CPSC 66 Final Report: Classifying Malignant and Benign Breast Cancer Histopathology Images

Ritika Rajamani RRAJAMA1@SWARTHMORE.EDU
Lalith Suresh LSURESH1@SWARTHMORE.EDU

Abstract

Breast cancer is aggressive and prevalent, and diagnosis often relies on labor-intensive pathology reports that can take days to weeks. Machine learning has the potential to reduce workload, accelerate diagnosis, and improve patient outcomes. In this project, we focus on minimizing false negatives, which carry high clinical costs. We evaluate a CNN with a ResNet-34 backbone for classifying breast tissue images as benign or malignant, comparing it to a baseline SVM. The CNN achieves 98.1% accuracy with few false negatives, substantially outperforming the SVM at 78.1% accuracy with numerous false negatives. However, generalizability is limited. Accuracy drops to 50.3% on data from an unseen patient. Future work is needed to improve robustness across diverse samples. Despite limitations, these results highlight the promise of CNN-based models for automated breast cancer diagnosis.

1. Introduction

Breast cancer is the most widely diagnosed cancer worldwide, with women accounting for over 99% of cases (World Health Organization, 2025b;a). Globally, 1 in 20 women will be diagnosed in their lifetime , and in the United States, this risk rises to 1 in 8 (Kim et al., 2025; Foundation, 2025). Breast cancer comprises multiple subtypes with varying tumor growth rates, and patient-specific kinetics further influence progression (Waks & Winer, 2019; Bleicher, 2018). Aggressive forms, such as triple-negative breast cancer, carry high risk for rapid advancement, making timely diagnosis critical to avoid metastasis (Waks & Winer, 2019).

Diagnosis typically begins with imaging techniques such as mammograms, ultrasounds, and MRIs, but biopsies are often required for definitive assessment (Jafari et al., 2018). Pathologists use histopathology to determine malignancy, stage, and grade (Vijayaraghavan et al., 2023). However, this process is complex and time-consuming, often taking weeks due to multiple tissue preparation and staining steps (Breast Cancer Now, 2025; Veta et al., 2014). The most labor-intensive step is manually identifying tissue features indicative of malignancy (Veta et al., 2014). Machine learning can accelerate this process by learning these features, enabling faster and more accurate classification of tissue. This has the potential to shorten diagnosis times, allowing earlier treatment and improved patient outcomes.

In this work, we evaluate a convolutional neural network (CNN) for classifying breast cancer histopathology images and compare it to a linear Support Vector Machine (SVM) baseline. Previous classification work has demonstrated significant improvement. Accuracies have increased from 91% to 98.43%, for classification of the BreakHis dataset (Yamlome et al., 2020; Balasubramanian et al., 2024). Our aim was to achieve a similar accuracy to the one in 2024, utilizing the same dataset. Furthermore, ethical considerations are crucial, as misclassifications, particularly false negatives, carry significant consequences. While machine learning is not yet ready for direct clinical deployment, this proof-of-concept demonstrates a promising avenue to improve diagnostic efficiency.

2. Methods

2.1. Model Overview

We compared two different classification models: a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). These models offer differing approaches and achieve distinct performance outcomes.

2.1.1. Support Vector Machine (SVM)

An SVM is a supervised machine learning algorithm that finds the best boundary that separates different classes from the data (GeeksforGeeks, 2025c). It is effective for high-dimensional data and provides a strong baseline because it learns a linear dimensional boundary(Scikit-learn Developers, 2025). However, the SVM is typically unable to cap-

CPSC 66 Machine Learning Proceedings, Swarthmore College, Fall 2025.

ture spatial patterns in images.

2.1.2. Convolutional Neural Network (CNN with ResNet-34 Backbone)

A CNN is designed to learn spatial features directly from images using convolutions, layers, and pooling. ResNet-34, specifically, is a deep residual network pretrained on ImageNet that uses skip connections to enable deeper architectures (GeeksforGeeks, 2025b). As histopathology images share a consistent overall structure, subtle morphological differences may be lost in shallower CNN architectures. By leveraging these residual connections, we can preserve low-level visual features while allowing deeper layers to learn minute deviations important for distinguishing malignant from benign tissue.

2.2. Dataset

We used the BreaKHis breast cancer histopathology dataset, downloaded through the KaggleHub API (Spanhol et al., 2016a). This dataset is widely used in the development of deep learning models for this classification task. After solidifying a root directory, we split the data into benign and malignant subsets. Each RGB image was loaded into Jupyter Notebook using OpenCV and resized to 224×224 pixels.

To evaluate the CNN, we split the dataset into training, validation, and test sets using train_test_split with stratification on the labels to preserve the benign/malignant ratio in each split. The final proportions are 60% train, 20% validation, and 20% test. [1]

2.3. Preprocessing

Flattened features for SVM Because SVM's operate on one-dimensional feature vectors, all images must be flattened prior to training. For this linear SVM specifically, we flatten each $224 \times 224 \times 3$ breast cancer image into a one-dimensional feature vector of length 150,528. Flattening preserves individual pixel intensities, but results in a high dimensional feature space. Each pixel is treated as an independent feature in our flattened image, so we standardize feature using a StandardScaler to ensure that all pixels contribute similarly. This prevents features with large values from disproportionately influencing the model, thereby reducing the risk of overfitting.

Tensor representation for CNN For our CNN, we retain the original structure of the images. We normalize each channel using the following normalization values to match the data distribution the ResNet-34 was originally trained on:

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225].$$

[1]The kaggle dataset can be found here: https://www.kaggle.com/datasets/anaselmasry/breast-cancer-dataset/data

2.4. Implementing Linear SVM

For the baseline classifier, we implemented a Linear SVM with Stochastic Gradient Descent. The model uses a scikit learn Pipeline consisting of a `StandardScaler` followed by an `SGDClassifier`. The `SGDClassifier` implements a linear SVM by optimizing hinge loss with L2 regularization, using stochastic gradient updates. We also pass in an alpha parameter that controls the amount of regularization we apply to the SVM.

2.5. Implementing CNN ResNet-34 Backbone

For our CNN classifier, we utilized a Resnet-34 Backbone. Each input for the CNN is represented as a tensor of shape $3 \times 224 \times 224$. The CNN processes the input through a series of residual blocks, ultimately producing a final feature map of $512 \times 7 \times 7$. We also use average pooling to reduce the feature-map to a one-dimensional feature vector. This vector is passed into a layer that outputs logits for how much the model thinks that the input is benign or malignant. During training, the model minimizes cross-entropy loss using the Adam optimizer, updating parameters.

3. Experimental Results

3.1. SVM

3.1.1. Choosing right SVM

Initially, we experimented with an SVM using a RBF kernel. Support Vector Machines (SVMs) with the Radial Basis Function (RBF) kernel are a popular choice for SVM because they can handle non-linear decision boundaries, making it suitable for a wide range of classification tasks (GeeksforGeeks, 2025a). While SVMs are generally used for high-dimensional data, the use of a distance-based kernel such as RBF becomes less effective when operating on flattened image representations, where Euclidean distances are less meaningful due to the loss of spatial structure. Since the flattened image representations produce a high-dimensional feature space, we transitioned to using a linear SVM optimized with stochastic gradient descent, which proved significantly more effective for this task. We hypothesize that this behavior arises because the linear SVM operates directly in the original feature space, while the RBF kernel projects the data into a more complex, non-linear space that is unnecessary for this representation.

3.1.2. 5-FOLD CROSS-VALIDATION

To reduce evaluation bias, we used 5-fold stratified cross-validation for the SVM. This approach provides a reliable estimate of generalization while accounting for class imbalance by preserving the original benign–malignant class distribution in each fold. In each iteration, four folds were used for training and one for testing.

3.1.3. REGULARIZATION PARAMETER TUNING

We tuned the SVM regularization parameter α, which controls the strength of the L2 penalty applied to the model weights. Hyperparameter selection was performed using grid search over the following array within the cross-validation pipeline:

$\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$

These values, in particular, were chosen to explore a range of both weaker and stronger penalties to balance model complexity and the model's ability to generalize. As a result, the optimal value of α was selected independently for each fold based on inner cross-validation performance.

3.2. CNN (with ResNet-34 backbone)

In contrast to the SVM baseline, our CNN model is capable of learning spatial feature directly from image data. The model uses convolutional operations to capture patterns and hierarchical feature representations within images. Because changing the value of one hyperparameter affects how the others behave, several training parameters were adjusted concurrently while improving accuracies. Therefore, these results should be interpreted as the result of tuning parameters as a whole rather than the effect of any single change.

3.2.1. FREEZING/UNFREEZING LAYERS

Since the dataset is already small when compared to large-scale natural image datasets, fine-tuning all layers of the CNN puts the model at risk of overfitting. To address this, we froze most of the ResNet-34 backbone and selectively unfroze deeper layers. This way, we can preserve the general low-level visual features learned from ImageNet, while unfreezing the higher layers to allows the model to adapt more specialized features needed to distinguish subtle differences between benign and malignant tissue.

Unfreezing progressively deeper layers consistently improved validation performance, with the best results obtained when both layer4 and layer3 were unfrozen. This configuration was ultimately used in the final model.

3.2.2. DATA AUGMENTATION

To improve generalization, we applied data augmentation to the training images using PyTorch, including random resized cropping, random horizontal flipping, and random vertical flipping. Initial experiments with color-based augmentation led to decreased accuracy. In histology imaging, color reflects standardized staining protocols and contains diagnostically relevant information(Veta et al., 2014). As such color variations are unlikely in real-world medical settings and degraded performance, we excluded color transformations.

3.2.3. HYPERPARAMETER TUNING

Multiple hyperparameters were tuned during the optimization process.

Learning Rate The Adam optimizer was chosen for its adaptability in adjusting learning rates for each weight update across layers (Kingma & Ba, 2017). A cosine annealing scheduler further improves convergence by taking larger steps far from the minima and smaller steps near the minima to enhance accuracy and smoothness, enabling fast and stable convergence (Mallick, 2024). This combination enables both fast and stable convergence to the optimal minima. A starting learning rate is required for both. Initially, 1×10^{-4} was tested but proved too small for efficient convergence without risking overfitting. We therefore settled on 5×10^{-4}, which balanced speed and reliable convergence.

Regularization Overfitting was a concern due to the relatively small dataset compared to ImageNet. To mitigate this, L2 regularization was applied, similar to the SVM. A value of 1×10^{-4} was used, which is commonly adopted in ResNet models.

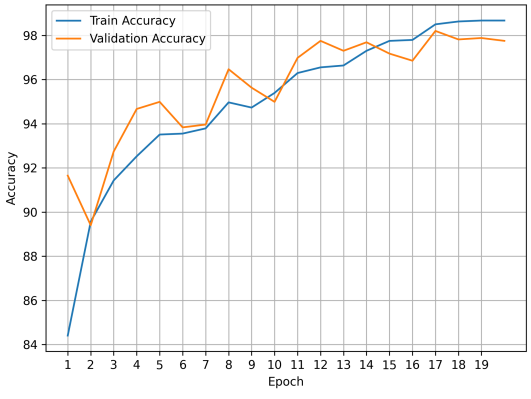

Figure 1. Train and Validation accuracy across 20 epochs. A validation accuracy was acquired before training; however, that data is withheld as to keep the axis scaled well.

Number of Epochs Choosing the number of epochs is critical. Too few causes underfitting, while too many leads to overfitting. Each epoch completes a full pass through the training dataset, allowing the model to refine weights. After extensive experimentation, we selected 20 epochs. Validation accuracy peaked around epoch 17 and then plateaued (Fig. 1). Beyond this, training loss approached zero and training accuracy neared 100%, indicating overfitting. (Figs. 1 & 2).

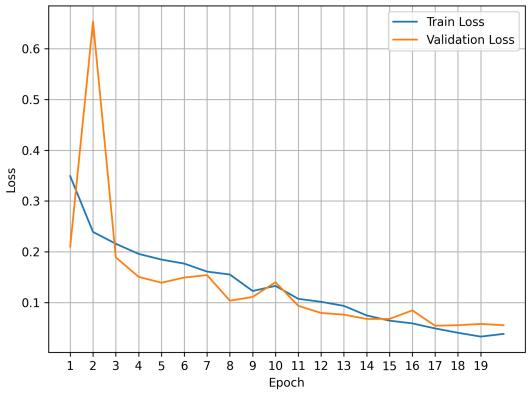

Figure 2. Train and Validation loss across 20 epochs. A validation accuracy was acquired before training; however, that data is withheld as to keep the axis scaled well.

Batch Size Mini-batch gradient descent was used, as it reduces the computational burden of batch gradient descent, which updates weights only after a full pass through all training data. It also provides smoother convergence than stochastic gradient descent, which updates after each observation. We selected a batch size of 15, balancing faster weight updates with smooth convergence to the minima.

3.3. Performance Metrics

To evaluate model performance, we display accuracy, F1-score, ROC-AUC and confusion matrices. These metrics provide additional perspectives on classification performance and are particularly important in a medical imaging context, where different types of errors carry varying consequences.

Accuracy We chose to measure accuracy to evaluate how correct our models are.

F1-Score We chose to track F1-score as it reflects the model's ability to limit both false negatives and false positives by balancing precision and recall. Specifically, it provides a reliable metric of performance for imbalanced data. In this case, the malignant class dominates, which skews the reliability of the accuracy metric.

Table 1. Performance comparison on the BreakHis dataset between the linear SVM baseline and the CNN model. SVM accuracy is reported as the mean across 5-fold cross-validation.

Model	Accuracy	F1-score	ROC-AUC
Linear SVM	78.14%	0.852	0.794
CNN	98.14%	0.986	0.998

ROC-AUC We measure ROC-AUC because it provides a measure of class separability, which is especially important given the class imbalance in our dataset. Furthermore, AUC is an especially useful metric for comparing different models.

3.4. Results and Discussion

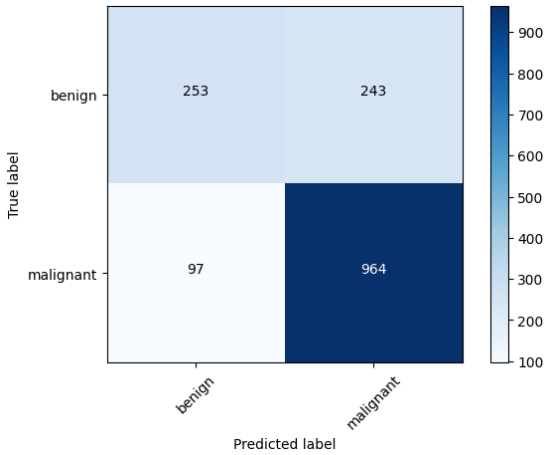

Figure 3. Confusion matrix for a representative fold of the 5-fold stratified cross-validation using the linear SVM.

Our results show that the CNN achieves substantially higher accuracy, F1-score, and ROC–AUC than the SVM baseline (Table 1). Unlike the SVM, which operates on flattened image representations and treats pixels independently, the CNN leverages the spatial tensor structure of the images. Its hierarchical architecture captures increasingly complex spatial patterns, enabling more effective differentiation between benign and malignant tissue.

The confusion matrices highlight this performance gap (Fig. 3 & 4). The SVM exhibits many false negatives, which are particularly dangerous in medical diagnosis. In contrast, the CNN greatly reduces false negatives while maintaining high F1-score and ROC-AUC.

Although the CNN performed well on the test set, the dataset included images from only 82 patients, resulting in a high image-to-patient ratio. Previous studies have mentioned that the small number of distinct patients in the BreakHis dataset leads to an inflated accuracy metric (Ben-

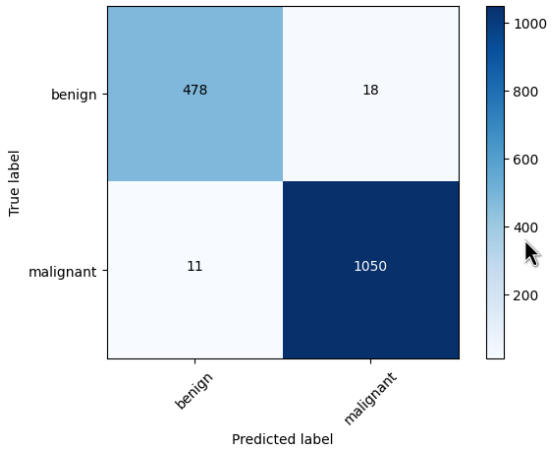

Figure 4. Confusion matrix for the CNN evaluated on the histopathology test set.

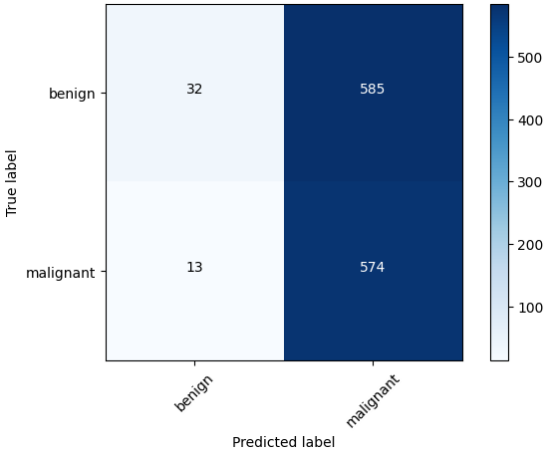

Figure 5. Confusion matrix for the CNN evaluated on images from an unseen patient.

hammou et al., 2020). In fact, studies have observed a large decrease in accuracy when moving from image-level to patient-level classification (Spanhol et al., 2016b). The reported accuracy may overestimate generalization, as many images from the same patients appear across training and test sets. The limited number of patients also restricts the model's ability to learn features that generalize, since many images share patient-specific visual characteristics. To better evaluate generalization, we tested our ResNet model on another Breast Histopathology dataset obtained using the KaggleHub API (Janowczyk & Madabhushi, 2016). [2] This dataset contains images from 279 previously unseen patients, with both classes represented. The same image processing procedures used for the BreakHis dataset were applied to these new images. Under this evaluation, the model's accuracy dropped to 50.3%, and its performance was particularly poor (Fig. 5), highlighting limited generalizability beyond the original training population. It is clear that the model predicts malignant for a large majority of the dataset, which reveals its inability to extract correct features in images from an unseen patient. While the number of false negatives is low, this is only the case due to the way the model predicted.

[2]The kaggle dataset can be found here: https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images/data

4. Conclusions and Future Considerations

For this project, we evaluated and compared the strength of an SVM (Support Vector Machine) and a CNN (Convolution Neural Network) with a ResNet-34 backbone for classifying breast cancer histopathology images as benign or malignant. Overall, our results demonstrate that our CNN outperforms the SVM across all evaluation metrics (accuracy, F1-score, ROC-AUC). This performance difference highlights the importance of learning spatial features directly from image data, which is not possible with image flattening processes required by models like SVMs. We want to emphasize, however, that our CNN's performance does not guarantee real-world efficiency, as the model's accuracy drops sharply when evaluated on images from unseen patients. Future work should prioritize training and evaluating models on larger and more diverse patient populations to improve generalization. Although this model is not yet ready for clinical use, the strong performance of our CNN in the evaluated data set demonstrates its potential as a decision-support tool in the diagnosis of breast cancer. By training and tuning our models on a more varied dataset, our model could help reduce the manual workload required for cancer diagnosis.

5. Social Implications

The primary stakeholders affected by this work include patients undergoing breast cancer screening, medical professionals, and other machine learning developers creating diagnostic tools. If successful, automated histopathology classifications could replace the labor it requires to manually diagnose breast cancer. This would ultimately improve patient health given that the patient would enter treatment earlier than if manual diagnosis was used.

However, our results also highlight dangerous risks associated with depending on a model for automated cancer diagnosis. As described in the previous subsection, there is a sharp performance decline when evaluating the CNN on images from previously unseen patients. In a clinical setting, failures from models that train on a dataset with low patient variability may disproportionately harm patients whose data distributions differ from those seen dur-

ing training. Ultimately, this may reinforce already existing healthcare disparities. While our model showed promising initial results, a long road of optimization is required before the social impacts of these machine learning models can become negligible.

Acknowledgments

We thank Professor Gabe Hope for his help in our work. We thank him for aiding us with our dataset along with helping us choose our models and best parameters for those models as well. We also thank Professor Andrew Danner who also assisted us with the downloading of our dataset.

References

Balasubramanian, Aadhi Aadhavan, Al-Heejawi, Salah Mohammed Awad, Singh, Akarsh, Breggia, Anne, Ahmad, Bilal, Christman, Robert, Ryan, Stephen T., and Amal, Saeed. Ensemble deep learning-based image classification for breast cancer subtype and invasiveness diagnosis from whole slide image histopathology. *Cancers*, 16(12), 2024. ISSN 2072-6694. doi: 10.3390/cancers16122222.

Benhammou, Yassir, Achchab, Boujemâa, Herrera, Francisco, and Tabik, Siham. Breakhis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. *Neurocomputing*, 375:9–24, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.09.044.

Bleicher, R. J. Timing and delays in breast cancer evaluation and treatment. annals of surgical oncology. *Annals of Surgical Oncology*, 25:2829–2838, 2018. doi: 10.1245/s10434-018-6615-2.

Breast Cancer Now. Your pathology results, 2025.

Foundation, National Breast Cancer. Breast cancer facts, 2025.

GeeksforGeeks. Rbf svm parameters in scikit-learn, 2025a.

GeeksforGeeks. Residual networks (resnet) - deep learning, 2025b.

GeeksforGeeks. Support vector machine (svm) algorithm, 2025c.

Jafari, Seyed Hamed, Saadatpour, Zahra, Salmaninejad, Arash, Momeni, Fatemeh, Mokhtari, Mojgan, Sadri Nahand, Javid, Rahmati, Majid, Mirzaei, Hamed, and Kianmehr, Mojtaba. Breast cancer diagnosis: Imaging techniques and biochemical markers. *Journal of Cellular Physiology*, 233(7):5200–5213, 2018. doi: 10.1002/jcp.26379.

Janowczyk, Andrew and Madabhushi, Anant. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7:29, 2016. doi: 10.4103/2153-3539.186902.

Kim, J., Harper, A., McCormack, V., and et al. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine*, 31:1154–1162, 2025. doi: 10.1038/s41591-025-03502-3.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization, 2017.

Mallick, Utkrisht. Cosine learning rate schedulers in pytorch, 2024.

Scikit-learn Developers. Support vector machines, 2025.

Spanhol, Fabio A., Oliveira, Luiz S., Petitjean, Caroline, and Heutte, Laurent. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016a. doi: 10.1109/TBME.2015.2496264.

Spanhol, Fabio Alexandre, Oliveira, Luiz S., Petitjean, Caroline, and Heutte, Laurent. Breast cancer histopathological image classification using convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2560–2567, 2016b. doi: 10.1109/IJCNN.2016.7727519.

Veta, Mitko, Pluim, Josien P. W., van Diest, Paul J., and Viergever, Max A. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014. doi: 10.1109/TBME.2014.2303852.

Vijayaraghavan, Gopal R., Guembou, Isabelle M., and Vedantham, Srinivasan. The current state of timeliness in the breast cancer diagnosis journey: Abnormal screening to biopsy. *Seminars in Ultrasound, CT and MRI*, 44(1):56–61, 2023. ISSN 0887-2171. doi: 10.1053/j.sult.2022.11.004.

Waks, Adrienne G. and Winer, Eric P. Breast cancer treatment: A review. *JAMA*, 321(3):288–300, 01 2019. ISSN 0098-7484. doi: 10.1001/jama.2018.19323.

World Health Organization. Breast cancer, 2025a.

World Health Organization. Cancer, 2025b.

Yamlome, Pascal, Akwaboah, Akwasi Darkwa, Marz, Aylin, and Deo, Makarand. Convolutional neural network based breast cancer histopathology image classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1144–1147, 2020. doi: 10.1109/EMBC44109.2020.9176594.