Beyond Accuracy: Enhancing Parkinson's Diagnosis with Uncertainty Quantification of Machine Learning Models

Asif Azad[1,3], Md. Saiful Islam[2,1], Ehsan Hoque[2,3], and M Saifur Rahman[1†]

[1] Department of Computer Science & Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
[2] Department of Computer Science,
University of Rochester, Rochester, New York, United States
[3] Ministry of Defence Health Services, Riyadh, Saudi Arabia

asifazad0178@gmail.com, mislam6@ur.rochester.edu,
mehoque@cs.rochester.edu, mrahman@cse.buet.ac.bd
[†]**Corresponding author**

Abstract. As deep learning and machine learning architectures have demonstrated considerable promise in clinical diagnosis, establishing their reliability has become imperative for responsible medical implementations. This research examines uncertainty estimation techniques to improve model reliability in Parkinson's disease detection. We assess Monte Carlo Dropout, Deep Evidential Classification, and Bayesian Neural Networks across three datasets representing finger tapping, facial expressions, and vocal patterns. Findings indicate that Deep Evidential Classification performs poorly in both diagnostic accuracy and uncertainty assessment, whereas Monte Carlo Dropout and Bayesian Neural Networks exhibit enhanced dependability. Integrating uncertainty estimation enables identification of ambiguous predictions, minimizing diagnostic errors and promoting secure AI adoption in medicine. Complete code and technical specifications are accessible through the official public github repository https://github.com/BRAINIAC2677/UQ4PD-ML.

Keywords: Uncertainty Quantification · Parkinson's Diagnosis · Bayesian Neural Network · MCDropout · Deep Evidential Classification.

1 Introduction and Related Literature

Machine Learning (ML) and Deep Learning (DL) have transformed medical diagnostics, offering unprecedented capabilities in disease detection and prediction. However, the clinical adoption of these models hinges not only on their accuracy but also on their reliability and interpretability. A critical yet often neglected aspect is the quantification of predictive uncertainty, which enables clinicians to assess the confidence of model outputs. Uncertainty Quantification (UQ) is particularly vital in high-stakes domains like Parkinson's Disease (PD) diagnosis,

where erroneous predictions can directly impact patient care [21]. Neural networks, despite their prowess, are frequently criticized as black-box systems due to their opaque decision-making processes and tendency toward overconfidence. Integrating UQ methods addresses these limitations fostering trust in AI-assisted diagnostics.

Parkinson's Disease exemplifies the challenges and opportunities for ML in healthcare. As the second most common neurodegenerative disorder, Parkinson's disease affected 6.1 million individuals globally in 2016, up from 2.5 million in 1990, with the prevalence more than doubling over this period due to aging populations and other contributing factors. [9,8]. The disease manifests through motor symptoms (e.g., tremor, bradykinesia) and non-motor features (e.g., cognitive decline, sleep disturbances), though clinical diagnosis remains predominantly symptom-based [19]. Traditional assessments like the MDS-UPDRS scale [13] suffer from subjectivity, while emerging biomarkers such as CSF α-synuclein assays [29] are invasive and costly. Recent advances leverage wearable sensors, video-based head pose tracking and gait analysis [3,23,34,22]. Despite their potential, wearable sensors encounter limitations in cost, comfort, and ease of use, restricting their scalability for global deployment. These limitations have spurred interest in non-invasive, scalable alternatives, particularly ML-driven analysis of multimodal data including voice, movement, and facial expressions [30].

The Health AI community has increasingly recognized the need for UQ in medical applications. While conventional models optimize for predictive performance, they often fail to signal when predictions are unreliable. This gap is especially problematic for PD, where symptom variability necessitates models that can discern ambiguous cases. Recent work, such as the Uncertainty-calibrated Fusion Network (UFNet) [17], demonstrates the potential of UQ-aware multimodal fusion, achieving robust performance by dynamically weighting task-specific predictions based on their uncertainties. However, comprehensive comparisons of UQ methods—such as Monte Carlo Dropout (MC Dropout) [10], Deep Evidential Classification (DEC) [28], and Bayesian Neural Networks (BNNs) [6]—remain underexplored for PD diagnosis.

This paper evaluates three UQ paradigms on three PD datasets across finger-tapping, facial expression, and speech modalities. Unlike prior studies emphasizing only accuracy, we systematically examine how UQ enhances diagnostic reliability by enabling models to "know when they don't know." Our findings underscore UQ's role in developing trustworthy AI systems for PD screening, particularly in resource-limited settings requiring accessible, non-invasive diagnostics.

2 Materials and Methods

2.1 Datasets and Feature Overview

In this study, we utilize three standardized datasets corresponding to the finger-tapping, smile, and speech tasks, originally curated by Islam et al. [17]. These

datasets include pre-extracted features derived from video and audio recordings of participants performing specific tasks using parktest.net [22], and are publicly released by the authors. Importantly, we do not access or analyze any raw video data or personally identifiable information. As such, Institutional Review Board (IRB) approval was not required for this study.

Finger-Tapping Task Participants were instructed to tap their thumb with the index finger ten times as quickly as possible, using both hands. The features, 130-dimensional per sample, were extracted by the Islam et al. [18] using MediaPipe Hand [14]. These features encapsulate motion attributes such as tapping frequency, amplitude, and movement interruptions, relevant to assessing bradykinesia—a hallmark motor symptom of PD [16].

Smile Task In this task, participants alternated between a neutral facial expression and a smile three times. The 42-dimensional features provided by Adnan et al. [2] were extracted using OpenFace [5] and MediaPipe, capturing metrics related to lip movement, mouth opening, facial muscle activation, and eye blinking. These features are indicative of hypomimia, or reduced facial expressiveness, often seen in PD [2].

Speech Task Participants read aloud a scripted passage including an English pangram and additional context to ensure sufficient speech length:

> "The quick brown fox jumps over a lazy dog. The dog wakes up and follows the fox into the forest. But again, the quick brown fox jumps over the lazy dog."

The shared 1024-dimensional features were extracted from a pre-trained WavLM model [7], which effectively captures acoustic and prosodic cues relevant for PD detection [1].

Table 1: Summary of the datasets and pre-extracted features used in this study.

Datasets			**Datapoints**		
Name	**Features**	**Split**	**Total**	**PD**	**Non-PD**
Finger-Tapping	130	Train	945	415	530
		Validation	221	83	138
		Test	208	69	139
Smile	42	Train	1021	339	682
		Validation	342	116	226
		Test	321	98	223
Speech	1024	Train	1007	355	652
		Validation	338	114	224
		Test	310	92	218

These pre-extracted features form a robust and comprehensive resource for multimodal behavioral analysis of PD. The diverse participant pool and well-structured splits support reliable machine learning development and evaluation.

2.2 Limitations of Softmax for Uncertainty Estimation

While softmax probabilities $p(y = k|\mathbf{x}) = e^{z_k} / \sum_i e^{z_i}$ are commonly interpreted as confidence measures, they are poorly calibrated for uncertainty estimation. Key limitations include:

- Overconfidence on out-of-distribution inputs due to exponential amplification of dominant logits [26]
- Inability to capture model uncertainty, being deterministic point estimates [15]

This makes softmax unreliable for medical diagnosis where accurate uncertainty assessment is crucial.

2.3 MC Dropout

Dropout is a widely used regularization technique in deep learning, introduced by Srivastava et al. [32], which randomly deactivates a fraction of neurons during training to prevent overfitting and improve generalization. While traditionally applied only during training, dropout's stochastic nature can also be exploited during inference to approximate Bayesian inference in neural networks. Monte Carlo Dropout (MC Dropout), proposed by Gal and Ghahramani [10], extends this idea by enabling dropout during the inference phase. By performing multiple forward passes with dropout enabled, MC Dropout generates a distribution of predictions, which can be used to quantify both aleatoric (data-related) and epistemic (model-related) uncertainty [20]. This approach provides a computationally efficient way to estimate uncertainty without requiring significant changes to the standard training pipeline, making it a practical tool for uncertainty quantification in deep learning models [31].

During training, standard dropout randomly sets activations to zero with probability p:

$$\tilde{h} = M \odot h, \tag{1}$$

where h represents the activations, $M \sim \text{Bernoulli}(1-p)$ is a dropout mask, and $\odot$ represents element-wise multiplication. During inference, instead of using deterministic forward passes, multiple stochastic passes are performed by applying dropout, yielding an ensemble of predictions:

$$\hat{y}^{(t)} = f_{\theta^{(t)}}(x), \quad t = 1, \ldots, T. \tag{2}$$

The predictive mean and variance are computed as:

$$\mathbb{E}[y] \approx \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}, \tag{3}$$

$$\mathrm{Var}(y) \approx \frac{1}{T}\sum_{t=1}^{T}(\hat{y}^{(t)} - \mathbb{E}[y])^2. \tag{4}$$

MC Dropout provides a computationally efficient approach to uncertainty estimation by leveraging existing dropout layers without requiring model retraining, making it easy to integrate into standard deep learning pipelines. Its simplicity and low computational overhead have led to widespread adoption, particularly in medical AI, where real-time uncertainty quantification is crucial.

2.4 Deep Evidential Classification

Deep Evidential Classification (DEC) models predictive uncertainty using evidential theory, offering a computationally efficient alternative to Bayesian approaches. Introduced by Sensoy et al. [28], DEC employs a Dirichlet distribution to estimate class probabilities and uncertainty in a single forward pass, making it suitable for real-time applications [24]. By interpreting network outputs as evidence for different classes, DEC bridges deep learning and probabilistic reasoning [4].

The Dirichlet distribution extends the Beta distribution to multiple categories, parameterized by a concentration vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$. Its probability density function is:

$$\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{5}$$

where $B(\boldsymbol{\alpha})$ is the multivariate Beta function. DEC replaces softmax with Dirichlet parameters derived from network outputs: $\alpha_k = f_\theta(x) + 1$. The class probability estimate is given by the Dirichlet mean:

$$p(y = k|x) = \frac{\alpha_k}{S}, \quad S = \sum_j \alpha_j. \tag{6}$$

DEC quantifies uncertainty by computing aleatoric uncertainty (data noise) $\sum_k \frac{\alpha_k}{S(S+1)}$, and epistemic uncertainty (model confidence) $\frac{K}{S}$. These measures help assess prediction reliability.

The loss function incorporates cross-entropy with regularization:

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} \log \frac{S_i}{\alpha_{ik}} + \lambda \sum_{i=1}^{N} \frac{1}{S_i}, \tag{7}$$

where the second term discourages overconfidence. Proper tuning of λ is crucial to balancing predictive accuracy and uncertainty estimation.

DEC enables efficient uncertainty estimation without multiple forward passes [28]. Unlike sampling-based methods, it directly models uncertainty, making it practical for safety-critical applications [24]. However, selecting λ is challenging, as improper tuning can lead to overconfidence or excessive uncertainty. Despite this, DEC remains a robust framework for uncertainty quantification, particularly where interpretability and efficiency are essential [12].

2.5 Bayesian Neural Network

Bayesian Neural Networks (BNNs) extend traditional neural networks by modeling weights and biases as probability distributions rather than fixed values. This enables uncertainty quantification, making BNNs particularly valuable in critical applications like medical diagnosis and autonomous systems. Instead of deterministic predictions, BNNs infer a posterior distribution over parameters given data $\mathcal{D}$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{8}$$

where $p(\mathcal{D}|\theta)$ is the likelihood, $p(\theta)$ the prior, and $p(\mathcal{D})$ the evidence. Computing this posterior is intractable in deep networks due to high-dimensional integration:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta. \tag{9}$$

Traditional neural networks approximate $p(\theta|\mathcal{D})$ by a point estimate θ^* found via maximum likelihood:

$$\theta^* = \arg\min_{\theta} \left[-\log p(\mathcal{D}|\theta)\right]. \tag{10}$$

This disregards parameter uncertainty, a limitation BNNs address using variational inference. Variational Bayesian inference approximates $p(\theta|\mathcal{D})$ with a tractable distribution $q(\theta;\phi)$ by minimizing the KL divergence:

$$\mathrm{KL}(q(\theta;\phi)||p(\theta|\mathcal{D})) = \mathbb{E}_{q(\theta;\phi)}\left[\log \frac{q(\theta;\phi)}{p(\theta|\mathcal{D})}\right]. \tag{11}$$

This leads to the evidence lower bound (ELBO):

$$\mathcal{L}(\phi) = \mathbb{E}_{q(\theta;\phi)}[\log p(\mathcal{D}|\theta)] - \mathrm{KL}(q(\theta;\phi)||p(\theta)). \tag{12}$$

A common choice for $q(\theta;\phi)$ is a Gaussian $\mathcal{N}(\mu,\sigma^2)$, allowing uncertainty propagation. The reparameterization trick facilitates optimization:

$$\theta = \mu + \sigma \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{13}$$

Predictions in BNNs are obtained via the posterior predictive distribution:

$$p(\mathbf{y}|\mathbf{x},\mathcal{D}) = \int p(\mathbf{y}|\mathbf{x},\theta)p(\theta|\mathcal{D})d\theta, \tag{14}$$

approximated using Monte Carlo sampling:

$$\mathbb{E}[\mathbf{y}|\mathbf{x},\mathcal{D}] \approx \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{x};\theta_s). \tag{15}$$

The variance of these predictions quantifies uncertainty. Bayesian Neural Networks (BNNs) provide principled uncertainty estimation and inherent regularization through prior distributions, improving robustness to noise and adversarial examples. However, they require computationally expensive inference methods, and their uncertainty quality depends on the approximation fidelity of the posterior. Despite these challenges, BNNs remain valuable for applications demanding reliable uncertainty quantification.

3 Experiments

3.1 Experimental Setup

All models were implemented using the PyTorch framework, leveraging the `torch-uncertainty` package for uncertainty-aware deep learning methods. The training procedures were executed on a single NVIDIA GeForce RTX 4070 GPU. Hyperparameter tuning was conducted using Optuna, an efficient framework for automated hyperparameter optimization, to systematically explore and identify the optimal configuration for each model. This process is crucial for achieving robust and high-performing models. The complete hyperparameter search space for result reproduction are available in the official github repository, providing full transparency and enabling verification of the reported results.

3.2 Evaluation Metrics

Classification Metrics We evaluate model performance using standard classification metrics, including Accuracy, AUROC, AUPR, and FPR95, which provide a comprehensive assessment of predictive performance.

- **Accuracy** measures the proportion of correct predictions among all samples and serves as a basic indicator of classification performance.
- **AUROC** (Area Under the Receiver Operating Characteristic Curve) evaluates the trade-off between true positive rate and false positive rate across thresholds. Higher AUROC indicates better discriminative ability.
- **AUPR** (Area Under the Precision-Recall Curve) focuses on the precision-recall trade-off and is particularly informative in class-imbalanced settings.
- **FPR95** (False Positive Rate at 95% True Positive Rate) quantifies the false positive rate when the true positive rate is fixed at 95%. It is commonly used in out-of-distribution and robustness evaluation to assess the rate of incorrect high-confidence predictions.

Uncertainty Quantification Metrics For uncertainty quantification (UQ), we employ five key metrics to assess the reliability and calibration of model predictions.

- **Expected Calibration Error (ECE)** [25] quantifies the mismatch between the predicted confidence scores and the actual correctness of model predictions. A well-calibrated model should have its confidence values align closely with empirical accuracy.
  To compute ECE, predictions are divided into M bins based on confidence scores. For each bin B_m, we calculate:
  - $\text{acc}(B_m)$: The empirical accuracy in bin B_m, defined as the proportion of correct predictions in that bin.
  - $\text{conf}(B_m)$: The average predicted confidence of samples in B_m.

The ECE is then computed as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{16}$$

where $|B_m|$ is the number of samples in bin B_m, and N is the total number of samples.
A high ECE indicates poor calibration, meaning the model's confidence does not match its actual accuracy. If a model is overconfident, it will have confidence values higher than empirical accuracy; if it is underconfident, the opposite occurs.

- **Adaptive Expected Calibration Error (aECE)** [27] extends ECE by employing adaptive binning, which ensures that each bin contains roughly the same number of samples rather than using fixed-width confidence intervals. This adaptive strategy helps mitigate binning artifacts, particularly when confidence values are unevenly distributed.
- **Area Under the Generalized Risk Curve (AUGRC)** [33] evaluates the model's uncertainty-awareness by considering risk over varying confidence thresholds. It is defined as:

$$AUGRC = \int_0^1 \text{Risk}(\tau) d\tau, \tag{17}$$

where τ represents a confidence threshold, and $\text{Risk}(\tau)$ is the classification risk at that threshold. The risk function typically considers incorrect predictions weighted by their confidence scores.
Lower AUGRC values indicate that the model maintains low risk across different uncertainty thresholds, meaning it effectively differentiates between certain and uncertain predictions.

- **Area Under the Risk Coverage Curve (AURC)** [11] measures how well uncertainty-aware rejection strategies improve prediction quality. It is defined as:

$$AURC = \sum_{i=1}^{N} \text{risk}(i) \cdot \Delta\text{coverage}(i), \tag{18}$$

where $\text{risk}(i)$ represents classification risk, and $\text{coverage}(i)$ refers to the proportion of samples retained based on confidence.
A lower AURC indicates that uncertain samples are correctly identified and discarded, leading to better overall performance when coverage is reduced.

- **Risk@80Cov** [11] measures the classification risk when retaining the top 80% most confident predictions. It provides an interpretable metric for assessing model reliability under constrained coverage settings. Formally:

$$\text{Risk@80Cov} = \text{Risk}(\tau|\text{Coverage} = 80\%). \tag{19}$$

A lower Risk@80Cov value indicates that the model maintains low risk even when most predictions are retained, reflecting strong confidence-awareness and well-calibrated uncertainty estimation.

3.3 Results

We evaluated the performance of three uncertainty quantification (UQ) methods: MC Dropout, Deep Evidential Classification (DEC), and Bayesian Neural Networks (BNN) on three distinct tasks relevant to Parkinson's disease assessment: finger-tapping, smile, and speech.

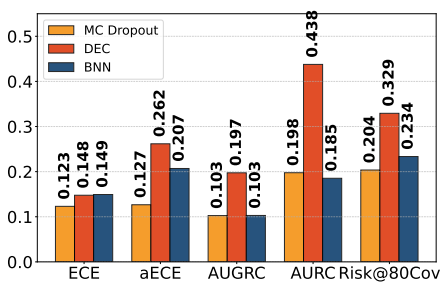
Overall, DEC consistently underperformed compared to both MC Dropout and BNN across all tasks and evaluation metrics. For instance, on the finger-tapping task, DEC achieved an accuracy of only 68.3%, markedly lower than MC Dropout (76%) and BNN (76.4%). This performance gap is further reflected in the AUROC (52.6% vs. 77.6% for BNN) and AUPR (38.2% vs. 60.5% for BNN). A similar trend is observed in the smile task, where DEC reached an accuracy of 70.1% compared to 81.3% and 77.6% for MC Dropout and BNN, respectively. Additionally, its FPR95 remained high at 94.2%, indicating poor calibration under high-recall conditions. The speech task further emphasizes this discrepancy. While MC Dropout and BNN achieved high accuracy (86.5% and 85.8%, respectively) and strong AUROC/AUPR scores, DEC lagged behind with an accuracy of 70.7%, AUROC of 51.2%, and AUPR of 30.8%. See Table 2 for detailed performance metrics.

Table 2: Classification performance of different UQ methods across Finger-tapping, Smile, and Speech tasks. Best results for each metric and task are highlighted in bold.

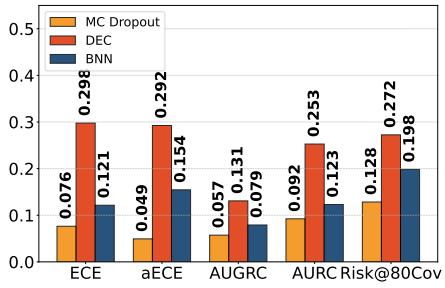
Task	Model	Accuracy	AUROC	AUPR	FPR95
Finger-tapping	MC Dropout	75.96%	77.58%	**60.70%**	82.73%
	DEC	68.27%	52.61%	38.21%	100.00%
	BNN	**76.44%**	**77.67%**	60.53%	**74.82%**
Smile	MC Dropout	**81.31%**	**85.57%**	**70.24%**	**50.22%**
	DEC	70.09%	60.13%	44.23%	94.17%
	BNN	77.57%	80.60%	62.40%	61.43%
Speech	MC Dropout	**86.45%**	85.97%	77.68%	61.01%
	DEC	70.65%	51.18%	30.81%	94.95%
	BNN	85.81%	**89.21%**	**81.56%**	**53.21%**

Figure 1 illustrate the uncertainty quantification histograms for the Finger Tapping, Smile, and Speech tasks, respectively. For the smile task (Figure 1b), we observe that MC Dropout achieves the lowest Expected Calibration Error (ECE) of 0.08, as well as favorable aECE and AURC values, indicating well-calibrated predictions. In contrast, DEC exhibits significantly poorer uncertainty calibration across all metrics (e.g., ECE = 0.30, Risk@80Cov = 0.27), suggesting

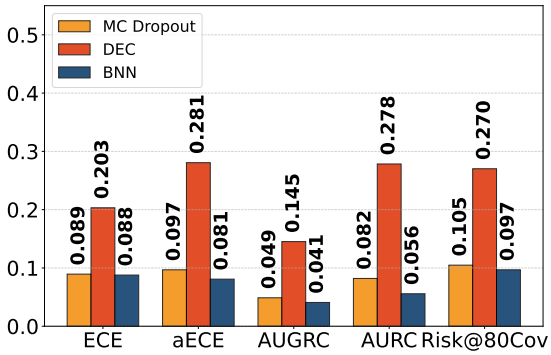
overconfident and unreliable uncertainty estimates. A similar trend is evident for the speech task (Figure 1c), where BNN outperforms both MC Dropout and DEC in terms of calibration (ECE = 0.09, AURC = 0.06). DEC again demonstrates high calibration error and risk, confirming its inferior uncertainty modeling. In the Finger Tapping task (Figure 1a), although all models show relatively higher uncertainty compared to the other tasks, MC Dropout and BNN still outperform DEC. Notably, DEC's aECE rises to 0.26 and its Risk@80Cov to 0.33, indicating unreliable predictions under high coverage. These results, consistent with the classification metrics, confirm that DEC struggles to model uncertainty effectively in these clinical tasks.



(a) Finger Tapping



(b) Smile



(c) Speech

Fig. 1: Uncertainty quantification visualizations for different models across (a) Finger Tapping, (b) Smile, and (c) Speech tasks.

4 Future Directions

Uncertainty quantification (UQ) in medical diagnosis is still an evolving field, with many opportunities for exploration and improvement. In this section, we outline three key areas for future research: expanding datasets to cover a wider

range of diseases, exploring additional UQ methods, and refining evaluation metrics for uncertainty estimation.

4.1 Expanding to More Diseases and Datasets

The application of uncertainty quantification extends beyond Parkinson's disease, holding promise for a wide array of medical conditions, including neurodegenerative disorders, cancer, cardiovascular diseases, and rare genetic conditions. Since different diseases exhibit varying patterns of uncertainty, future research should focus on expanding datasets to encompass diverse medical conditions. Collecting and curating extensive, well-annotated datasets will improve model generalizability and enhance robustness. Additionally, establishing benchmark datasets for medical uncertainty estimation, akin to ImageNet in computer vision, can standardize evaluation and foster comparative research. Close collaboration with medical professionals is crucial to ensure that these datasets are representative of real-world clinical scenarios, ultimately improving the reliability and applicability of AI-driven diagnosis.

4.2 Exploring More Uncertainty Estimation Methods

While existing uncertainty estimation techniques such as MC Dropout and Bayesian Neural Networks have demonstrated efficacy, they exhibit notable limitations. Our findings highlight the challenges faced by Deep Evidential Classification (DEC) in maintaining both classification accuracy and reliable uncertainty estimation, suggesting the need for alternative or hybrid approaches. Exploring hybrid models that integrate multiple uncertainty estimation techniques could provide more stable predictions. Transformer-based architectures, particularly for sequential medical data, may offer novel insights into uncertainty modeling. Furthermore, self-supervised learning and data augmentation strategies could enhance the reliability of uncertainty quantification. Improving model interpretability will be essential in understanding the underlying reasons behind uncertainty variations, fostering greater trust in AI-based medical diagnostics.

4.3 Developing Better Uncertainty Evaluation Metrics

The lack of universally accepted metrics for uncertainty quantification in medical AI remains a significant challenge. While metrics such as Expected Calibration Error (ECE), Adjusted ECE (aECE), and Risk@80Cov provide useful insights, they fail to comprehensively capture clinical uncertainty. Future efforts should focus on designing new metrics specifically tailored to medical applications, ensuring that they align closely with real-world diagnostic needs. Evaluating how different metrics correlate with clinical decision-making can enhance the practical utility of uncertainty estimation. Simulation-based frameworks may also prove valuable in modeling various types of uncertainty and assessing their impact. Establishing standardized benchmarks for uncertainty quantification will

facilitate fair comparisons across models, fostering advancements in reliable AI-driven medical diagnosis.

A well-defined set of UQ metrics can significantly enhance the reliability of AI-based decision-making in healthcare.

5 Conclusion

In this study, we investigated the role of uncertainty quantification (UQ) in Parkinson's disease diagnosis using deep learning models, comparing Monte Carlo (MC) Dropout, Bayesian Neural Networks (BNNs), and Deep Evidential Classification (DEC). Our findings demonstrate that MC Dropout and BNNs outperform DEC in both classification accuracy and uncertainty reliability, reinforcing the importance of robust UQ methods in medical AI. The quality of uncertainty estimation was closely linked to classification performance, with DEC's underperformance suggesting a need for methodological refinements or alternative approaches. Uncertainty-aware models can enhance clinical trust and decision-making, yet the field still requires standardized and well-validated evaluation metrics. This work advances the development of more reliable and interpretable AI for healthcare, though further research is needed to fully realize its potential in real-world clinical settings.

Acknowledgment

This research was supported by ICT Innovation Funds (FY 2023-24, Round 1), administered by the ICT Division, Government of the People's Republic of Bangladesh, the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number P50NS108676, Gordon and Betty Moore Foundation and Google Faculty Research Award. Md Saiful Islam is a recipient of the Google PhD Fellowship. M Saifur Rahman is partially supported by Basic Research Grant from BUET. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

1. Adnan, T., Abdelkader, A., Liu, Z., Hossain, E., Park, S., Islam, M., Hoque, E.: A novel fusion architecture for pd detection using semi-supervised speech embeddings. arXiv preprint arXiv:2405.17206 (2024)
2. Adnan, T., Islam, M.S., Rahman, W., Lee, S., Tithi, S.D., Noshin, K., Sarker, I., Rahman, M.S., Hoque, E.: Unmasking parkinson's disease with smile: An ai-enabled screening framework (2024), `https://arxiv.org/abs/2308.02588`
3. Ali, M.R., Sen, T., Li, Q., Langevin, R., Myers, T., Dorsey, E.R., Sharma, S., Hoque, E.: Analyzing head pose in remotely collected videos of people with parkinson's disease. ACM Transactions on Computing for Healthcare **2**(4), 1–13 (2021)

4. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
5. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). pp. 1–10. IEEE (2016)
6. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight uncertainty in neural network. In: International conference on machine learning. pp. 1613–1622. PMLR (2015)
7. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (2022). https://doi.org/10.1109/JSTSP.2022.3188113
8. De Lau, L.M., Breteler, M.M.: Epidemiology of parkinson's disease. The Lancet Neurology **5**(6), 525–535 (2006)
9. Dorsey, E.R., Elbaz, A., Nichols, E., Abbasi, N., Abd-Allah, F., Abdelalim, A., Adsuar, J.C., Ansha, M.G., Brayne, C., Choi, J.Y.J., et al.: Global, regional, and national burden of parkinson's disease, 1990–2016: a systematic analysis for the global burden of disease study 2016. The Lancet Neurology **17**(11), 939–953 (2018)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
11. Geifman, Y., Uziel, G., El-Yaniv, R.: Bias-reduced uncertainty estimation for deep neural classifiers. arXiv preprint arXiv:1805.08206 (2018)
12. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: A review of challenges and opportunities in machine learning for health. AMIA Annual Symposium Proceedings **2021**, 743 (2021)
13. Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., et al.: 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)
14. Grishchenko, I., Bazarevsky, V., Zanfir, A., Bazavan, E.G., Zanfir, M., Yee, R., Raveendran, K., Zhdanovich, M., Grundmann, M., Sminchisescu, C.: Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation. arXiv preprint arXiv:2206.11678 (2022)
15. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
16. Hughes, A.J., Daniel, S.E., Kilford, L., Lees, A.J.: Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. Journal of neurology, neurosurgery & psychiatry **55**(3), 181–184 (1992)
17. Islam, M.S., Adnan, T., Freyberg, J., Lee, S., Abdelkader, A., Pawlik, M., Schwartz, C., Jaffe, K., Schneider, R.B., Dorsey, E.R., Hoque, E.: Accessible, at-home detection of parkinson's disease via multi-task video analysis. Proceedings of the AAAI Conference on Artificial Intelligence (2025)
18. Islam, M.S., Rahman, W., Abdelkader, A., Lee, S., Yang, P.T., Purks, J.L., Adams, J.L., Schneider, R.B., Dorsey, E.R., Hoque, E.: Using ai to measure parkinson's disease severity at home. npj Digital Medicine **6**(1) (Aug 2023). https://doi.org/10.1038/s41746-023-00905-9, http://dx.doi.org/10.1038/s41746-023-00905-9

19. Jankovic, J.: Parkinson's disease: clinical features and diagnosis. Journal of neurology, neurosurgery & psychiatry **79**(4), 368–376 (2008)
20. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems. pp. 5574–5584 (2017)
21. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis. Artificial Intelligence in Medicine **150**, 102830 (2024). https://doi.org/https://doi.org/10.1016/j.artmed.2024.102830, https://www.sciencedirect.com/science/article/pii/S0933365724000721
22. Langevin, R., Ali, M.R., Sen, T., Snyder, C., Myers, T., Dorsey, E.R., Hoque, M.E.: The park framework for automated analysis of parkinson's disease characteristics. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(2) (Jun 2019). https://doi.org/10.1145/3328925, https://doi.org/10.1145/3328925
23. Liu, Y., Zhang, G., Tarolli, C.G., Hristov, R., Jensen-Roberts, S., Waddell, E.M., Myers, T.L., Pawlik, M.E., Soto, J.M., Wilson, R.M., et al.: Monitoring gait at home with radio waves in parkinson's disease: A marker of severity, progression, and medication response. Science Translational Medicine **14**(663), eadc9669 (2022)
24. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. In: Advances in Neural Information Processing Systems. pp. 7047–7058 (2018)
25. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)
26. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 427–436 (2015). https://doi.org/10.1109/CVPR.2015.7298640
27. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR workshops. vol. 2 (2019)
28. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
29. Siderowf, A., Concha-Marambio, L., Lafontant, D.E., Farris, C.M., Ma, Y., Urenia, P.A., Nguyen, H., Alcalay, R.N., Chahine, L.M., Foroud, T., et al.: Assessment of heterogeneity among participants in the parkinson's progression markers initiative cohort using α-synuclein seed amplification: a cross-sectional study. The Lancet Neurology **22**(5), 407–417 (2023)
30. Skaramagkas, V., Pentari, A., Kefalopoulou, Z., Tsiknakis, M.: Multi-modal deep learning diagnosis of parkinson's disease—a systematic review. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 2399–2423 (2023)
31. Smith, L., Gal, Y.: Uncertainty quantification in deep learning for safer ai. arXiv preprint arXiv:1811.12709 (2018)
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
33. Traub, J., Bungert, T.J., Lüth, C.T., Baumgartner, M., Maier-Hein, K.H., Maier-Hein, L., Jaeger, P.F.: Overcoming common flaws in the evaluation of selective classification systems. arXiv preprint arXiv:2407.01032 (2024)
34. Yang, Y., Yuan, Y., Zhang, G., Wang, H., Chen, Y.C., Liu, Y., Tarolli, C.G., Crepeau, D., Bukartyk, J., Junna, M.R., et al.: Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. Nature medicine **28**(10), 2207–2215 (2022)