 27, no. 3, pp. 236–239, 1984.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.
[54] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[55] S. Ruder, "An overview of gradient descent optimization algorithms," 2016. [Online]. Available: arXiv:1609.04747.
[56] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," in *Neural Networks: Tricks of the Trade*. Heidelberg, Germany: Springer, 2012, pp. 437–478.
[57] M. D. Zeiler, "AdaDelta: An adaptive learning rate method," 2012. [Online]. Available: arXiv:1212.5701.
[58] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Jul. 2011.
[59] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[60] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognit.*, vol. 73, pp. 65–75, Jan. 2018.
[61] T. M. Cover, *The Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[62] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, Nov. 2017.
[63] H. Bonab and F. Can, "Less is more: A comprehensive framework for the number of components of ensemble classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2735–2745, Sep. 2019.
[64] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 mb model size," 2016. [Online]. Available: arXiv:1602.07360.
[65] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: arXiv:1905.11946.

Matej Gazda received the M.Sc. degree in applied computer science from the Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Kosice, Slovakia, in 2019.

His research interests include biomedical image analysis, feature selection, and biomedical decision support systems.

Máté Hireš received the M.Sc. degree in computer science from the Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Kosice, Slovakia, in 2019, where he is currently pursuing the Ph.D. degree in computer science.

His research interests include biomedical signal analysis and analysis of interval-Monge fuzzy matrices.

Peter Drotár (Member, IEEE) received the M.Sc. and Ph.D. degrees in electronics from the Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Kosice, Slovakia, in 2007 and 2010, respectively.

From 2010 to 2012, he was with the Honeywell International, Advanced Technology Europe, as a Scientist for Communication and Surveillance Systems. From 2012 to 2015, he was with SIX Research Centre, Brno University of Technology, Brno, Czech Republic, as a Postdoctoral Research Assistant. He is currently an Associate Professor with the Department of Computers and Informatics, Technical University of Kosice. He leads research and development projects concerning biomedical decision support systems. His research interests include biomedical signal processing, feature selection, and pattern recognition.